A Diagnostic Classification Analysis of Problem-Solving Competence using Process Data: An Item Expansion Method*

Peida Zhan

pdzhan@gmail.com

Department of Psychology, College of Teacher Education, Zhejiang Normal University, China

Xin Qiao**

xinqiao@umd.edu

Measurement, Statistics, and Evaluation, Department of Human Development and Quantitative Methodology, University of Maryland, College Park, Maryland, United States

* This work was supported by the National Natural Science Foundation of China (Grant No. 31900795).

** Corresponding author: Xin Qiao, 1230 Benjamin Building, University of Maryland, College Park, 20742, MD, USA (e-mail: xinqiao@umd.edu).

A Diagnostic Classification Analysis of Problem-Solving Competence using Process Data: An Item Expansion Method

Abstract

Process data refers to data recorded by computer-based assessments (CBA) that reflect respondents' problem-solving processes and provide greater insight into how respondents solve problems, instead of merely how well they solve them. Using the rich information contained in process data, this study proposed an item expansion method to analyze action-level process data from the perspective of diagnostic classification in order to comprehensively understand respondents' problem-solving competence. The proposed method not only can estimate respondents' problem-solving ability along a continuum, but also can classify respondents according to their problem-solving skills. To illustrate the application and advantages of the proposed method, a Programme for International Student Assessment (PISA) problem-solving task was used. The results indicate that (a) the estimated latent classes provided more detailed diagnoses of respondents' problem-solving skills than the observed score classes; (b) although only one item was used, estimated higher-order latent ability reflected the respondents' problem-solving ability more accurately than the estimated unidimensional latent ability taken from the outcome data; and (c) the interactions between problem-solving skills may follow the conjunctive condensation rule, which assumes that only when a respondent has mastered all the required problem-solving skills can the specific action sequence appear. Overall, the main conclusion drawn from this study was that using diagnostic classification is a feasible and promising method for analyzing process data.

Keywords: process data, diagnostic classification model, problem-solving competence, cognitive diagnosis, PISA

Introduction

Computer-based assessments (CBAs) with innovative item types have been developed rapidly in recent years. Compared to traditional item types (e.g., multiple-choice items), CBA items usually require multiple decision-making steps that eventually lead to the solution of a problem. Therefore, CBA items can measure more complex cognitive processes. These items can often be seen as the problem-solving in technology-rich environments items in various large-scale international survey programs, including the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the Programme for the International Assessment of Adult Competencies (PIAAC).

Despite the appealing attributes of CBA items, a major challenge lies in the interpretation of process data obtained from CBAs. The most common type of process data obtained from CBAs is long-format and similar to longitudinal datasets in which each respondent has multiple rows of data that record his or her sequential actions and corresponding time stamps, along with identification (ID) indicators (e.g., school and person ID). Given that CBAs usually assess higher-order thinking skills and involve intricate problem-solving processes, process data recording how participants solve tasks play a crucial role in providing information about the latent construct being measured. However, it is difficult to cope with its non-standard format that respondents' response processes are action sequences (e.g., mouse clicks) with lengths varying across respondents. As a result, existing psychometric models, such as item response theory (IRT) models and diagnostic classification models (DCMs), are not readily applicable to process data. Thus, statistical methods that can draw meaningful inferences from process data are vitally necessary. To cater to this need, the current study aims to propose a novel method for the analysis process data.

An increasing number of statistical methods for analyzing process data have been proposed in the past decade. Two categories of statistical methods now exist: (a) data analysis methods that classify or explore respondents' problem-solving processes and (b) psychometric models that draw statistical inferences about respondents' latent proficiency. Data analysis methods mainly include data mining techniques (e.g., He & von Davior, 2016; Liao et al., 2019; Qiao & Jiao, 2018), social network methods (e.g., Zhu et al., 2016), diagraphs (e.g., DiCerbo et al., 2011), and process mining (e.g., Howard et al., 2010). Despite the useful applications of data analysis methods, they are exploratory in nature and cannot be used to explain latent proficiency.

By contrast, a few efforts have been made to develop psychometric models for process data (e.g., LaMar, 2018; Levy, 2014; Liu et al., 2018; Shu et al., 2017; Xu et al., 2020). A shared characteristic of these models is that they relate to both hidden Markov models (HMMs; Rabiner, 1989) and traditional measurement models (e.g., IRT models). Such a combination makes these models capable of handling the longitudinal data structures in time series data and simultaneously drawing inferences about respondents' latent proficiency. Furthermore, some joint psychometric models have been proposed to incorporate item response times, which constitute a special type of process data at item-level rather than at action-level (e.g., van der Linden, 2007; Wang & Chen, 2020; Zhan et al., 2018).

The applications of the existing modeling approaches, however, are restricted in four ways. First, multidimensionality issues cannot be adequately accommodated using current methodologies (e.g., LaMar, 2018) while multiple latent constructs are likely to exist in complex assessment forms. Second, the lack of readily available software makes some methods (e.g., Xu et al., 2020) inaccessible to practitioners. Third, some models are limited to the analysis of item-level process data (e.g., Zhan et al., 2018). Lastly, most existing models aim to estimate respondents' problem-solving abilities along a continuum but not to classify respondents based on the different problem-solving skills they adopted (e.g., Liu et al., 2018).

In the current study, we propose an item expansion method to analyze action-level process data from the perspective of diagnostic classification based on DCMs that can be estimated using readily available software programs. Therefore, its analytical process is easy to implement and practitioner-friendly. The proposed method has two main uses, (a) estimating respondents' problem-solving abilities along a continuum and (b) simultaneously classifying respondents according to their multidimensional problem-solving skills. The former, as a comprehensive indicator, can be used to determine the problem-solving levels of respondents and locate their relative positions in a group. The latter, as a fine-grained indicator, can be used to identify the problem-solving skills mastered by the respondents for diagnosis purposes.

The rest of this paper is organized as follows. In section 2 we introduce the DCM framework to acquaint readers with the analysis models used in this study. The proposed diagnostic classification method for process data is outlined in Section 3. Section 4 describes an empirical study of process data using a PISA 2012 item to demonstrate the proposed method and to show that process data contain richer diagnostic information than traditional outcome data (i.e., item responses). Finally, some concluding remarks are made and further research directions are discussed in Section 5.

Diagnostic Classification Model Framework

DCMs are a family of restricted or confirmatory latent class psychometric models that model relationships between several fine-grained discrete latent attributes and observed item responses (von Davier & Lee, 2019). A latent attribute can either correspond to a concrete knowledge point/skill or refer to a more abstract latent construct (de la Torre & Chiu, 2016). DCMs are developed to make statistical inferences about respondents' statues of these latent attributes (e.g., "mastery" versus "non-mastery", "advanced" versus "beginner", and "able to implement" versus "unable to implement") and to separate respondents into several latent classes (i.e., attribute patterns), according to their observed item responses. Because of the multidimensional nature, DCMs have the potential to yield richer diagnostic information than traditional psychometric models (Ma & de la Torre, 2020a), which is especially helpful for supporting instruction (Chen et al., 2017; Chen et al., 2018; Tang & Zhan, 2021; Wu, 2019; Zhan et al., 2019; Zhan, 2020).

A Q-matrix is an important component of DCMs that specify the relationships between items and latent attributes (Tatsuoka, 1983) and reflects the cognitive specification of a test (Leighton et al., 2004). A correct response to an item may depend on one or more latent attributes. For example, in a test that measures a total of K latent attributes, each of the I items requires a distinct subset of relevant attributes to be answered correctly. These specific item-attribute associations are collected into a binary $I \times K$ Q-matrix. The element q_{ik} indicates whether or not the item i ($i = 1, \ldots I$) requires the application of latent attribute k ($k = 1, \ldots K$) to give a correct response. Hence, for K latent attributes, 2^K attribute patterns can be constructed. Appropriately specifying the attributes required by each item is one of the necessary conditions for diagnostic classification analysis (Chen, Culpepper, Chen et al., 2018; Liu et al., 2012). Furthermore, the completeness of the Q-matrix (Chiu, 2013) and the model identifiability requirements of the Q-matrix (Gu & Xu, 2019) are necessary in the correct estimation of the model parameters.

