General Cognitive Diagnosis Model for Response Time

Abstract: This study proposes a general cognitive diagnosis model for response times (CDM-RT). The idea of the model is to consider person parameters defined on fine-grained attributes to improve the model fit and skill diagnosis related to RTs. The model framework is specified under two distinct motivation conditions and two treatments of attribute profiles. By comparing the attribute profile estimated from CDM for RA and CDM for RT, we can investigate students' problem-solving behavior and detect possible abnormal patterns. Both simulation study and real data analysis are conducted via a fully Bayesian approach with Markov Chain Monte Carlo (MCMC) method.

Keywords: Response time, Total Time limits, Adaptive Truncation, Item Position Effect

Introduction

The introduction of computer-based assessments in education enables researchers and practitioners to collect and analyze process data in standardized operational settings. For instance, students' response times (RTs) to items would be recorded in order to obtain more comprehensive information about the amount of labor needed to complete an item as well as the speed at which tasks are completed (Partchev et al., 2013; van der Linden, 2009). Meanwhile, the information included in RTs may well be utilized to improve standard testing techniques such as item calibration, adaptive item selection, and latent ability estimate, as well as to explore and evaluate test performance-related aspects (Fox, Entink, & van der Linden, 2007).

It has long been recognized that RTs, being most widely discussed process information in research and practice, are critical sources of data for describing examinees' behavioral patterns and mental activity (Schnipke & Scrams, 2002). The study of statistical modeling of RT stretches all the way back to the 1960s, when Rash (1960) used a Gamma distribution for the RT and a Poisson distribution for the number of items completed to model the tests. Thissen (1983) combined a two-parameter logistic item response model with a linear model for the logarithm of response time, including a normally distributed error factor, to account for both speed and power. With RTs as the outcome variable, the Weibull and lognormal distributions (Rouder, Sun, Speckman, Lu, & Zhou, 2003) are used to examine latent parameters. Klein Entink, van der Linden, and Fox (2009), on the other hand, used a family of Box-Cox transformations to approximate data produced by Weibull, Gamma, and exponential models. However, these models only give a one-dimensional representation of examinees' speed along a latent continuum, ignoring the assessment's various fine-grained skills. Meanwhile, examinees' motivations and the nature of the evaluation are often ignored.

Although the construct of speed has a long history in both individual differences psychology and educational measurement (e.g., Gulliksen, 1950; Kelley, 1927; Thorndike et al., 1926), the majority of previous advancements have placed a larger focus on response accuracy (RAs) than on other process data during assessments (Lee & Chen, 2011). Cognitive diagnostic models (CDMs) have been created to represent the multidimensional evaluation of students' abilities. Numerous specific and general formulations have been created in psychometrics that focus only on RAs. Among these models are the deterministic input, noisy "and" gate (DINA; Junker & Sijtsma, 2001), the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006), the log-linear CDM (Henson, Templin, & Willse, 2009), the generalized DINA model (G-DINA; de la Torre, 2011), and the general diagnostic model (GDM; von Davier, 2005). With the exception of Minchen and de la Torre (2018), the discussion and application of CDM to a continuous response such as RT remain restricted.

We propose a general CDM framework for RTs (CDM-RT) in this article, which may be thought of as a combination of G-DINA (de la Torre, 2011) and the lognormal model (van der Linden, 2006). The model's goal is to include both attribute-based item characteristics and examinee parameters in order to improve model fit and skill diagnosis for RTs. We begin by defining the model framework for different motivation conditions. Meanwhile, two treatment are explored for attribute profiles. Then, Bayesian estimation is used to estimate model parameters. The purpose of simulation studies is to determine the robustness of item parameter estimations and the predictive accuracy of individual parameters. Finally, a real-world data analysis is undertaken utilizing PISA 2012 computer-based mathematics data.

