An Empirical Research on Identifiability and Q-matrix Design for DINA model

Peng Xu
Polytechnique Montreal
peng.xu@polymtl.ca

Michel C. Desmarais
Polytechnique Montreal
michel.desmarais@polymtl.ca

ABSTRACT

In most contexts of student skills assessment, whether the test material is administered by the teacher or within a learning environment, there is a strong incentive to minimize the number of questions or exercises administered in order to get an accurate assessment. This minimization objective can be framed as a Q-matrix design problem: given a set of skills to assess and a fixed number of question items, determine the optimal set of items, out of a potentially large pool, that will yield the most accurate assessment. In recent years, the Q-matrix identifiability under DINA/DINO models has been proposed as a guiding principle for that purpose. We empirically investigate the extent to which identifiability can serve that purpose. Identifiability of Q-matrices is studied throughout a range of conditions in an effort to measure and understand its relation to student skills assessment. The investigation relies on simulation studies of skills assessment with synthetic data. Results show that identifiability is an important factor that determines the capacity of a Q-matrix to lead to accurate skills assessment with the least number of questions.

1. INTRODUCTION

Consider a set of items intended to assess a student's mastery over a set of skills, or knowledge components (KC). These items, along with the set of skills, can be designed to test a single skill at once. Or, they can be designed to involve two or more skills. A test composed of a fixed number of items can either be composed of a mixture of single and multiple skills items, or composed of one type of items only. Skills can themselves be defined so as to facilitate the creation of task/problem items that involve single skill per item, or multiple skills per items. By which principles should a teacher choose among these different options?

This paper addresses this question, with the general objective of designing a test that will bring the most accurate assessment of a student's skill mastery state with the least number of questions items.

The investigation is framed within the DINA model, where question items can involve one or more skills, and where all skills are required in order to succeed the question, while a success can still occur through a guessing factor, and failure can also occur through a slip factor.

2. Q-MATRIX, DINA MODEL AND IDENTIFIABILITY

The mapping of items to skills is referred to as a Q-matrix, where items are mapped to latent skills whose mastery is deemed necessary in order for the student to succeed at the items. An item can represent a question, an exercise, or any task that can have a positive or negative outcome. In the DINA model, the conjunctive version of the Q-matrix is adopted: all skills are considered necessary for success.

In the last decade, a number of papers have been devoted to deriving a Q-matrix from student test results data (Barnes, 2010; Liu, Xu, & Ying, 2012; Desmarais, Xu, & Beheshti, 2015; P. Xu & Desmarais, 2016). Another line of research on Q-matrices has been devoted to refine or to validate an expert-given Q-matrix (de la Torre & Chiu, 2015; Chiu, 2013; Desmarais & Naceur, 2013). While the problems of deriving or refining a Q-matrix from data are related to Q-matrix design, they do not provide insight into how best to design them.

In parallel to these investigations, some researchers have looked at the question of the identifiability. The general idea behind identifiability is that two or more configurations of model parameters can be considered as equivalent. Sets of parameters will be considered equivalent if, for example, their likelihood is equal given a data sample. Or, conversely, if the parameters are part of a generative model, two sets of equivalent parameters would generate data having the same characteristics of interest, in particular equal joint probability distributions (see Doroudi & Brunskill, 2017, for more details).

The issue of identifiability for student skills assessment was first raised for the Bayesian Knowledge Tracing (BKT) model by Beck & Chang, 2007 and later discussed by van De Sande, 2013, and Doroudi & Brunskill, 2017. In this paper, we consider the identifiability of the Q-matrix with regards to the DINA model, which was studied by G. Xu & Zhang, 2015 and Qin et al., 2015. Detailed statistical analysis has been made under the DINA/DINO models, which was first discussed under the situation that slip and guess is zero (Chiu,

Douglas, & Li, 2009), and then the case that slip and guess exist but is known (Liu, Xu, & Ying, 2013), and finally the case that slip and guess is unknown (Chen, Liu, Xu, & Ying, 2015). These studies provide theoretical analysis of formal properties of Q-matrices and apply them to the problem of Q-matrix derivation from data, but not to the design problem itself.