In this paper, we used the generalized deterministic inputs, noisy "and" gate (GDINA), model (de la Torre, 2011) to illustrate the proposed analysis method. The GDINA model is a general DCM that can be used to specify a host of reduced DCMs by using different parameterizations, such as the DINA model (Junker & Sijtsma, 2001), the deterministic inputs, noisy "or" gate (DINO) model (Templin & Henson, 2006), and the additive cognitive diagnostic model (ACDM) (de la Torre, 2011), thus, increasing the generalizability of the proposed analysis method. Like most, if not all, DCMs, the

GDINA model relies on a Q-matrix to specify the associations between the *K* latent attributes and *I* items. The item response function of the GDINA model is given by:

$$P(y_{ni}=1\,|\,\boldsymbol{\alpha}_{nk}^{*})=\delta_{i0}+\sum_{k=1}^{K_i^*}\delta_{ik}\alpha_{nk}+\sum_{k'>k}^{K_i^*}\sum_{k=1}^{K_i^*-1}\delta_{ikk'}\alpha_{nk}\alpha_{nk'}+\ldots+\delta_{i(K_i^*)}\prod_{k=1}^{K_i^*}\alpha_{nk}\ , \tag{1}$$

where y_{ni} is the observed item response of respondent *n* to item *i*, $P(y_{ni}=1\,|\,\boldsymbol{\alpha}_{nk}^{*})$ is the correct response probability of respondent *n* to item *i* by given $\boldsymbol{\alpha}_{nk}^{*}$, and $\boldsymbol{\alpha}_{nk}^{*}$ denotes the *n*th reduced attribute pattern among the $2^{K_i^*}$ possible reduced attribute patterns, where $K_i^*=\sum_{k=1}^{K}q_{ik}$ denotes the number of required attributes for item *i*. The reduced attribute patterns with respect to item *i* contain only the required attributes for the item. δ_{i0} denotes the intercept of item *i*, δ_{ik} denotes the main effect of attribute α_{nk}, $\delta_{ikk'}$ denotes the two-way interaction effect of attributes α_{nk} and $\alpha_{nk'}$, and $\delta_{i(K_i^*)}$ denotes the highest-way interaction effect for all the required attributes. Among them, the intercept parameter reflects the guessing probability of the item, the main effect parameters reflect the individual contribution of each attribute to the correct response probability of the item, and the interaction effect parameters reflect the joint influence of multiple attributes on the correct response probability of the item.

As mentioned earlier, the GDINA model can be constrained to yield several reduced DCMs. For example, the DINA model can be obtained from the GDINA model by setting all the parameters, except δ_{i0} and $\delta_{i(K_i^*)}$, to zero. In such cases, the item response function of the DINA model can be expressed as:

$$P(y_{ni}=1\,|\,\boldsymbol{\alpha}_{nk}^{*})=\begin{cases}\delta_{i0}\equiv g_i & \text{if } \prod_{k=1}^{K^*}\alpha_{nk}^{q_{ik}}=0\\ \delta_{i0}+\delta_{i(K_i^*)}\equiv 1-s_i & \text{if } \prod_{k=1}^{K^*}\alpha_{nk}^{q_{ik}}=1\end{cases}, \tag{2}$$

where g_i is the guessing parameter and s_i is the slipping parameter, respectively. More details about the GDINA model and its special cases can be found in de la Torre (2011).

In practice, since the required latent attributes in a test are often conceptually related and statistically correlated, a higher-order latent structural model (de la Torre & Douglas, 2004) can be constructed to link them:

$$P(\alpha_{nk} = 1 \mid \theta_n) = \frac{\exp(\gamma_k \theta_n - \lambda_k)}{1 + \exp(\gamma_k \theta_n - \lambda_k)}, \tag{3}$$

where $P(\alpha_{nk} = 1 \mid \theta_n)$ is the marginal probability of attribute *k* for person *n* by given θ_n. θ_n is the higher-order latent ability of respondent *n*, which denotes the general problem-solving ability of respondent *n* in this study; and γ_k and λ_k are the slope and difficulty parameters for attribute *k*, respectively. The higher-order latent structure can be incorporated into many DCMs. For example, combining Equations 1 and 3 creates the higher-order GDINA (HO-GDINA) model and combining Equations 2 and 3 creates the higher-order DINA (HO-DINA) model (de la Torre & Douglas, 2004).

More importantly, a readily available GDINA package (Version 2.8.0; Ma & de la Torre, 2020b) can be utilized with R software by practitioners to carry out the analysis. Overall, based on this general DCM, we provide further theoretical extensions and practical applications of the proposed diagnostic classification analysis method in this paper.

Diagnostic Classification Analysis of Process Data: An Item Expansion Method

The DCM framework is a promising modeling approach in CBA settings where multiple problem-solving skills need to be measured. However, to our knowledge, no research has attempted to introduce the idea of diagnostic classification into action-level process data analysis (cf. Jiao et al., 2019). In this section, we provide an item expansion method for action-level process data analysis from the perspective of diagnostic classification that can be generalized to a broad scenario of CBA items.

The proposed method is an item-specific analysis method that is normally used in process data analysis given that the problem-solving scenarios differ widely among

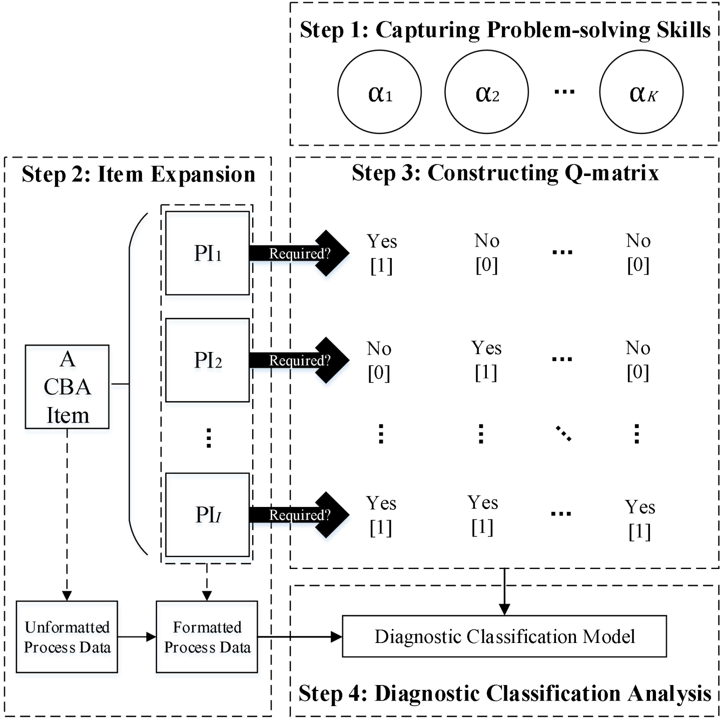
CBA items (e.g., Greiff et al., 2016; Kroehne & Goldhammer, 2018; Liu et al., 2018; Qiao & Jiao, 2018). The detailed analysis demonstration using a specific item will be presented in the next section.

Diagnostic Classification Analysis of Process Data

The DCM framework for process data analysis is portrayed as a tetrad flowchart as presented in Figure 1. The four dotted rectangles represent the key components of the proposed method: (1) the problem-solving skills that are required by a specific CBA item; (2) the phantom items, expanded from the CBA item, representing key actions or action sequences produced by respondents when completing the CBA item; (3) the Q-matrix that connects the problem-solving skills and phantom items; its entries consist of 1s and 0s, indicating whether a problem-solving skill is required by a phantom item; and (4) a DCM that leads to meaningful interpretation of the process data. The four components are described and explained in detail in the following paragraphs.

Figure 1

Illustration of Key Components in the DCM Framework for Process Data Analysis.



Note. α = latent attribute; K = number of attributes; PI = phantom item; I = number of phantom items.

Problem-Solving Skills

In this study, latent attributes refer to the problem-solving skills that are required to produce various action sequences for a CBA item. In such a case, $\alpha_{nk} = 1$ indicates that respondent n is able to implement the kth problem-solving skill, and $\alpha_{nk} = 0$ otherwise.

Typically, CBA items are designed to measure general cognitive processes involved in the problem-solving tasks. For example, the PISA 2012 problem-solving items measure four main cognitive processes: (a) exploring and understanding, (b) representing and formulating, (c) planning and executing, and (d) monitoring and reflecting. Given that individual items have one of these cognitive processes as their main focus, each main cognitive process can be further decomposed into several concrete problem-solving skills for a particular item.

In practice, we can use the test blueprint, scoring rules, and assessment framework to capture the required problem-solving skills for an item. This process is similar to that of using expert judgement to determine latent attributes in traditional diagnostic assessments (e.g., Roduta Roberts et al., 2014).

Item Expansion: From One CBA Item to Several Phantom Items

Let $\boldsymbol{A} = \{a_1, \ldots, a_A\}$ denote the set of all distinct actions for a specific CBA item (i.e., an action space), where A is the total number of distinct actions and each element in $\boldsymbol{A}$ is a unique action. Typically, these recorded A actions directly relate to item solving, while other irrelevant actions (e.g., mouse clicks on unrelated areas), if existing, are considered as noises and eliminated at the beginning of the analysis. One or more distinct actions can be further combined to form various action sequences:

$$\begin{aligned}\boldsymbol{S} = \{v_1, \ldots, v_x, \ldots, v_X\} = \{&a_1, \ldots, a_A, && \#\text{one - step action sequences} \\ &a_1a_2, \ldots, a_{A-1}a_A, && \#\text{two - steps action sequences} \\ &a_1a_2a_3, \ldots, a_{A-2}a_{A-1}a_A, && \#\text{three - steps action sequences} \\ &\ldots\}\end{aligned}, \tag{3}$$

where ν_x denotes the xth distinct action sequence and X is the total number of distinct action sequences. Since CBA items may allow respondents to repeatedly carry out certain actions, the value of X (i.e., the volume of $\boldsymbol{S}$) is usually large.