Model

Model specification

CDMs, as a restricted latent class model, provide detailed information about examinees' learning strengths and weaknesses, and assign examinees into more qualitative groups based on a set of cognitive skills. Let us assume there are K distinct attributes (or called skills) being measured in an assessment. To identify which attribute is required in each item, the Q-matrix (Tatsuoka, 1983) is an integral component of almost all CDM frameworks. Q-matrix is of dimension $J \times K$ with J representing the number of items. Entries of Q-matrix are binary (i.e., $q_{jk} \in \{0,1\}$) to indicate whether item j requires the attribute k or note. To apply the CDMs for RT, we follow the general framework of G-DNIA, which allows for a unique probability of RT for each of the possible latent groups. The general framework of CDM-RT can be expressed as:

$$\log\left(T_{ij}|\alpha_{ik}^{*}\right) = \mu_{ij} + \varepsilon_j \tag{1}$$

$$= \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_{ik}^{*}$$

$$+ \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \beta_{jkk'} q_{jkk'} \alpha_{ik}^{*} \alpha_{ik'}^{*} + \cdots \beta_{j12\ldots K} q_{j12\ldots K} \prod_{k=1}^{K} \alpha_{ik}^{*} + \varepsilon_j$$

Here, T_{ij} represents the RT of examinee i on item j. The lognormal distribution is used to describe RT with random error $\varepsilon_j \sim N\left(0, \sigma_j^{-2}\right)$, where σ_j is item discrimination (van der Linden, 2006). The mean of the log-normal distribution (μ_{ij}) is described as the linear combination of person parameters and item parameters, based on the Q-matrix. Following the definition of G-DINA, $\alpha_{ij}^{*} \in \{0,1\}$ is defined to represents whether examinee i master the required attributes for item j. With K attribute, there will be 2^K possible patterns in an unrestricted attribute space (e.g., no attribute hierarchy is specified), ranging from lacking all of K attributes to mastering all attributes.

For item parameters, baseline item intensity β_{j0} (intercept) represents the amount of labor required by the items indirectly, through its effect on the observed RTs. β_{jk} represents the main effect due to attribute k. $\beta_{jkk'}$ represents the first-order interaction effect due to attributes k and k'. And, $\beta_{j12\ldots K_j^*}$ represents the highest-order interaction effect due to all required attributes. Many popular CDMs, such as DINA and NIDO, could be viewed as a specific cases of G-DINA with different model assumptions.

Model Constraints and Its Relationship with Motivation

For G-DINA, the main effects are usually constrained to be nonnegative since we do not expect to see the probability of success could decrease when students master one additional attribute. However, constraints could be more complex for RT modeling. For the low-stake condition, examinees have low motivation and aim to spend as less unnecessary time as possible. Then, baseline item intensity (β_{0j}) could be viewed as the RT for examinees to read the item and make the decision of whether additional effort should be given. In this case, the main effects (δ_{jk}) should be non-negative, which represents the additional time examinees are willing to spend. Then, the baseline item intensity is the lower bound, and mastering the required attribute is a *stimulus* for longer RTs.

For the high-stake condition, a different monotonicity constraint should be specified since examinees have high motivation and aim at giving the correct answer as long as it does not time out. Then, baseline item intensity (β_{0j}) is the RT for examinees who do not master any required attribute to finish the item. The main effect (δ_{jk}) should be non-positive, which represents the time examinees could save given the attribute k. In this case, baseline item intensity is the upper bound, and mastering the required attribute is a reward for shorter RTs.

Two Treatments of Attribute Profile: Manifest and Latent

There are two different possible treatments of attribute profile. Traditionally, attribute profile is defined as a binary vector $\boldsymbol{\alpha}$ for each student, which indicates whether or not the student master attribute. Given RAs, examinees' attribute profile could be estimated by a variety of CDMs based on different model assumptions. Thus, estimated attribute profile could be taken as a manifest variable in the CDM-RT. In this way, CDM-RT becomes the multivariate regression with monotonicity constraints. Item features, defined as baseline item time intensity, main effects, and interactions effects, are the main results. The main purpose of this specification is to analyze the speed-accuracy trade off in detail. Based on the manifest information of attribute profile, we can have a deeper understanding about how RTs are generated.