Identifiability is a general concept for statistical models. Its formal definition is:

Definition (1) (Casella & Berger, 2002) A parameter θ for a family of distribution $f(x|\theta : \theta \in \Theta$ is *identifiable* if distinct values of θ correspond to distinct pdfs or pmfs. That is, if $\theta \neq \theta'$, then $f(x|\theta)$ is not the same function of x as $f(x|\theta')$.

The DINA model has parameters $\theta = \{Q, p, s, g\}$, where Q is the Q-matrix. p represents a student profile. That is, it indicates the probability that a student belongs to each profile category. For example, in a 3-skill case, there are $2^3 = 8$ categories for students to belong to, and the 8-component probability vector of students belongs to each of these categories is the model parameter p. Finally, s and g are both vectors denoting the slip and guess of each item.

The identifiability of all parameters in DINA model have been thoroughly investigated and several theorems are given (G. Xu & Zhang, 2015). But for the Q-matrix design problem that is the focus of this paper, we solely need to ensure that the model parameter p is identifiable, meaning that we can distinguish different profile categories. Fortunately, for the case when s and g are known, the requirement easily satisfied, since it only requires the Q-matrix to be *complete*.

Definition (2) (Chen et al., 2015) The matrix Q is *complete* meaning that $\{e_i : i = 1, ..., K\} \subset R_Q$, where R_Q is the set of row vectors of Q and e_i is a row vector such that the i-th element is one and the rest are zero (i.e. a binary unit vector).

And the heart of the current investigation is based on the following proposition:

Proposition (Chen et al., 2015) Under the DINA and DINO models, with Q, s and g being known, the population proportional parameter p is *identifiable* if and only if Q is *complete*.

In the next section, we investigate empirically the Q-matrix design options in light of the *completeness* requirement, using synthetic student performance data with the DINA model. Synthetic data is essential for this investigation because we need to know the underlying ground truth and return to the issue of using real data in the conclusion.

3. EXPERIMENT

The Q-matrix design problem is essentially an optimization problem. Basically, we have a pool of Q-matrices, and each of them is formed by a selection with replacement from a pool of q-vectors. Each Q-matrix will yield some capacity to diagnose students, as measured by a loss function. We aim to choose a Q-matrix that minimizes the loss function.

In our experiments, to use Q-matrix to diagnose students under DINA model, we follow a Bayesian framework. First, we use one-hot encoding to denote all profile categories. Set M to be the number of profile categories. Then, in 3-skill case, the $M = 8$ profile categories pc_i are:

$$\begin{array}{c} \\ pc_1 \\ pc_2 \\ pc_3 \\ pc_4 \\ pc_5 \\ pc_6 \\ pc_7 \\ pc_8 \end{array} \begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \left[\begin{array}{ccc} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{array}\right] \end{array}$$

Therefore, a student belonging to profile pc_1 is encoded as a binary unit vector $\alpha_1 = (1, 0, 0, 0, 0, 0, 0, 0)$, and so on for pc_2 encoded $\alpha_2 = (0, 1, 0, 0, 0, 0, 0, 0)$, ..., and pc_8 encoded $\alpha_8 = (0, 0, 0, 0, 0, 0, 0, 1)$. For this case, the DINA model parameter $p = (p_1, p_2, ..., p_8) = (P(\alpha_1), P(\alpha_2), ..., P(\alpha_8))$ where P means probability. Then, we set the prior of each student profile to be

$$\alpha_0 = (1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)$$

With the conditional independence assumed (i.e, conditioned on a given profile category, the probability to answer each question correct is independent), the likelihood is given by (De La Torre, 2009) (Chen et al., 2015)