However, not all action sequences are meaningful or informative, given the noisy nature of process data. If some action sequences are conducted by too few (or too many) respondents, these rare (or excessive) action sequences are uninformative in the inference of problem-solving skills. Additionally, when an item does not examine problem-solving efficiency, the repeated appearance of an action sequence can be considered as one appearance (e.g., $a_1a_2a_1a_2 \equiv a_1a_2$). Therefore, it is necessary to filter the action sequences in $\boldsymbol{S}$ to form a set of reduced set of action sequences:

$$\boldsymbol{S}^* = \{\nu_{1^*}, \ldots, \nu_{x^*}, \ldots, \nu_{X^*}\} \subseteq \boldsymbol{S}, \tag{4}$$

where ν_{x^*} is the xth action sequence and X^* is the total number of action sequences in the reduced set of action sequences $\boldsymbol{S}^*$. In this study, the action sequences in $\boldsymbol{S}^*$ are referred to as phantom items. In the rest of this paper, the term “phantom item” is used interchangeably with the term “action sequence”. Typically, one CBA item can be expanded to several phantom items. We call this phase as item expansion.

In practice, we can use one or more approaches to obtain $\boldsymbol{S}^*$ from $\boldsymbol{S}$. Several common approaches include: (1) action sequences with an appearance rate of less than 5% or more than 95% should be excluded from the analysis (Tang et al., 2020); (2) action sequences generated from process data should theoretically relate to the latent construct being measured to achieve better classification results (Sao Pedro et al., 2012); and (3) it is important that $\boldsymbol{S}^*$ should include phantom items that meet the requirements of the Q-matrix for model parameter identifiability purposes. For example, three conditions need to be satisfied for the identifiability of the DINA model (Gu & Xu, 2019): (a) there are items that solely measure each one of the attributes, namely, an identity matrix in the Q-matrix; (b) each attribute is measured by at least three items; and (c) any column of the Q-matrix apart from the identity matrix is distinct. Once $\boldsymbol{S}^*$ is obtained, the distinct action sequences it contained are treated as phantom items for subsequent diagnostic classification analysis.

Finally, the original long-format process data for the CBA item, referred to as *unformatted process data* as shown in Figure A1 in the Appendix, can be transformed into *formatted process data* as shown in Figure A2 in the Appendix to be used in the subsequent diagnostic classification analysis. A formatted process data is a matrix with each row representing a respondent and each column representing a phantom item. Furthermore, when an action sequence appears during a respondent's problem-solving process, the corresponding element in the matrix is coded as 1, otherwise, the corresponding element in the matrix is coded as 0. In this way, a correct response to a phantom item (y_{ni} = 1) indicates that the respondent conducts an action or an action sequence. In terms of the GDINA model (Equation 1), $P(y_{ni}=1\,|\,\boldsymbol{\alpha}_{nk}^*)$ is the probability that respondent n conducts action sequence i conditional on requisite problem-solving skills $\boldsymbol{\alpha}_{nk}^*$.

Q-matrix

Once the phantom items (i.e., action sequences in $\boldsymbol{S}^*$) and problem-solving skills (i.e., latent attributes) are determined, the Q-matrix can be constructed to specify the associations among them. In this study, each column of the Q-matrix represents a problem-solving skill and each row of it represents a phantom item. In such cases, q_{ik} = 1 indicates that the ith action sequence requires kth problem-solving skill, while q_{ik} = 0 otherwise.

In addition, our approach in constructing the Q-matrix is based on expert judgement and is essentially retrofitting. That is, we aligned the phantom items to the problem-solving skills in a post hoc manner.

Diagnostic Classification Analysis

Existing reduced DCMs (e.g., DINA and DINO models) can be divided into three categories, based on the condensation rules for how latent attributes influence respondents' item responses: conjunctive, disjunctive, and compensatory models (Maris, 1999). The selection of reduced DCMs requires a clear understanding of the theoretical interactions between the latent attributes, which are usually judged by experts during the item/test development phase. However, theoretical interactions between problem-solving skills are not very clear in the analysis of process data. In such a case, it is suggested to use a general DCM that is not limited to a specific condensation rule, although multiple DCMs can be used for simultaneous analysis using model–data fit indices to select the most appropriate one (Chen et al., 2013). In addition, higher-order DCMs are suggested in this study. The higher-order DCMs can be used to estimate the problem-solving ability of a respondent along a continuum and to diagnose the problem-solving skills used by this respondent simultaneously. Parameter estimation for higher-order DCMs can be carried out using the readily available GDINA package.

Procedure

Based on the above discussion, the diagnostic classification analysis of process data consists of the following steps:

1. Collect and preprocess the unformatted process data from the raw log files.
2. Define the problem-solving skills (i.e., latent attributes), including their number and meaning.
3. Create the phantom items (i.e., action sequences in $\boldsymbol{S}^*$).
4. Code the formatted process data matrix from the unformatted process data.

5. Construct the Q-matrix (i.e., relate each phantom item to different problem-solving skills).
6. Select an appropriate DCM and analyze the formatted process data.
7. Interpret the diagnostic results.

It can be seen that the analytical procedure for process data is in general consistent with that for outcome data in traditional diagnostic assessments. An empirical data analysis is presented in the next section to demonstrate the analytical procedure described above.

Empirical Study

Item Description

PISA 2012 contains 48 problem-solving items, which assess respondents' cognitive ability in solving real-life problems using computer-based simulated scenarios. One of these items, TICKETS task 2 (CP038Q01),[1] was used in the current study. Figure 2 shows the opening page of this item. This item asks respondents to buy the cheapest ticket that allows them to travel around the city four times on the subway using the virtual ticket machine. It also informs the respondents that concession fares are available for respondents.

Several decisions need to be made to solve this item. First, a respondent must choose a train network (either "city subway" or "country trains", as shown in Figure 2). Then, according to the train network chosen, the respondent chooses between "full fare" and "concession fare", as shown in Figure A3 in the Appendix. Last, after a fare type is chosen, the respondent must buy either a "daily" or "individual" ticket, as shown in Figure A4 in the Appendix. Figures A5 and A6 in the Appendix show the screenshots when a "daily" ticket or an "individual" ticket is chosen, respectively.

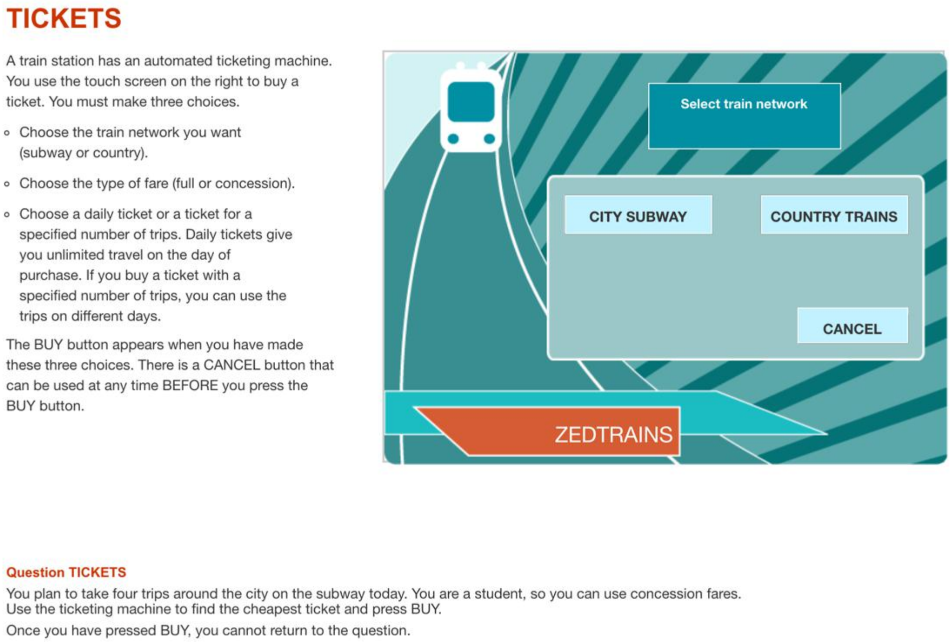
[1] More details about this item can be found in https://www.oecd.org/pisa/test-2012/testquestions/question5/ retrieved on July 11th, 2020.

Specifically, the respondent can choose to buy one to five individual tickets if the "individual" ticket type is chosen. The prices of tickets are shown on the screen for comparison purposes. The respondent can click on "Cancel" at any of the above steps to restart the item from the beginning. Both a daily city subway concession fare and four individual city subway concession fares allow the respondent to travel around the city in a day, but the latter option is cheaper; therefore, this item requires respondents to compare the prices between the two and choose to buy four individual city subway concession fares.

According to the scoring rule, the outcome data for this item has three score categories: 0, 1, and 2. A respondent who buys four individual city subway concession fares and compares the price with the daily concession fare receives a full score of 2. A respondent who buys either four individual city subway concession fares *or* a daily city subway concession fare without comparing the prices receives a partial credit of 1. A respondent who makes any other decisions receives a score of 0.