Secondly, we could take attribute profile as the latent variable as CDMs. With specified Q-matrix, we assign examinees with the attribute profile which could best predict observed RTs. Thus, the estimated attribute profile should *not* be defined upon the mastery condition of attribute since the information of RA is not incorporated. Instead, it purely indicates whether student spend time on (or attempt to solve the problem using) the required attribute during the assessment in general. Compared with all the other examinees, when a student spends less RTs on the items related to a specific attribute, the estimated latent attribute profile would be zero. Meanwhile, main stimulus and reward time effects indicate the additional time examinees could spend or save if they indeed attempt to use the knowledge of corresponding attributes.

Table 1. Contingency Table for Latent Attribute Profile

	$\alpha_{ik} = 1$	$\alpha_{ik} = 0$
$\alpha_{ik}^{*} = 1$	Reasonable Attempt	Unreasonable Attempt
$\alpha_{ik}^{*} = 0$	Unreasonable Neglect	Reasonable Neglect

Ideally, examinees should only spend time on the attribute they master and not waste time on the attribute they do not master. However, the latent attribute profile of CDM-RT (α_{ik}^{*}) does not always be the same as the latent attribute profile of CDMs (α_{ik}). *Table 1* summarized four conditions of matching between two attribute profiles. When $\alpha_{ik} = 1$ and $\alpha_{ik}^{*} = 1$, student i is expected to master the attribute k and spend adequate RTs upon attribute k. This represents a common problem-solving behavior. We name this case as a reasonable attempt, which indicates that student attempts to apply the attribute that they master to solve the related items. When $\alpha_{ik} = 0$ and $\alpha_{ik}^{*} = 0$, student i has low response accuracy and spend few RT on the items related to attribute k. We name this case as reasonable neglect since students could avoid wasting time on the items that they are less likely to get correct. However, if a student has reasonable neglect behaviors upon almost all attributes, it will be a signal of rapid guessing since he/she takes almost all items at a relatively fast speed with low accuracy. Let us name the proportion of reasonable neglect across all attributes for a student as the reasonable neglect index (RNI; $\frac{1}{k}\sum_{k} I(\alpha_{ik}^{*} = 0)I(\alpha_{ik} = 0)$), which range from 0 to 1. When $\alpha_{ik} = 1$ and $\alpha_{ik}^{*} = 0$, student i has high response accuracy and fast speed. We name this case the unreasonable neglect. This indicates that students' proficiency on attribute k is high. However, a student has unreasonable neglect behaviors upon almost all attributes, it could also be viewed as a signal of cheating. In practice, it will be challenging to distinguish the student with the significant high ability from the cheater. We name the proportion of unreasonable neglect across all attributes for a student as unreasonable neglect index (UNI; $\frac{1}{k}\sum_{k} I(\alpha_{ik}^{*} = 0)I(\alpha_{ik} = 1)$). It relies on more additional information to decide whether a high NNI is due to high proficiency or cheating. Finally, when $\alpha_{ik} = 0$ and $\alpha_{ik}^{*} = 1$, student i spend the unignorable amount of time on the attribute k even he/she does not master it. We name this case as the unreasonable attempt. We name the proportion of unreasonable attempt

across all attributes for a student as the unreasonable attempt index (UAI; $\frac{1}{k}\sum_k I(\alpha_{ik}^* = 1)I(\alpha_{ik} = 0)$). A large value of NAI is a strong single of high motivation. However, low NAI does not necessarily mean students have low motivation.

Estimation

Parameters in CDM-RT can be estimated via the Bayesian approach with the Markov Chain Monte Carlo (MCMC) method. In Bayesian estimation, the prior distribution of parameters and observed data likelihood produce a joint posterior distribution. In this study, the Gibbs sampler (Gelfand & Smith, 1990) is used with R2jags package (Version 0.6.1; Su & Yajima, 2015) in R (Version 4.0.2; R Core Team, 2016). Assuming local independence, $\log(T_{ij})$ and α_{ik} are conditionally and independently distributed as $\log(T_{ij}) \sim^{iid} N(\mu_{ij}, \sigma_j^{-2})$ and $\alpha_{ik} \sim^{iid} Bernoulli(p(\alpha_{ik} = 1))$. The prior distribution of item parameters is specified as a weekly informative distribution. We assume baseline item intensity $\beta_{j0} \sim^{iid} N(0,1)$, item discrimination $\sigma_j^2 \sim^{iid} InvGamma(1,1)$, and all interactive time effect $\beta_{jkk'} \sim^{iid} N(0,10)$. For main time effect, we assume $\beta_{jk} \sim^{iid} N(0,10)I(\beta_{jk} \leq 0)$ in high motivation condition, and $\beta_{jk} \sim^{iid} N(0,10)I(\beta_{jk} \geq 0)$ for low motivation. For the person parameter, we follow the setting of Zhan, Jiao, Man, & Wang (2019). We assume that each examinees' attribute profile is randomly generated from all 2^{K^*} possible patterns of attribute profile. A vector of probability ($\boldsymbol{P}$) is first generated by Dirichlet distribution with all hyperparameters as 1 ($\boldsymbol{P}$~Dirichlet (**1**)). Then an index (c) is generated from the categorical distribution ($c \sim Categorical(\boldsymbol{P})$). c represents which attribute from all 2^K possible attribute profiles is sampled.