$$\begin{aligned} L(p, Q, s, g|X) &= P(X|p, Q, s, g) \\ &= \prod_{i=1}^{I} \sum_{\alpha} p_\alpha P(X_i|\alpha, Q, s, g) \\ &= \prod_{i=1}^{I} \sum_{\alpha} p_\alpha \prod_{j=1}^{J} P_j(\alpha)^{X_{ij}} [1 - P_j(\alpha)]^{1-X_{ij}} \end{aligned} \tag{1}$$

in which X is the response matrix and X_i is the i-th row, I is the number of records(students), J is the number of questions. $P_j(\alpha)$ is the probability of student profile α to answer correctly of question j, notice α in 3-skill case has only 8 possible values, for any of them $\alpha_m, m = 1, ..., 8$, the probability is given by DINA model

$$P_j(\alpha_m) = P(X_{ij} = 1|\alpha_m) = g_j^{1-\eta_{mj}} (1 - s_j)^{\eta_{mj}}$$

in which η_{mj} is the latent response of profile α_m to question j, that is, the response when slip and guess is 0. It can be calculated by

$$\eta_{mj} = \prod_{k=1}^{K} \alpha_{mk}^{q_{jk}}$$

where K is the number of skills and q_{jk} is the (j, k)-th element of Q-matrix Q.

After having the prior and likelihood, the posterior $\hat{\alpha}$ for each student can be calculated. It has the form

$$\hat{\alpha} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{p}_6, \hat{p}_7, \hat{p}_8)$$

and then we calculate the loss between this posterior and the true profile α_{true}, which is one of the one-hot encoding vector.

In short, for any Q-matrix configuration, the loss function is defined by

$$loss(Q) = \sum_{i \in \text{students}} \|\hat{\alpha}_i - \alpha_{\text{true}}\|^2$$

To implement the experiment, for each Q-matrix configuration, we generate a response matrix based on DINA model given fixed slip and guess parameters, using function 'DINAsim' from R package 'DINA' (Culpepper, 2015), then calculate the posterior estimation for all students and evaluate the total loss. The reported result is an average loss of 100 runs.

In our experiments, we consider the 3-skills case and 4-skills case respectively. For 3-skills case, we conduct it with $N = 200$ students, of which 25 students falling into each of 8 categories. For 4-skills case, we conduct it with $N = 400$ students, of which 25 students falling into each of 16 categories.

3.1 Experiment 1: Comparison of three strategies

In the first experiment, we compare three different Q-matrix design strategies. They are all based on repetition of a specific pool of q-vectors.

- Strategy 1: Using the identifiability condition (definition (1)) by only repeatedly using the vectors $\{e_i : i = 1, ..., K\}$ (binary unit vectors, or one-hot encodings). Q-matrix used in this strategy is denoted as Q-matrix 1.
- Strategy 2: Using the vectors $\{e_i : i = 1, ..., K\}$ plus an all-one vector $(1, 1, 1)$ (in 3-skill case) or $(1, 1, 1, 1)$ (in 4-skill case). This is inspired by orthogonal array design, which is a commonly seen design of experiments (Montgomery, 2017). Q-matrix used in this strategy is denoted as Q-matrix 2.
- Strategy 3: Repeatedly using all q-vectors. Q-matrix used in this strategy is denoted as Q-matrix 3.

For the 3-skills case, all these three Q-matrices are shown in Figure 1. The general pattern is to recycle the rows above the lines denoted by ...[..., ..., ...].

The 4-skills case is similar, which is omitted here. Results of these two cases are shown in Figure 2 and Figure 3.

3.2 Experiment 2: Find best configuration

The second experiment takes the brute force approach. We directly examine all possible Q-matrix configurations. First, for a given pool of q-vectors to choose from and an integer indicating the number of questions, we need to know the number of possible configurations of Q-matrices we have. This is equivalent to a classical combinatorial problem, that is, to allocate distinguished balls (q-vectors) to indistinguished cells (questions). It can be easily computed by combinatorial coefficients and interpreted by using stars and bars methods. For example, in 3-skills case, we have 7 q-vectors, and if we have 4 questions to allocate them, then