Figure 2

Opening Page for the TICKETS Task 2 (CP038Q01) in PISA 2012.



TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.

Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Participants and Datasets

Respondents from the United States, Singapore, Austria, and Turkey were selected from *CBA_cp038q01_logs12_SPSS.csv* (see online supporting materials) for the analysis in the current study, constituting a representative sample of people with a wide spectrum of problem-solving ability based on their PISA 2012 results (OECD, 2014). Both unformatted process data and outcome data (i.e., item responses) from these respondents were used in the current study (see online supporting materials). Respondents with missing IDs and not-reached item responses were excluded, resulting in a sample size of 3,760. As mentioned earlier, unformatted process data recorded the sequential actions and corresponding time stamps for the respondents. No missing data existed in the process data, while missing data in the outcome dataset were accommodated by the full information maximum likelihood (FIML) estimation.

In this paper, the observed item response categories from CP038Q01 (i.e., 0, 1, 2) served as observed classes that were compared with the latent classes estimated by the DCMs. However, the latent classes indicated different problem-solving skill patterns, which were expected to have more categories than that of the observed classes and were perceived as more fine-grained classifications of the respondents. The detailed analysis is described in the following section.

Analysis

To increase the replicability of this study, all relevant data and analysis code used in this study are available at https://osf.io/nwtfz. By running the code, readers can reproduce the analysis and obtain the same results. The analysis presented below serves as a representative example of the DCM framework for process data analysis as illustrated as the flowchart in Figure 1. Specifically, the construction of the problem-solving skills and phantom items, the construction of the Q-matrix, and the procedure of the diagnostic classification analysis are described in detail.

Capturing Problem-solving Skills

Based on the item instructions, scoring rules, and the targeted cognitive process (i.e., exploring and understanding) for this specific PISA problem-solving item (OECD, 2014), five problem-solving skills are required for respondents to respond correctly: (a) understanding the city subway and the correct train network, (b) understanding that concession fares are available, (c) understanding that either a daily or four individual trip tickets would allow them to travel four times around the city, (d) comparing the two ticket prices to find the cheapest, and (e) making the decision to buy. These five problem-solving skills are denoted as α_1, α_2, α_3, α_4, and α_5 in the remainder of the paper. In such cases, the higher-order latent ability in higher-order DCMs represents a respondent's problem-solving ability regarding the targeted cognitive process (i.e., exploring and understanding).

Obtaining Phantom Items and Formatted Process Data Coding

The major task in obtaining the phantom items is to get the reduced action sequence space $\boldsymbol{S}^*$ from the full action sequence space $\boldsymbol{S}$. The action space $\boldsymbol{A}$ for the phantom item used in the current study included 13 actions: city subway (city, for short), country train, concession (con, for short), full fare, daily, individual (ind, for short), trip 1, trip 2, trip 3, trip 4, trip 5, cancel, and buy. The $\boldsymbol{S}$ was generated by adding combined adjacent actions to $\boldsymbol{A}$, with action/action sequence lengths ranging from one to five (all X = 228 action sequences in $\boldsymbol{S}$ are presented in *allactionsequence.csv* and can be generated by running the shared code). Given that the PISA item we used does not assess problem-solving efficiency, repeated actions/action sequences from one respondent were considered to be the same as actions/action sequences that have appeared only once. Hence, we dichotomized all repeated actions (i.e., all non-zero count data was modified to 1).

We determined the $\boldsymbol{S}^*$ based on appearance rate (between 5% to 95%), theory, and the requirements for the identifiability of the DCMs. First, rare action sequences in $\boldsymbol{S}$

with appearance rate less than 5% were removed, resulting in 60 remaining action sequences. No excessive action sequences (i.e., appearance rate > 95%) existed in the current example. Second, we retained phantom items that could theoretically reflect the problem-solving skills defined in the previous paragraph, according to Sao Pedro et al. (2012). Therefore, some action sequences in S that contain actions that do not reflect the required problem-solving skills (e.g., country train and full fare) were removed, resulting in 14 action/action sequences. Lastly, the 14 action/action sequences were examined to have satisfied the identifiability conditions of the DCM (Gu & Xu, 2019), as mentioned in the previous section. Therefore, a set of reduced set of action sequences S^*, including 14 phantom items, were retained to measure the five problem-solving skills.

Several issues we encountered in the above procedure specifically for this PISA item were addressed. First, when respondents intended to buy tickets for four individual trips (as allowed by the item), they could repeatedly click on other numbers of trips (i.e., one, two, three, or five) to check the prices before choosing four trips. Given that this item does not assess problem-solving efficiency, we considered these less efficient action sequences to be the same as clicking “trip 4” immediately after “individual” ticket. Therefore, phantom items “ind→other→trip 4”, where “other” indicated clicking on a number of trips other than four, and “ind→trip 4” were considered as the same. Second, according to the scoring rule for this item, we considered both choices of “daily” ticket and “trip 4” (i.e., “individual” ticket with four trips) as a reflection of the mastery of α_3 , given that both of them allowed the respondents to travel around the city four times. Therefore, phantom items such as “daily/trip4_buy”, “con_daily/trip4_buy”, and “city_con_daily/trip4_buy” were generated, with respondents who bought either a daily or four individual tickets receiving correct scores for these phantom items.

As a result, we coded the formatted process data matrix for 3,760 respondents who responded to the 14 phantom items. As mentioned earlier, all responses to the phantom items were dichotomous.

Constructing Q-matrix

Based on the five problem-solving skills and 14 phantom items, the Q-matrix was constructed as shown in Table 1. In the current study, we assumed there was no attribute hierarchy (Leighton et al., 2004) among the five problem-solving skills. For example, respondents who understood that concession fares were available (α_2) may not also understand the city subway and the correct train network (α_1). Even if respondents made a buy decision, it does not mean that they mastered one or more of the first four problem-solving skills. For example, the appearance of an action sequence "country train→full fare→ind→trip 2→buy" only requires α_5.

In addition, all phantom item responses (i.e., action sequence appearances) are assumed to be conditional independent given requisite latent attributes (i.e., problem-solving skills). For example, the response to phantom item "city" given attribute α_1 is assumed to be conditional independent from the response to phantom item "city→con" given attributes α_1 and α_2, as shown in Table 1. In terms of the GDINA model, for respondent n, $y_{n\text{“city”}} \equiv y_{n1}$ given α_{n1} is assumed to be conditional independent of $y_{n\text{“city→con”}} \equiv y_{n6}$ given α_{n1} and α_{n2}. The local independence between different phantom items can also be reflected by (item-level) model-data fitting as shown in the results below.

Table 1

Q-matrix Created for PISA 2012 Problem-Solving Item TICKETS Task 2.

Item Number	Phantom Items	Problem-solving Skills				
		α_1	α_2	α_3	α_4	α_5
1	city	1	0	0	0	0
2	con	0	1	0	0	0
3	daily/trip4	0	0	1	0	0
4	cancel	0	0	0	1	0
5	buy	0	0	0	0	1
6	city→con	1	1	0	0	0
7	con→daily/trip4	0	1	1	0	0
8	city→con→daily/trip4	1	1	1	0	0
9	city→con→daily→cancel	1	1	1	1	0
10	daily→cancel	0	0	1	1	0
11	con→daily→cancel	0	1	1	1	0
12	daily/trip4→buy	0	0	1	0	1
13	con→daily/trip4→buy	0	1	1	0	1
14	city→con→daily/trip4→buy	1	1	1	0	1

Note: city = city subway, con = concession, ind = individual, other = number of individual trips other than four, trip4 = four individual trips, α_1 = understanding the city subway and the correct train network, α_2 = understanding that concession fares were available, α_3 = understanding that either a daily or four individual tickets allowed them to travel four times around the city, α_4 = comparing the two ticket prices to find the cheapest, α_5 = making a decision to buy.

Diagnostic Classification Analysis

The GDINA package was used to conduct diagnostic classification analysis on the formatted process data. First, the test- and item-level model-data fit were examined to determine whether the HO-GDINA model (Equations 1 and 3) fitted the data. For the test-level model-data fit evaluation, the M_2 root mean squared error of approximation (RMSEA$_2$) and the standardized root mean squared residual (SRMSR) (Maydeu-Olivares, 2013) were calculated by the GDINA package. Specifically, RMSEA$_2$ ≤ 0.05 and SRMSR ≤ 0.05 indicated satisfactory approximate and absolute model fit (Maydeu-

Olivares, 2013). For the item-level absolute model-data fit evaluation, the Fisher-transformed correlation of item pairs (Chen et al., 2013) was used and the adjusted *p*-value > 0.05 indicated adequate fit.