Given the priors specified above, the joint posterior probability for the can be expressed as:

$$P(\boldsymbol{\theta}|\log(\boldsymbol{T})) \propto L(\log(\boldsymbol{T})\,|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}) \times p(\boldsymbol{\alpha}) \times p(\boldsymbol{\beta}) \times p(\boldsymbol{\sigma^2})$$

, where $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}\}$ is a set of all latent variables, and

$$L(\log(\boldsymbol{T}) \,|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}) = \prod_{i=1}^{I} \prod_{j=1}^{J} p(\log(T_{ij}) | \alpha_i, \boldsymbol{\beta_j}, \sigma_j^2)$$

is the likelihood. In this study, we take the posterior median as the point estimate. For MCMC estimation, four Markov chains are used for sampling with 12000 iterations and 10000 burn-in for each chain.

Simulation

In this section, we examine the performance of CDM-RT in two different aspects. First, we relax the model assumption to check the robustness of item parameter estimations for CDM-RT with manifest attribute profile. Then, we investigate how accurately could we estimate the attribute profiles under different conditions for CDM-RT with latent attribute profiles. Without loss of generality, we focus on the high-stakes setting.

Simulation Study for Manifest Attribute Profile

With manifest attribute profile α_ik from CDM, CDM-RT makes a strong assumption that examinees spend time on and only on the attribute they master consistently during the assessment (named as 'mastery attempt matching' assumption). In this simulation study, we assume that there are 10 items with 100, 500, and 1000 examinees. Table 2 indicates the proposed Q-matrix, which would not bring identification issues, based on the validation studies in CDMs (i.e., Xu & Zhang, 2016; Fang, Liu, & Ying, 2019). We first generate the attribute profile for each examinee randomly from all possible attribute patterns. Then, we randomly select 5%, 20%, and 50% elements of the attribute profile matrix and change their values. These random errors represent the chance of examinees either spending time on the attribute they do not master or do not spend the time on the attribute they master. This revised attribute profile describes whether examinees indeed spend time

on (or attempt to solve the problem using) the required attributes. Baseline item intensities (β_{j0}) are sampled from the standard normal distribution. Item time discrimination (σ_j^2) is sampled from the truncated normal distribution with a mean as 1.875, standard deviation as 1, and lower bound of 0. Main time effects (β_{jk}) are sampled from the standard normal distribution with upper bound as 0, and standard deviation as 1. Interaction effects are sampled from the standard normal distributions. Using the revised attribute profile and generated true item parameters, we simulate the RTs. Then, generated RTs and *unrevised* attribute profile is used in CDM-RT to estimate the item parameters, which mimics the situation when the mastery attempt matching assumption is not satisfied.

Table 2. Q-matrix for simulation study

Item	α_1	α_2	α_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	1	0
5	1	0	1
6	0	1	1
7	1	1	1
8	1	1	0
9	0	1	1
10	1	0	1

Table 3 summarizes the performance of item parameter estimates under different conditions. To measure the performance of the estimations, we calculate the median of mean absolute error (MSE) for different types of item parameters estimators across 100 independent replications. In general, the biases under all conditions are unignorable compared with the original scale of true parameters. As expected, the biases for all item parameters tend to be larger with a higher level of random error. The estimators of item time intensity and main time effects are more robust than interactive time effects and item discrimination. Meanwhile, increasing the number of examinees reduces the bias for all item parameters. This result indicates that CDM-RT with

manifest attribute profile may bring unignorable biases in item parameter estimation if the mastery attempt matching assumption is not satisfied.