Q-matrix 1
(binary unit vectors)

	k_1	k_2	k_3
q_1	1	0	0
q_2	0	1	0
q_3	0	0	1
...	...	...	...
q_{19}	1	0	0
q_{20}	0	1	0
q_{21}	0	0	1

Q-matrix 2
(binary unit + all-1s vectors)

	k_1	k_2	k_3
q_1	1	0	0
q_2	0	1	0
q_3	0	0	1
q_4	1	1	1
...	...	...	...
q_{17}	1	0	0
q_{18}	0	1	0
q_{19}	0	0	1
q_{20}	1	1	1
q_{21}	1	0	0

Q-matrix 3
(all combinations)

	k_1	k_2	k_3
q_1	1	0	0
q_2	0	1	0
q_3	0	0	1
q_4	1	1	0
q_5	1	0	1
q_6	0	1	1
q_7	1	1	1
...	...	...	...
q_{15}	1	0	0
q_{16}	0	1	0
q_{17}	0	0	1
q_{18}	1	1	0
q_{19}	1	0	1
q_{20}	0	1	1
q_{21}	1	1	1

Figure 1: Q-matrix design strategies

we have $\binom{4+7-1}{7-1} = 210$ possible configurations. This number grows up sharply as a number of questions increases or number of patterns increases. As a comparison, in the 4-skills case, if we have 5 questions to allocate them, then we have $\binom{5+15-1}{15-1} = 11628$ possible configurations.

For each configuration, we calculate the MAP estimation for all categories of each student, and compare with the one-hot encoding for their true categories. The total loss is reported as the performance index.

We show the results of 6 combinations of different numbers of skills and questions:

- 3-skills case, 4 questions: Figure 4, Figure 5
- 3-skills case, 8 questions: Figure 6, Figure 7
- 4-skills case, 5 questions: Figure 8, Figure 9

4. RESULT

All results are given in form of figures.

5. DISCUSSION

From the result of experiment 1 we can see that strategy 1 always works better than the other two strategies, meaning that simply repeating the vectors $\{e_i : i = 1, ..., K\}$ in Q-matrix design, without using any combination of skills, yields better student diagnosis performance.

From the result of experiment 2, when slip and guess parameters are as low as 0.01, we can see obvious graded patterns among different configurations. This can be explained by the distinguishability of Q-matrix. For example,

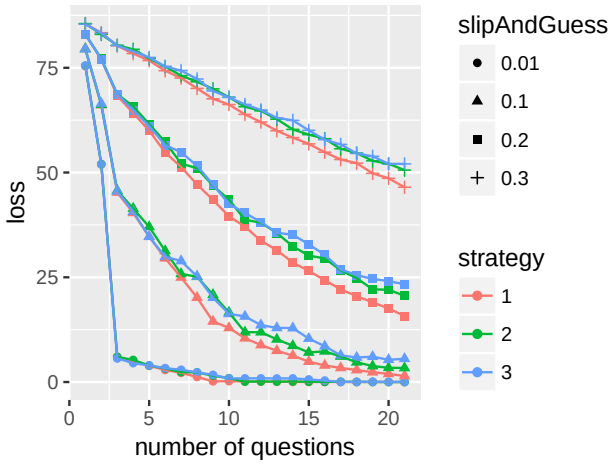

Figure 2: Three Strategy Comparison on 3-skills case

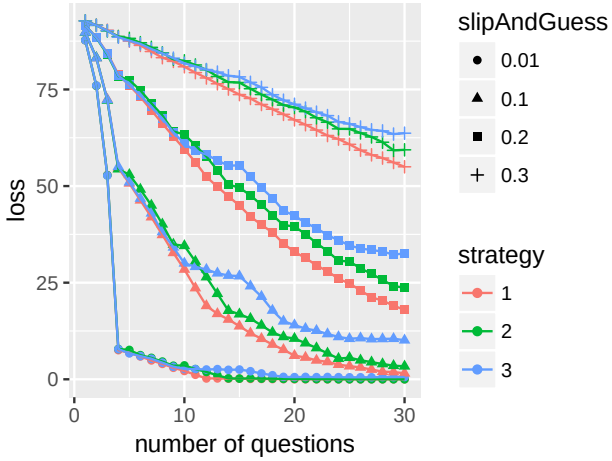

Figure 3: Three Strategy Comparison on 4-skills case

in Figure 4, we can see there are 7 layers. In fact, the first layer consisted of Q-matrix that can only cluster students into 2 categories. One example of such kind Q-matrix is

$$\begin{array}{c} \\ q_1 \\ q_1 \\ q_1 \\ q_1 \end{array} \begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