Typically, a misspecified Q-matrix may also cause model-data misfit in addition to misfitted items. In this study, the proportion of variance accounted for (PVAF) method (de la Torre & Chiu, 2016) was used for the Q-matrix validation. In addition, to explore the empirical interactions between the five problem-solving skills, three reduced models with different condensation rules were also fitted to the data: the HO-DINA model with a conjunctive rule, the HO-DINO model with a disjunctive rule, and the HO-ACDM with a compensatory rule. The relative model-data fit indices were computed for each model to evaluate the relative model-data fit, including the Akaika information criterion (AIC; Akaike, 1981), the Bayesian information criterion (BIC; Schwarz, 1978), the consistent AIC (CAIC; Bozdogan, 1987), and the sample size adjusted BIC (SABIC; Sclove, 1987). Likelihood ratio tests were also conducted given that the three reduced models were nested within the HO-GDINA model.

Outcome Data Analysis

For the outcome data analysis, the partial credit model (PCM; Masters, 1982) was fitted to the dataset using the full information maximum likelihood (FIML) with the expectation-maximization (EM) computation algorithm (Bock & Aitkin, 1981) using the TAM R package (Version 3.5-19; Robitzsch et al., 2020). We compared the estimated higher-order latent ability (denoted as θ_1) from the best fitting DCM and the estimated unidimensional latent ability (denoted as θ_2) from the PCM to explore their consistency in reflecting respondents' problem-solving abilities. Ability estimates from both models were expected a posteriori (EAP) estimates. The correlation coefficient were computed. Since there were several phantom items in the process data, but only one item in the outcome data, it was expected that the standard errors of θ_1s would generally be smaller than those of θ_2s. In other words, the best fitting DCM could provide more accurate problem-solving ability estimates than the PCM.

Results

Diagnostic Classification Analysis Results

The HO-GDINA model was first fitted to the formatted process data with the original Q-matrix (see Table 1). The test-level model-data fit indices are presented in Table 2. The RMSEA$_2$ and SRMSR suggested adequate fit. Additionally, the model-data fit became worse when using the revised Q-matrix based on PVAF. Combined with expert judgement (e.g., Ravand, 2016), the original Q-matrix was used for the next step.

Table 2

Summary of Test-level Absolute Model–data Fit.

Q-matrix	RMSEA$_2$ [90% CI]	SRMSR
Original	0.032 [0.025, 0.041]	0.033
Revised	0.085 [0.079, 0.091]	0.017

Note: Full = 14 phantom items, Reduced = 13 phantom items, M_2 = M_2 statistic, *df* = degree of freedom, RMSEA$_2$ = M_2 root mean squared error of approximation, CI = confidence interval, SRMSR = standardized root mean squared residual.

Item fit was further examined based on the HO-GDINA model. According to the heatmap for the adjusted *p*-values of the transformed correlation presented in Figure A7 in the Appendix, the fifth phantom item (i.e., "buy") was a misfit possibly because only 7% respondents did not conduct this action. However, the phantom item "buy" was necessary for the identifiability of the DINA model (i.e., Q-matrix contained an identity matrix) and was retained for subsequent analyses.

Table 3 presents the relative model-data fit indices for four models. Specifically, AIC, BIC, CAIC, and SABIC all suggested the HO-DINA model to be the best fitting model, and the likelihood ratio test showed that there was no significant difference between the HO-DINA model and the HO-GDINA model. Therefore, the more parsimonious HO-DINA model was used in subsequence analyses. Such results indicate that the interactions between problem-solving skills followed the conjunctive condensation rule: only when the respondent had mastered all the required problem-solving skills did the specific action sequence appear.

Table 3

Summary of Relative Model-data Fit.

Model	#par	–2LL	AIC	BIC	CAIC	SABIC	χ^2	df	*p*-value
HO-GDINA	92	16,951.31	17,135.31	17,708.67	17,800.67	17,416.33			
HO-DINA	38	17,014.83	17,090.83	17,327.65	17,365.65	17,206.91	63.52	54	0.180
HO-DINO	38	31,304.61	31,380.61	31,617.44	31,655.44	31,496.69	14,353.31	54	<0.001
HO-ACDM	54	20,179.61	20,287.61	20,624.15	20,678.15	20,452.57	3,228.31	38	<0.001

Note: HO-GDINA = higher-order generalized deterministic-inputs, noisy "and" gate model; HO-DINA = higher-order deterministic-inputs, noisy "and" gate model; HO-DINO = higher-order deterministic-inputs, noisy "or" gate model; HO-ACDM = higher-order additive cognitive diagnostic model; #par = number of estimated parameters; –2LL = –2 log-likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = Consistent AIC; SABIC = sample-size adjusted BIC; *df* = degree of freedom; in the likelihood ratio tests, models were tested against the HO-GDINA model.

Table 4 presents the item parameter estimates based on the HO-DINA model. The guessing parameter g_i indicates the probability of performing action *i* for respondents who did not master the problem-solving skills required for action *i*, while the slipping parameter s_i indicates the probability of not performing action *i* for respondents who have mastered the problem-solving skills required for action *i*. All item-level aberrant responses probabilities were quite small except for phantom item "buy", which means that most items had a high discrimination parameter (i.e., IDI$_i$). One possible reason for the relatively high guessing probability of item 5 is that "to buy a ticket" was written on the item description, thus, this action did not require much cognitive processes. Overall speaking, high quality of the majority of the phantom items provide the necessary guarantee for the high accuracy of diagnostic classification (de la Torre et al., 2010).

Table 4
Phantom Item Parameter Estimates based on the HO-DINA Model.

Item Number	Phantom Item	g_{i}	s_{i}	IDI_i
1	city	0.054 (0.030)	0.000 (0.045)	0.946
2	con	0.021 (0.012)	0.000 (0.090)	0.979
3	daily/trip4	0.025 (0.010)	0.000 (0.090)	0.975
4	cancel	0.107 (0.008)	0.000 (0.181)	0.893
5	buy	0.442 (0.029)	0.000 (0.026)	0.558
6	city→con	0.016 (0.012)	0.000 (0.121)	0.984
7	con→daily/trip4	0.000 (0.015)	0.000 (0.170)	1.000
8	city→con→daily/trip4	0.000 (0.171)	0.001 (0.002)	0.999
9	city→con→daily→cancel	0.000 (0.358)	0.021 (0.004)	0.979
10	daily→cancel	0.015 (0.004)	0.000 (0.166)	0.985
11	con→daily→cancel	0.000 (0.354)	0.000 (0.009)	1.000
12	dail/trip4→buy	0.023 (0.007)	0.000 (0.043)	0.977
13	con→daily/trip4→buy	0.000 (0.192)	0.000 (0.002)	1.000
14	city→con→daily/trip4→buy	0.000 (0.206)	0.009 (0.002)	0.991

Note: HO-DINA = higher-order deterministic-inputs, noisy "and" gate model; g_i = guessing parameter; s_i = slipping parameter; IDI_i = phantom item discrimination index, which equals to $1 – s_i – g_i$; standard errors in parenthesis.

Respondents' latent classes, estimated by the HO-DINA model, were compared to their observed item scores. As expected, there were multiple latent classes for each observed score category. The number of respondents in the three observed score categories (i.e., 2, 1, 0) in each latent class is shown in Table 5. The number of respondents who received scores of 2, 1, and 0 were 1,093, 1,637, and 1,030, respectively. By contrast, the theoretical number of all possible latent attribute patterns was $2^5 = 32$, given 5 latent attributes. Although only 26 latent attribute patterns were diagnosed in this study, the number of latent classes was still larger than that of the observed score category.

Table 5 shows the distribution of respondents and their latent attribute patterns with respect to their observed score categories. Respondents who received a full credit of 2 were estimated to have mastered all the problem-solving skills (i.e., pattern = 11111, *n* = 1,093), which was consistent with their observed score category.

For the respondents who received a partial credit of 1, the majority of them were estimated to have mastered all the problem-solving skills except α_4 (i.e., pattern = 11101, *n* = 1,481), which was supposed to be the latent attribute pattern for respondents who received a partial credit; that is, they made the correct decision to buy a ticket that allowed them to travel around the city four times, but did not compare the prices between tickets. A small proportion of respondents were estimated to have mastered all the problem-solving skills (i.e., pattern = 11111, *n* = 156). By taking a closer look at their actual sequential actions, we found that these respondents may have compared the prices of a daily ticket and four individual trips, but still decided to buy the daily ticket, which was more expensive.

Respondents who scored 0 (*n* = 1,030) were expected to have latent attribute patterns other than 11111 (i.e., the pattern for the full credit respondents) and 11101 (i.e., the pattern for the partial credit respondents). This meant that, if they scored 0, respondents did not understand what available tickets would allow them to travel around the city in four trips and did not compare the prices of these options. The main advantage of diagnostic classification analysis was that it can diagnose respondents' erroneous problem-solving skill patterns. For example, 72 respondents were classified as latent attribute pattern 01001, meaning that these respondents only understood the concession fares (α_2) and made the decision to buy (α_5), but did not master other problem-solving skills. One possible problem-solving process was that respondents could choose to buy a single individual concession ticket for the country train, which did not allow them to have four trips around the city. Furthermore, 157 respondents were estimated to have the latent attribute pattern 11001, which indicated that they did not understand that the ticket must allow them to travel four times around the city and did not compare the prices of possible options. One such answer was that a respondent bought two individual concession fares for the city subway, which also did not allow them to have four trips around the city. However, respondents with pattern 11001 mastered more problem-solving skills than respondents with pattern 01001, despite that they all received 0 scores. This indicated that using the diagnostic classification analysis method proposed in the current study provided a more fine-grained diagnosis of the error types in respondents' problem-solving processes than do the observed scores. Additionally, although 3 respondents were estimated to have latent profile 11101, and 29 respondents were estimated to have latent profile 11111 in the current study, such small amount of misclassification may have been due to estimation error, given that only 14 phantom items were used in the diagnostic classification analysis.