Table 3. Mean Absolute Bias for Item Parameter Estimates (100 Replications)

# of Examinees	Random Error %	Item Time Intensity	Main Time Effects	Two-way Interactive Effect	Three-way Interactive Effects	Item Discrimination
100	5%	0.139	0.198	0.326	0.417	0.165
	20%	0.341	0.421	0.588	0.849	0.303
	50%	0.739	0.716	1.028	0.803	0.427
500	5%	0.109	0.135	0.2	0.284	0.174
	20%	0.359	0.411	0.516	0.601	0.26
	50%	0.833	0.863	0.873	0.780	0.398
1000	5%	0.101	0.130	0.189	0.247	0.183
	20%	0.368	0.412	0.521	0.609	0.301
	50%	0.821	0.918	0.799	0.609	0.369

Simulation Study for Latent Attribute Profile

Different from CDM-RT with manifest attribute profile, latent attribute profile describes whether examinees attempt to solve the problem using the required attributes purely with the information of RTs. To explore the classification accuracy of latent attribute profiles, two different factors are considered: the number of examinees and the length of the exam. Item parameters and true attribute profiles are simulated in the same way as in the last section. Figure 1 indicates the Q-matrix. The first, second, and third conditions take the first 7, first 21, and all 35 items in the Q-matrix. The longer tests are generated by duplicating the items with the same Q-matrix as the first 7 items (3 and 5 times). For each test, the number of examinees is 100, 500, and 1000. In total, there are 9 conditions, and each one is replicated 100 times.

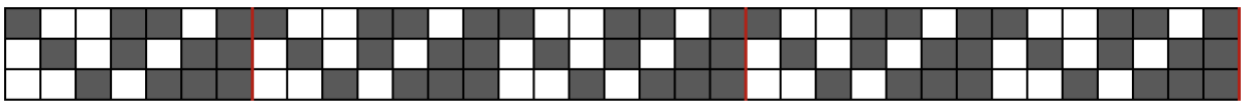

Figure 1. $K \times I$ Q-matrix

To measure the quality of latent attribute profile estimation, we use two measurements: attribute profile classification accuracy (ACA) and pattern-wise classification accuracy (PCA). As

shown in Table 4, ACAs are higher than PCA. Both ACAs and PCAs are at a high level for all conditions, which means good personal parameter recovery performance. With the same length of the exam, the ACAs and PCAs have similar classification accuracy across a different number of examinees. With the same number of examinees, we see a higher classification accuracy with more items in the exam.

Table 4. Median of Classification Rate across 100 replications

# of examinees	Length of Exam	ACAs			PCA
		α_1	α_2	α_3	
100	7	.950	.960	.940	.910
	21	1.00	.990	1.00	.990
	35	1.00	1.00	1.00	1.00
500	7	.940	.905	.970	.858
	21	1.00	.995	1.00	.995
	35	1.00	1.00	1.00	1.00
1000	7	.940	.905	.974	. 858
	21	1.00	.995	1.00	.995
	35	1.00	1.00	1.00	1.00

Real Data Analysis

In this section, a sample from PISA 2012, which includes both binary RA and RT, is used. Since PISA is a low-stake assessment, we pick the specification of CDM-RT for low-stack condition. This dataset has been used for some previous studies, which aimed at modeling RAs and RTs simultaneously (Zhan, et al. 2019). Table 5 indicates the Q-matrix used in this study. Seven attributes are assessed: change and relationships (α_1), quantity (α_2), space and shape (α_3), uncertainty and data (α_4), occupational (α_5), societal (α_6), and scientific (α_7). The first four attributes are related to mathematical content knowledge, while the last other three are mathematical context needed to place additional demands on the problem-solver (OECD, 2013; Watson & Callingham, 2003). Based on the Wald test (Ma & de la Torre, 2020), there is no modification needed for the Q-matrix. Since all ten items require two attributes (see Q-matrix in

Table 1), there are one baseline time-intensity (β_{j0}), two main attribute *stimulus* effects (non-negative; β_{jk}), and one interaction attribute effect ($\beta_{jkk'}$) for each item. In this real data analysis, we will focus on the case when attribute profile of RT is specified in a latent approach.