This Q-matrix can only discriminate between a student that mastered skill 1 or not. We know that there are in fact 8 categories of students, the 7 layers in Figure 4 from top to bottom correspond to the Q-matrix that can separate students into 2 to 8 categories. We can see that complete Q-matrices always fall in the bottom layer, which concurs with the proposition of Section 2. The 4-skills case is similar in Figure 8.

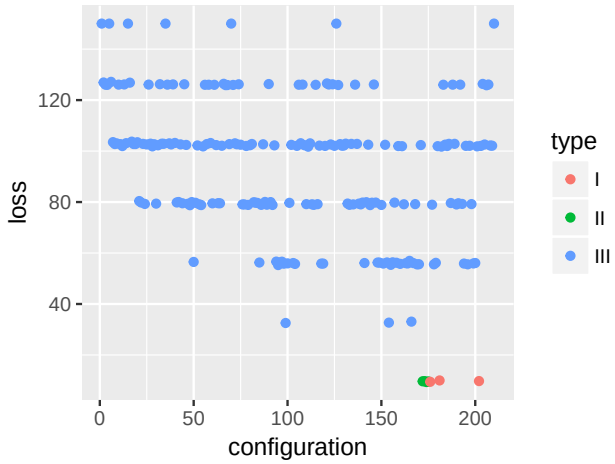

Figure 4: 3-skills case, slip=guess=0.01, J=4

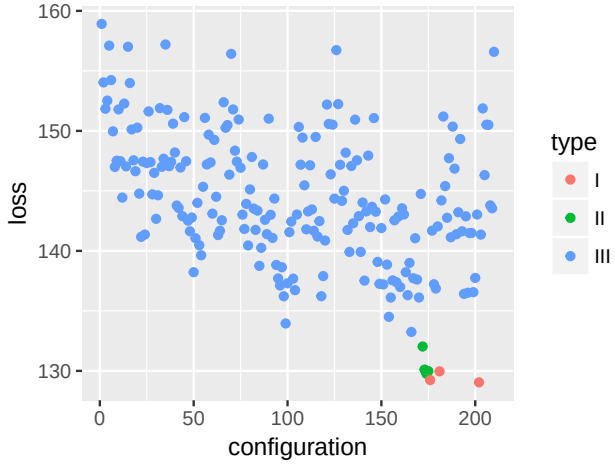

Figure 5: 3-skills case, slip=guess=0.2, J=4

When slip and guess parameter increase, it makes the points becomes more divergent which can be easily seen by comparison between Figure 4 and Figure 5. In order to see some more details, we distinguish three types of Q-matrices.

- Type I: Complete and confined, meaning it is only consisted of vectors $\{e_i : i = 1, ..., K\}$.
- Type II: Complete but not confined, meaning it not only contains all vectors $\{e_i : i = 1, ..., K\}$, but also contains at least one other q-vector.
- Type III: Incomplete Q-matrix.

Type I and Type II Q-matrices performs the same when slip and guess are low (Figure 4, Figure 8), but when they get higher, Type I Q-matrices show a better performance (Figure 5, Figure 9).