Table 5

Distribution of Respondents' Attribute Patterns with Respect to Their Observed Score Category.

Observed Score Category	Latent Attribute Pattern	Frequency
2	11111	1,093
1	11111	156
	11101	1,481
0	00000	87
	00001	22
	00100	10
	00101	39
	00110	2
	00111	1
	01000	6
	01001	72
	01100	21
	01101	88
	01110	13
	01111	2
	10000	9
	10001	29
	10100	10
	10101	107
	10110	2
	10111	6

11000	19
11001	157
11010	6
11011	35
11100	98
11101	3
11110	157
11111	29

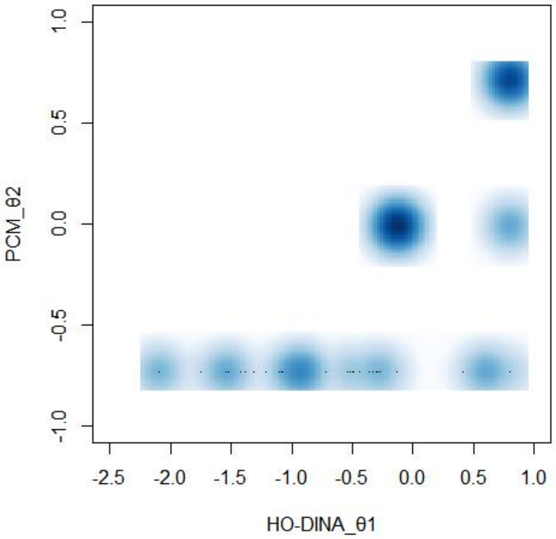
In addition, some supplementary diagnostic classification analysis results are reported in the Appendix, including attribute mastery proportions (Table S1), attribute pattern mixing proportions (Table S2), and attribute correlations (Table S3). Furthermore, a sensitivity analysis was conducted to examine the impact of different choices of phantom items in the construction of Q-matrix and formatted process data. Specifically, two Q-matrices and formatted process datasets that included less or more items than that used in the current study were constructed. More details can be found in the section S1 in the Appendix. The results indicate that the classification results remained similar to the original analysis in terms of the number of classes and the number of respondents in each class.

Outcome Data Analysis Results

For the problem-solving ability estimate, the correlation coefficient between θ_1s and θ_2s was 0.826 (p < 0.001). Such a significantly high positive correlation indicated that θ_1 and θ_2 were likely to measure the same latent construct. Additionally, Figure 3 displays the smooth scatter plot for the problem-solving ability estimates using the HO-DINA model and the PCM. It can be seen that the PCM estimated respondents' problem-solving ability as one of three values according to their item responses: scores of 0, 1, and 2 were estimated to be -0.733, -0.012, and 0.709, respectively. By contrast, the HO-DINA model can clearly showed the differences between respondents' problem-solving abilities, especially for respondents who scored 0. Furthermore, for the estimation standard error, the results of the dependent *t*-test with a one-tailed alternative hypothesis showed that the average of the standard errors for θ_1s were significantly less than those for θ_2s ($t(3759) = -115.58$, $p < 0.001$).

Figure 3

Smooth Scatter Plot for Problem-solving Ability Estimates using the HO-DINA model and the PCM.



Note. Darker color indicates more respondents.

Reliability and Validity

When new methods are used to analyze existing data, the reliability and validity of the analysis results should also be considered. In this study, the classification accuracy index (Wang et al., 2015) was used for evaluating the reliability of classification results. In addition, validity evidence was provided in the interpretation of the problem-solving abilities and the problem-solving skill patterns obtained using the proposed method. In sum, the proposed method can be used to assess problem-solving competence through the process data analysis with adequate reliability and validity. More details can be found in the section S2 in the Appendix.

Conclusion and Discussion

To comprehensively understand respondents' problem-solving competence through the rich information contained in process data, this study proposed an item expansion method from the perspective of diagnostic classification based on DCMs. In the proposed method, action sequences for a specific CBA item were treated as phantom items, and the problem-solving skills required to produce these action sequences were treated as latent attributes. In such cases, original unformatted process data can be transformed into formatted process data, and an additional Q-matrix can be used in the application of DCMs. By incorporating the idea of diagnostic classification into process data analysis, the proposed item-specific method cannot only estimate respondents' problem-solving ability along a continuum, but also classify respondents according to their problem-solving skills. More importantly, the data analysis in the proposed method can easily be handled by readily available software, which is very practitioner-friendly.

To illustrate the application and advantages of the proposed method, a PISA 2012 problem-solving item (i.e., TICKETS task 2 (CP038Q01)), was used in this study. The results indicated that (a) the estimated latent classes provided more detailed diagnoses of respondents' problem-solving skills than observed score classes; (b) although only one item was used, the estimated higher-order latent ability could reflect respondents' problem-solving ability more accurately than the estimated unidimensional latent ability from the outcome data; and (c) the interactions between problem-solving skills appeared to follow the conjunctive condensation rule. Furthermore, reliability of classification was reported and validity evidence was provided for the problem-solving ability and the problem-solving skill patterns, respectively. Overall, the main conclusion drawn from this study was that using DCMs is a feasible and promising method for analyzing process data and problem-solving competency measured by CBA items.

Despite the promising findings, further studies are still needed. First, as an item-specific method, the proposed method requires a deep understanding of respondents' cognitive processes during problem-solving. A considerable amount of expert judgement is required for the data preparation and Q-matrix construction. Thus, it may be 'expensive' for applications in a large number of items (Tang et al., 2020). In future research, we can try to integrate more generic automated methods (e.g., He & von Davier, 2016; Tang et al., 2020) into the proposed method. Second, although the results of this study have shown that even a single CBA item can be used to conduct a detailed diagnostic analysis of problem-solving competence using the proposed method, a means to carry out simultaneous analysis of multiple CBA items is still worth exploring in the future. Third, this study has shown that, through item expansion, the applications of DCMs can be extended from the item-level outcome data analysis to the action-level process data analysis. Similarly, through item expansion, the applications of other psychometric models (e.g., IRT models) can also be explored in the analysis of process data, which deserves future research. Fourth, process data often includes time stamps of the actions in addition to actions. The proposed method does not make use of this information. Thus, future studies are needed to incorporate time information into the current method. Fifth, given that the item analyzed in this study did not examine problem-solving efficiency, we dichotomized all recurring actions. In future studies, information about recurring actions can be considered in relation to problem-solving efficiency. Sixth, background variables (Liao et al., 2019) were not considered in the current method and future studies could explore the utilities of such auxiliary information. Lastly, the CBA item (CP038Q01) used in this study only required one problem-solving strategy that led to the correct response. In fact, by properly setting the Q-matrix, the item expansion method can also be used for CBA items with multiple problem-solving strategies. For example, we can first capture different problem-solving skill patterns for different strategies and then retain action sequences required by different strategies in ***S****. The feasibility of the method needs to be validated using CBA items allowing multiple problem-solving strategies in the future.

In sum, the work presented in this paper is an initial attempt to analyze action-level process data from the perspective of diagnostic classification via an item expansion method. It shows the great potential of using DCMs to analyze process data which can provide more fine-grained diagnostic information on the respondents than using outcome data alone.

References

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of econometrics, 16*(1), 3–14.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. https://doi.org/10.1007/BF02294361

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika, 83*(1), 89–108.

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement, 42*, 1, 5–23.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Recommendation system for adaptive learning. *Applied Psychological Measurement, 42*(1), 24–41. doi:10.1177/0146621617697959

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618. https://doi.org/10.1177/0146621613488436

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). 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. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*(2), 227-249.

DiCerbo, K. E., Liu, J., Rutstein, D. W., Choi, Y., and Behrens, J. T. (2011). "Visual analysis of sequential log data from complex performance assessments," [Paper presentation]. Annual meeting of the American Educational Research Association, New Orleans, LA, United States.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Gu, Y., & Xu, G. (2019). *Partial identifiability of restricted latent class models*. arXiv preprint. https://arxiv.org/abs/1803.04353v1

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (Vol. 2, pp. 749–776). Hershey: Information Science Reference. http://dx.doi.org/10.4018/978-1-4666-9441-5.ch029

Howard, L., Johnson, J., and Neitzel, C. (2010). “Examining learner control in a structured inquiry cycle using process mining.” in *Proceedings of the 3rd International Conference on Educational Data Mining*, 71–80. Available online at: https://files.eric.ed.gov/fulltext/ED538834.pdf#page=83 (Accessed August 26, 2018).