Table 5. Q-matrix

Item	α_1	α_2	α_3	α_4	α_5	α_6	α_7
1	0	1	0	0	1	0	0
2	1	0	0	0	1	0	0
3	1	0	0	0	1	0	0
4	0	0	1	0	0	0	1
5	0	0	1	0	0	0	1
6	0	0	1	0	0	0	1
7	0	0	1	0	0	0	1
8	0	0	0	1	0	1	0
9	0	0	0	1	0	1	0
10	0	0	0	1	0	1	0

We analyze the RTs using CDM-RT with latent attribute profiles. Table 6 shows the estimated values of item parameters. Item 5, 10, and 4 have the biggest baseline time intensity. The biggest main stimulus effect comes from the first attribution of item 5 (i.e., space and shape α_3). The interactive effect is still small. Item 5 has the biggest upper and lower bound of RT. Item 2 has the smallest lower bound of RT, and item 10 has the smallest upper bound of RT. Beyond the item parameters, we also estimate the $I \times K$ attribute profile matrix. For person parameters, 977 (69.73%) of the examinees would spend time on all attributes (i.e., attribute profile is all of 1), and 30 (2.14%) of the examinees would not spend time on any attributes (i.e., attribute profile is all of 0). The number of examinees attempt to solve the problem with first to seventh attributes are: 1014 (72.37%), 1015 (72.44%), 997 (71.16%), 990 (70.66%), 1355 (96.71%), 1310 (93.50%), and 1332 (95.07%). The first four attributes have a relatively lower value than the last three attributes. This may be because they are defined upon different sub-domains.

Table 6. Result of CDM-RT with Latent Attribute Profile.

Item	β_{j0}	β_{j1}	β_{j2}	β_{j12}	Exp (β_{j0})	Exp (β_{j1})	Exp (β_{j2})	Exp (β_{j12})	Upper Time Bound
1	3.127 (0.216)	3.747 (1.181)	0.689 (0.275)	-3.127 (1.143)	22.797	42.398	1.991	0.044	67.230
2	2.454 (0.198)	3.866 (0.875)	1.682 (0.252)	-3.381 (0.983)	**11.640**	47.774	5.377	0.034	64.825
3	3.198 (0.136)	3.989 (0.986)	0.777 (0.194)	-3.209 (0.936)	24.491	53.998	2.176	0.040	80.705
4	3.431 (0.104)	4.564 (0.608)	0.867 (0.117)	-3.934 (0.62)	**30.917**	95.931	2.379	0.020	129.247
5	3.547 (0.125)	4.576 (0.454)	1.043 (0.104)	-4.19 (0.46)	**34.714**	**97.118**	2.838	0.015	**134.685**
6	2.994 (0.096)	3.728 (0.625)	0.5 (0.09)	-3.156 (0.625)	19.959	41.612	1.649	0.043	63.263
7	2.646 (0.094)	4.382 (0.731)	1.288 (0.116)	-3.778 (0.74)	14.095	80.001	3.624	0.023	97.743
8	2.868 (0.119)	3.551 (0.561)	0.607 (0.086)	-3.063 (0.568)	17.600	34.849	1.834	0.047	54.330
9	2.742 (0.084)	4.134 (0.506)	1.162 (0.133)	-3.528 (0.547)	15.524	62.41	3.195	0.029	81.158
10	3.487 (0.462)	2.086 (1.923)	0.83 (0.42)	-0.331 (0.405)	**32.691**	8.054	2.294	0.718	**43.757**

Note: β_{j1} and β_{j2} refers to the main effect of the first and second attributed required by the item; Point estimator is based on the median of posterior samples.