However, when more questions are involved in high slip and

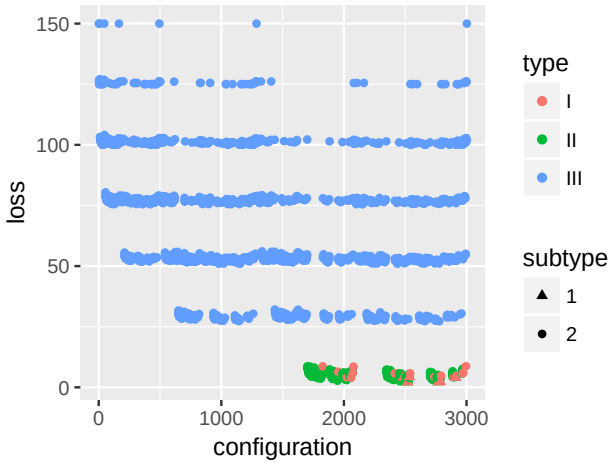

Figure 6: 3-skills case, slip=guess=0.01, J=8

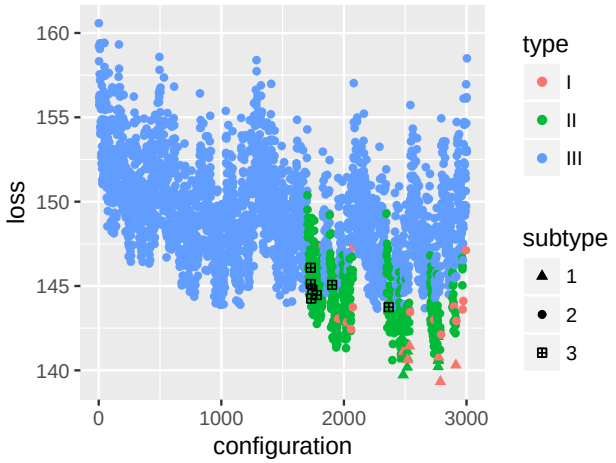

Figure 7: 3-skills case, slip=guess=0.3, J=8

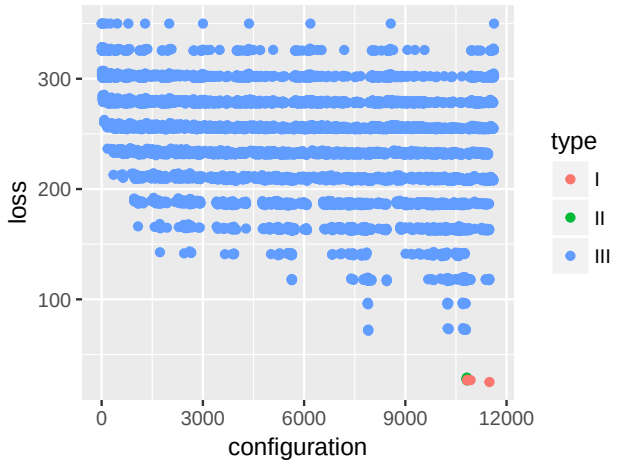

Figure 8: 4-skills case, slip=guess=0.01, J=5

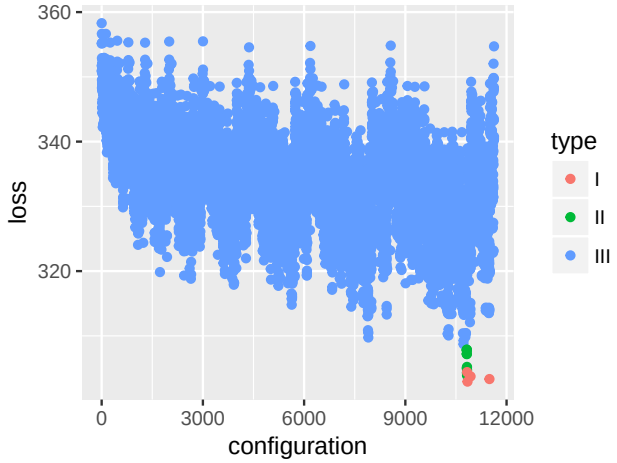

Figure 9: 4-skills case, slip=guess=0.2, J=5

guess, the performance becomes more unstable. Therefore, we again consider more subtypes. In 3-skills case for 8 questions, we consider three subtypes below.