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Jiao, H., Liao, D., & Zhan, P. (2019). *Utilizing process data for cognitive diagnosis*. In von Davier, M., & Lee, Y.-S. (Eds.). Handbook of Diagnostic Classification Models. New York: Springer.

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika, 83*(1), 67–88. https://doi.org/10.1007/s11336-017-9570-0

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CRESST Report No.837). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved August 26, 2018, from https://files.eric.ed.gov/fulltext/ED555714.pdf

Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of us adults’ employment status in PIAAC. *Frontiers in Psychology, 10*: 646.

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

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology. 9*: 1372. https://doi.org/10.3389/fpsyg.2018.01372

Ma, W., & de la Torre, J. (2020a). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement Interdisciplinary Research and Perspectives,* 18(2), 87–96. https://doi.org/10.1080/15366367.2019.1697122

Ma, W. & de la Torre, J. (2020b). GDINA: An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software, 93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Masters, G. N. (1982). A Rasch model for partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–174.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 11*(3), 71–101.

OECD (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. PISA, OECD Publishing: http://doi.org/10.1787/9789264208070-en

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: a didactic. *Frontiers in psychology*, *9*: 2231.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE, 77*(2), 257–285.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782–799. https://doi.org/10.1177/0734282915623053

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis modules. [R package version 3.5-19]. Retrieved from http://CRAN.R-project.org/package=TAM

Roduta Roberts, M., Alves, C. B., Chu, M.-W., Thompson, M., Bahry, L. M., & Gotzmann, A. (2014). Testing expert-based versus student-based cognitive models for a grade 3 diagnostic mathematics assessment. *Applied Measurement in Education, 27*(3), 173–195. https://doi.org/10.1080/08957347.2014.905787

Sao Pedro, M. A., Baker, R. S. J. d., & Gobert, J. D. (2012). *Improving construct validity yields better models of systematic inquiry, even with less information*. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), User modeling, adaptation, and personalization: Proceedings of the 20th UMAP conference (pp. 249–260). Berlin Heidelberg, Germany: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling, 59*(1), 109–131.

Tang, F., & Zhan, P. (2021). *Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment*. ResearchGate Preprint. https://doi.org/10.13140/RG.2.2.23511.19365

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2020). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1-33. https://doi.org/10.1111/bmsp.12203

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72(3),* 287–308. http://dx.doi.org/10.1007/s11336-006-1478-z

von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages.* New York, NY, USA: Springer.

Wang, S & Chen, Y. (2020). Using response times and response accuracy to measure fluency within cognitive diagnosis models, *Psychometrika, 85*, 600–629.

Wu, H.-M. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educational Psychology, 39*(10), 1218–1232. https://doi.org/10.1080/01443410.2018.1494819

Xu, H., Fang, G., & Ying, Z. (2020). A latent topic model with Markovian transition for process data. *British Journal of Mathematical and Statistical Psychology.* Advanced Online. URL https://doi.org/10.1111/bmsp.12197

Zhan, P. (2020). A Markov estimation strategy for longitudinal learning diagnosis: Providing timely diagnostic feedback. *Educational and Psychological Measurement, 80(6)*, 1145–1167.

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., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics, 44*(3), 251-281.

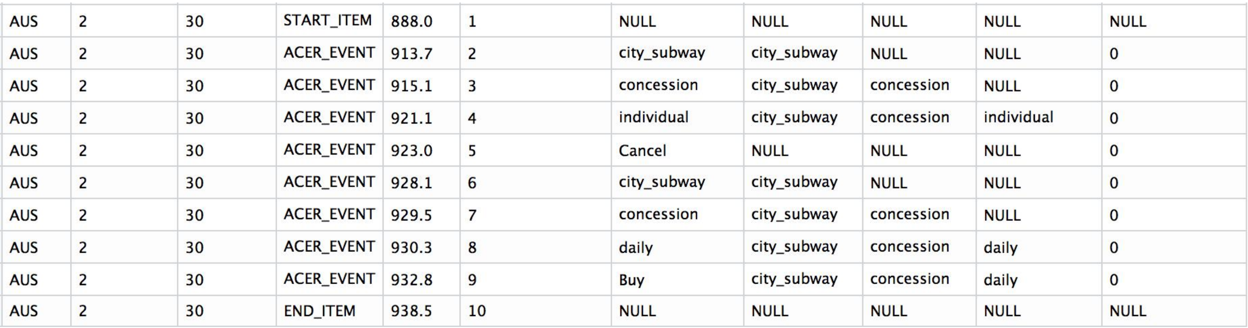
Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement*, *53*(2), 190–211.

Online Appendix

A Diagnostic Classification Analysis of Problem-Solving Competence using Process Data: An Item Expansion Method

Figure A1

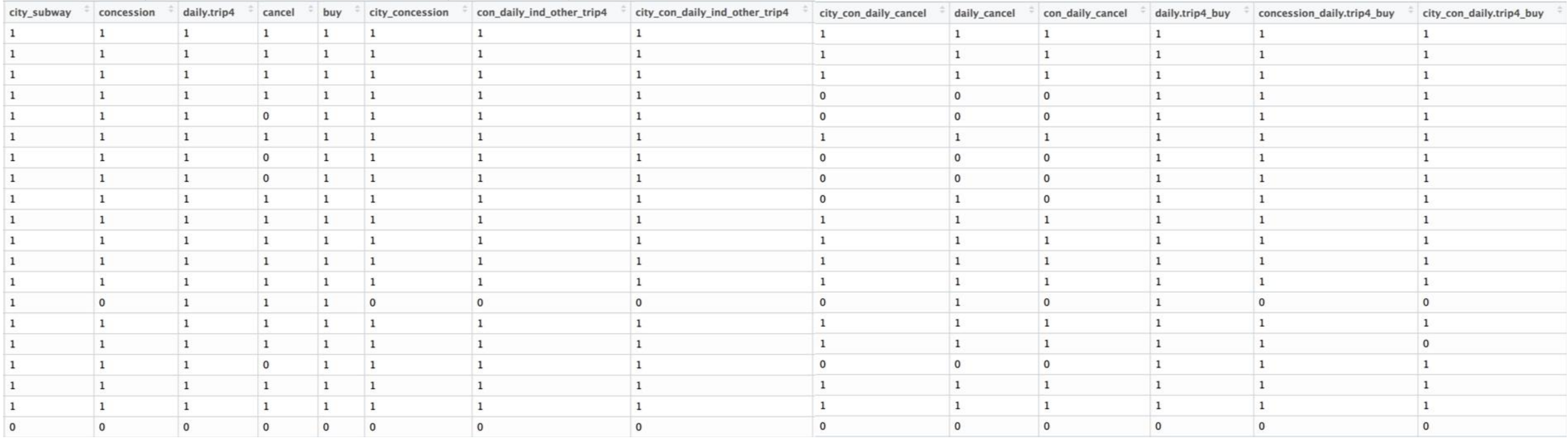
Screenshot of Unformatted Process Data for One Student From the TICKETS Task2 (CP038Q01) in PISA 2012.



AUS	2	30	START_ITEM	888.0	1	NULL	NULL	NULL	NULL	NULL
AUS	2	30	ACER_EVENT	913.7	2	city_subway	city_subway	NULL	NULL	0
AUS	2	30	ACER_EVENT	915.1	3	concession	city_subway	concession	NULL	0
AUS	2	30	ACER_EVENT	921.1	4	individual	city_subway	concession	individual	0
AUS	2	30	ACER_EVENT	923.0	5	Cancel	NULL	NULL	NULL	0
AUS	2	30	ACER_EVENT	928.1	6	city_subway	city_subway	NULL	NULL	0
AUS	2	30	ACER_EVENT	929.5	7	concession	city_subway	concession	NULL	0
AUS	2	30	ACER_EVENT	930.3	8	daily	city_subway	concession	daily	0
AUS	2	30	ACER_EVENT	932.8	9	Buy	city_subway	concession	daily	0
AUS	2	30	END_ITEM	938.5	10	NULL	NULL	NULL	NULL	NULL