Furthermore, we compared the attribute profiles estimated from CDM-RT and CDM. R package GDINA (Ma & de la Torre, 2020) is used to estimate the attribute profile based on the saturate format of G-DINA model. Figure 2 shows the histogram of UNI, RNI, and UAI across students. Most student have small UNI. For the limited number of students with UNI larger than 0.5, it is likely that these students have strong abilities on some of required attributes and is able to answer the items related to these attributes accurately and quickly. However, if UNI equals to 1, it is also likely these students have cheating behavior. We do not found students with UNI equals to 1 in this dataset. There are many students have RNI larger than 0.5, which means these students have low accuracies on more than half of required attributes and small RTs. This might indicate that these students might have low motivations. There is no student have UAI larger than 0.6, which means we do not detect student with strong evidence of high motivation.

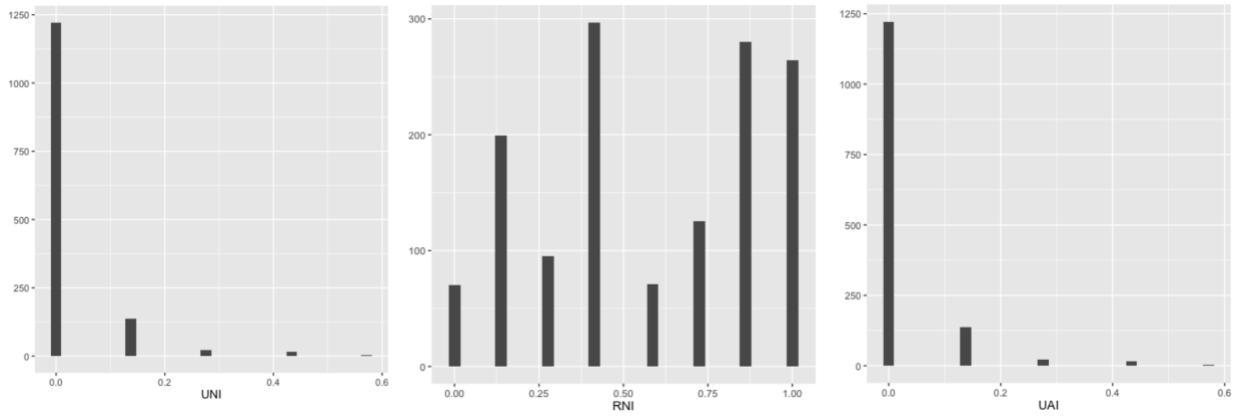

Figure 3. Histogram of three RT Behavior Indexes

Discussion

This paper proposes the general cognitive diagnosis model for RTs (CDM-RT). As a general model, CDM-RT is interpretable and flexible to chapter different model assumptions. Meanwhile, the monotonicity constraints are specified based on different motivation conditions. Positive correlations can be found for nearly any pair of examinees' response time on any kind of ability test (van der Linden, 2016). Using CDM-RT, these positive correlations could be partially explained by the similarity of the two examinees' attribute profiles. Compared with the conventional modeling of response time (i.e., log-normal model, and Box-Cox transformation models), CDM-RT explains the observed pattern of RTs by attribute combination in a multidimensional approach. In this way, CDM-RT has the potential to provide teachers with information about students' strengths and identify their instructional needs.

CDM-RT with manifest attribute profile gives different item parameter estimates from CDM-RT with latent attribute profile. Meanwhile, the estimated higher-order interaction effects from the two models sometimes are negatively correlated. In reality, we recommend estimating the attribute time effects with latent attribute profile, unless there is reliable evidence to support the use of manifest or designated attribute profile. Two treatments of attribute profiles are discussed. CDM-RT with manifest attribute profile assumes that examinees would spend time on

and only on the attributes they mater (mastery attempt matching assumption). CDM-RT with latent attributes describes whether examinees spend time on (or attempt to solve the problem using) the required attributes. Comparing the latent and manifest attribute profiles, we generate the contingency table (at attribute and entire assessment level) that describes whether examinees attempt to solve the problem with the attributes they master and do not master. Meanwhile, we designed UNI and RNI to detect the potential risk of cheating and rapid guessing. The UAI could be used to find the students with high motivation. However, these three indexes only show some signal to detect the protentional abnormal behaviors. More information is in need for a reliable conclusion.