- Subtype 1: Q-matrix contains each component of $\{e_i : i = 1, ..., K\}$ at least twice.
- Subtype 2: Other situations (e.g A complete Q-matrix but all the other vectors are just repeated e_1).
- Subtype 3: Q-matrix contains all q-vectors.

From Figure 7 we can see that the subtype 1 (denoted by triangle) shows better performance than subtype 2, meaning that repeating the whole set of $\{e_i : i = 1, ..., K\}$ is a better strategy just like the strategy 1 we used in experiment 1. Subtype 3 corresponds to the strategy 3 in experiment 1, it has only 7 possible configurations in 8-question setting and we can see that they do not perform well.

Therefore, we argue that the best Q-matrix design is to use only the vectors $\{e_i : i = 1, ..., K\}$ since it offers quicker convergence speed (as shown in experiment 1) and better robustness against slip and guess (as shown both in experiments 1 and 2).

6. CONCLUSION

This work is still in an early stage and has limitations, in particular because it is conducted with synthetic data, but the main finding is wide reaching and warrants further investigations. The support for designing Q-matrices that satisfy the identifiability condition by single-skill items is compelling in the experiments conducted with synthetic data. The results clearly show such matrices yield more accurate student skills assessment. In particular, they show that Q-matrices that contains items that span the whole range of potential combinations of skills tend to yield lower skills assessment than Q-matrices that simply repeat the pattern of single-skill items.

The finding that tests composed of single-skill items are better for skills assessment is somewhat counter-intuitive, as intuition suggests that a good test should also include items with combinations of skills. But intuition also suggests that items that involve combination of skills are more difficult, and it may not simply be because they involve more than one skill. It might be that solving items that combine different skills in a single problem is a new skill in itself. This conjecture is in fact probably familiar to a majority of educators, and the current work provides formal evidence to support it. And the immediate consequence is that Q-matrices, as we currently conceive them, fail to reflect that a task that combines skill involves a new skill.

Ideally, future work should be conducted with real data. However, given that we do not know the real Q-matrix that underlies real data, investigating the questions raised by the current study is non trivial. Meanwhile, further experiments with synthetic data can be considered with different choices on student profiles distribution, and different number of skills involved. Besides, the case where slip and guess are unknown should also be considered, which involves a different identifiability requirement (G. Xu & Zhang, 2015).

References

Barnes, T. (2010). Novel derivation and application of skill matrices: The Q-matrix method. *Handbook on educational data mining*, 159–172.

Beck, J. E., & Chang, K.-m. (2007). Identifiability: A fundamental problem of student modeling. In *International conference on user modeling* (pp. 137–146).

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633.

Culpepper, S. A. (2015). Bayesian estimation of the dina model with gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454-476. Retrieved from `http://www.hermanaguinis.com/pubs.html` doi: 10.3102/1076998615595403

De La Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*, 1–21.

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In *Artificial intelligence in education* (pp. 441–450).

Desmarais, M. C., Xu, P., & Beheshti, B. (2015). Combining techniques to refine item to skills q-matrices with a partition tree. In *Educational data mining 2015*.

Doroudi, S., & Brunskill, E. (2017). *The misidentified identifiability problem of bayesian knowledge tracing*. International Conference on Educational Data Mining, EDM2017.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied psychological measurement*, *36*(7), 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, *19*(5A), 1790.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Qin, C., Zhang, L., Qiu, D., Huang, L., Geng, T., Jiang, H., ... Zhou, J. (2015). Model identification and Q-matrix incremental inference in cognitive diagnosis. *Knowledge-Based Systems*, *86*, 66–76.

van De Sande, B. (2013). Properties of the bayesian knowledge tracing model. *Journal of Educational Data Mining*, *5*(2), 1–10.

Xu, G., & Zhang, S. (2015). Identifiability of diagnostic classification models. *Psychometrika*, 1–25.

Xu, P., & Desmarais, M. (2016). Boosted decision tree for Q-matrix refinement. In *Edm* (pp. 551–555).