When the number of attributes required by each item becomes larger, the number of item parameters increases exponentially. Thus, general CDM-RTs with all higher-order interactive effects may not be an appropriate choice. To estimate model fit statistically, three types of tests could be used. For the frequentist approach, one is a CDM-RT model as a baseline model tested against an alternative model that incorporates a violation of the specific assumption that is tested. Then, two asymptotically equivalent tests could be used: the likelihood ratio test or the Wald test. For the information criteria approach, Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Deviance Information Criterion (DIC) could be used to measure to what extent, a model both provide a good fit to the data but also has the capability for predicting future or different data. For the Bayesian approach, Bayesian residual analysis, prior predictive checks, and posterior predictive checks are three main tests for model selection.

Except for the validation of attribute profile treatment, the performance of the CDM-RT model strongly depends on the quality of the Q-matrix. It is possible that there are some alternative strategies for solving the same item (de la Torre & Douglas, 2008), or attributes may have some

hierarchical relationships (Gierl, 2007), or the level of attribute could be polytomous (chen & de la Torre, 2013). Meanwhile, previous studies have proposed several fully or semi-data-driven Q-matrix to improve the model identification (Liu, Xu, & Ying, 2012; DeCarlo, 2012; de la Torre & Chiu, 2015). Since the CDM-RT uses the continuous RT as the outcome variable, the conventional discussion of Q-matrix validation for categorical data needs to be extended.

The proposed framework represents an extension of the current application of cognitive diagnosis modeling of RAs and the lognormal modeling of RTs. Collaborative efforts with diverse expertise are needed in developing the assessments that can support diagnostic inferences of both RAs and RTs, selecting the appropriate psychometric tools for real data analysis, and interpreting the scores and other process information for decision making.

Reference

Chen, J., & de la Torre, J. (2013). A General Cognitive Diagnosis Model for Expert-Defined Polytomous Attributes. *Applied Psychological Measurement*, *37*(6), 419–437. https://doi.org/10.1177/0146621613479818

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, 36, 447–468.

de la Torre, J., & Chiu, C. Y. (2015). A General Method of Empirical Q-matrix Validation. *Psychometrika*, *81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. A. (2008). Model Evaluation and Multiple Strategies in Cognitive Diagnosis: An Analysis of Fraction Subtraction Data. *Psychometrika*, *73*(4), 595–624. https://doi.org/10.1007/s11336-008-9063-2

Fang, G., Liu, J., & Ying, Z. (2019). On the Identifiability of Diagnostic Classification Models. *Psychometrika*, *84*(1), 19–40. https://doi.org/10.1007/s11336-018-09658-x

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409. https://doi.org/10 (1080/01621459), 1990, 10476213

Gierl, M. J. (2007). Making Diagnostic Inferences About Cognitive Attributes Using the Rule-Space Model and Attribute Hierarchy Method. *Journal of Educational Measurement*, *44*(4), 325–340. https://doi.org/10.1111/j.1745-3984.2007.00042.x

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

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, 62, 621-640.

Lee, Y. -H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, 53(3), 359–379.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. Applied Psychological Measurement, 36, 548–564. Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187–212.

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

Minchen, N., & de la Torre, J. (2018). A general cognitive diagnosis model for continuous-response data. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 30–44. https://doi.org/10.1080/15366367.2018.1436817

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68, 589–606.

Su, Y.-S., & Yajima, M. (2015). *R2jags: Using R to run 'JAGS'*. R package version 0.6.1. Retrieved from http://CRAN.R-project.org/package=R2jags

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C.N. Mills, M. Potenza, J.J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237- 266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287–305.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.). New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing (pp. 179–203). New York: Academic Press.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31, 181–204.

van der Linden, W. J. (2016). Lognormal Response-Time Model. In *Handbook of Item Response Theory: Vol. One Model* (pp. 261–282). Taylor & Francis Group, LLC.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287-308.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton: Educational Testing Service.

Xu, G., & Zhang, S. (2015). Identifiability of Diagnostic Classification Models. *Psychometrika*, *81*(3), 625–649. https://doi.org/10.1007/s11336-015-9471-z

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, 71(2), 262– 286.

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian Cognitive Diagnosis Modeling: A Tutorial. *Journal of Educational and Behavioral Statistics*, *44*(4), 473–503. https://doi.org/10.3102/1076998619826040