Figure A2

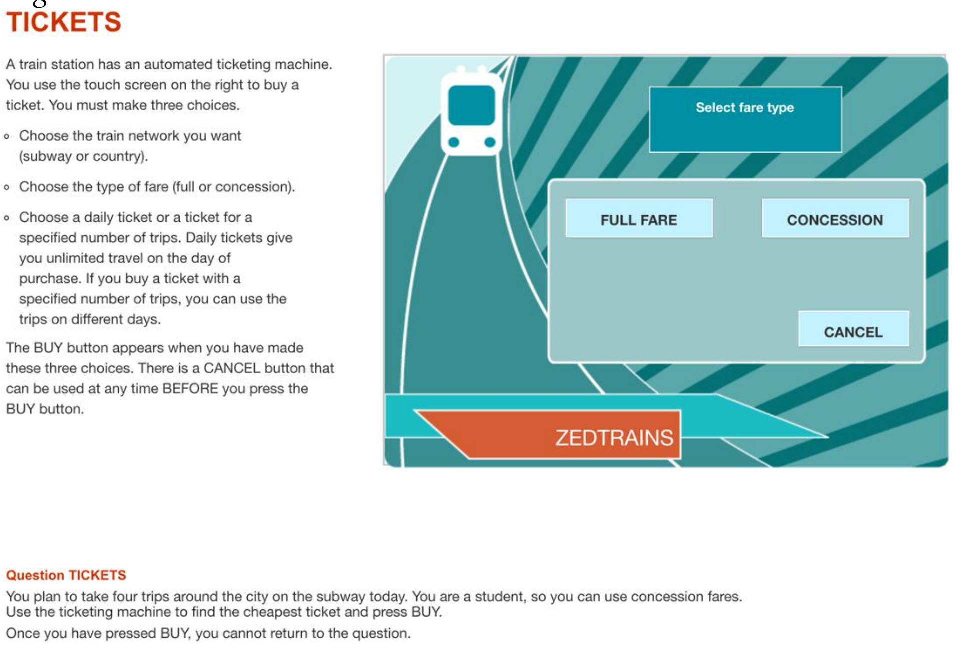
Screenshot of Formatted Process Data for Twenty Students.



city_subway	concession	daily.trip4	cancel	buy	city_concession	con_daily_ind_other_trip4	city_con_daily_ind_other_trip4	city_con_daily_cancel	daily_cancel	con_daily_cancel	daily.trip4_buy	concession_daily.trip4_buy	city_con_daily.trip4_buy
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	1	1	1
1	1	1	0	1	1	1	1	0	0	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	0	0	0	1	1	1
1	1	1	0	1	1	1	1	0	0	0	1	1	1
1	1	1	1	1	1	1	1	0	1	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	1	1	0	0	0	0	1	0	1	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	1	1	0	1	1	1	1	0	0	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure A3

Screenshot for the TICKETS Task2 (CP038Q01) in PISA 2012 after Clicking on A Train Network in Figure 2.



TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.

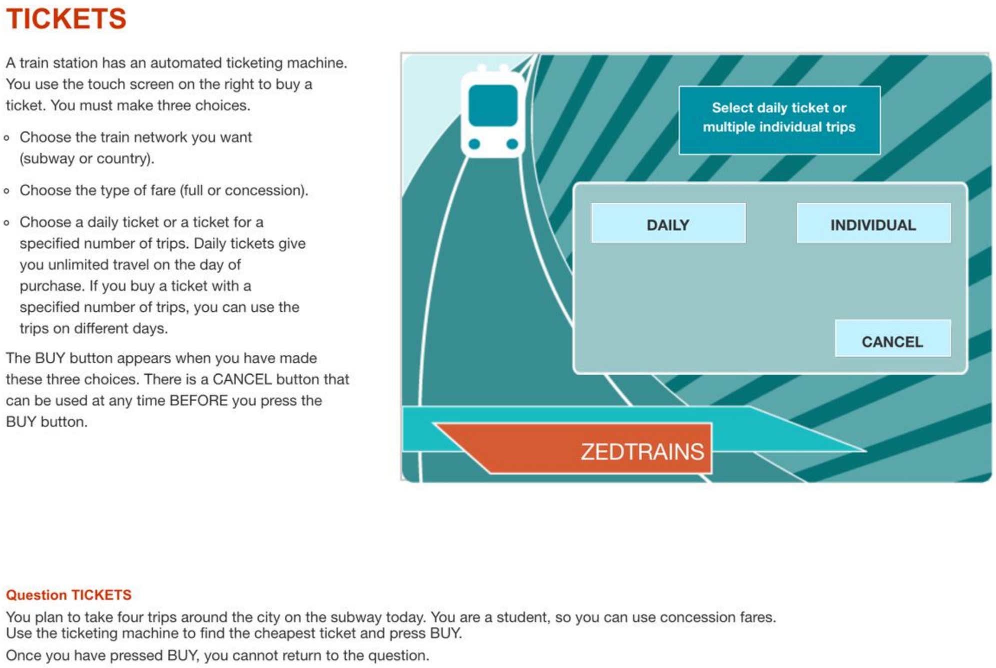
Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Figure A4

Screenshot for the TICKETS Task2 (CP038Q01) in PISA 2012 after Clicking on A Fare Type in Figure A3.



TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.

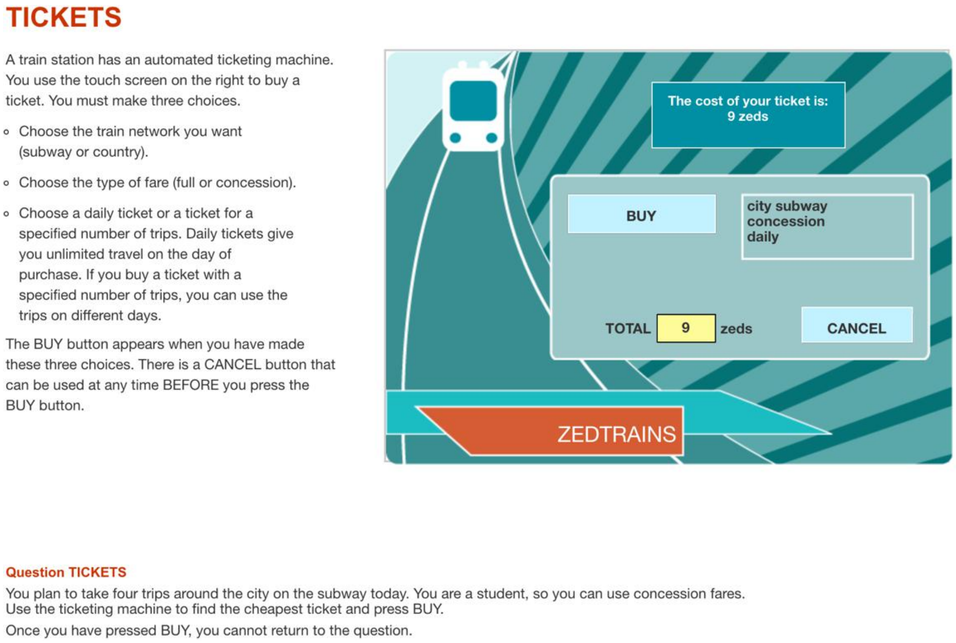
Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Figure A5

Screenshot for the TICKETS Task2 (CP038Q01) in PISA 2012 after Clicking on A Daily Ticket Type in Figure A4.



TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.

The cost of your ticket is:
9 zeds

city subway
concession
daily

TOTAL 9 zeds

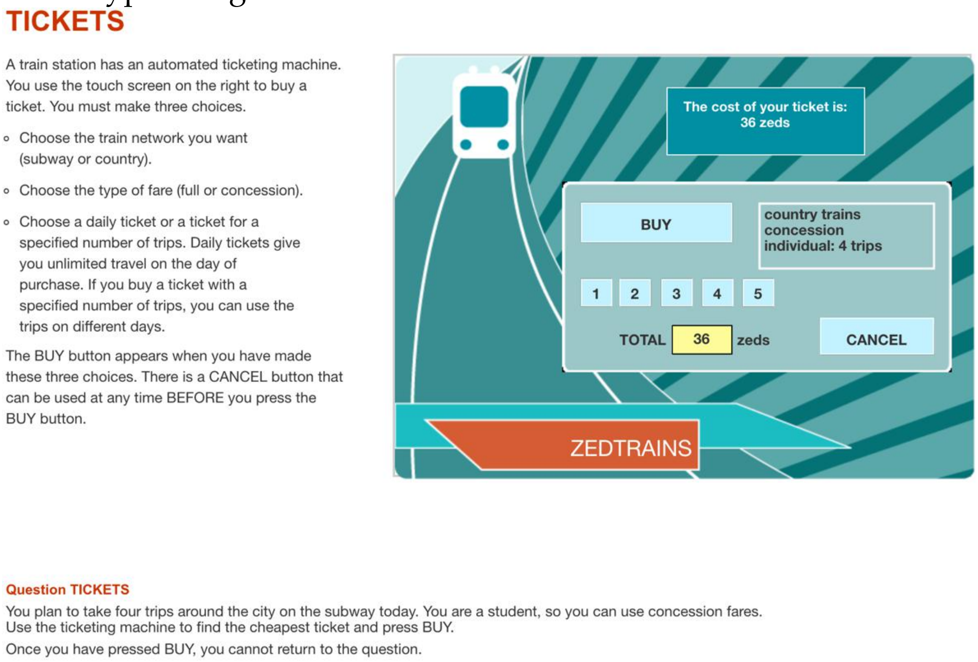
Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Figure A6

Screenshot for the TICKETS Task2 (CP038Q01) in PISA 2012 after Clicking on An Individual Ticket Type in Figure A4.



TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.

The cost of your ticket is:
36 zeds

country trains
concession
individual: 4 trips

TOTAL 36 zeds

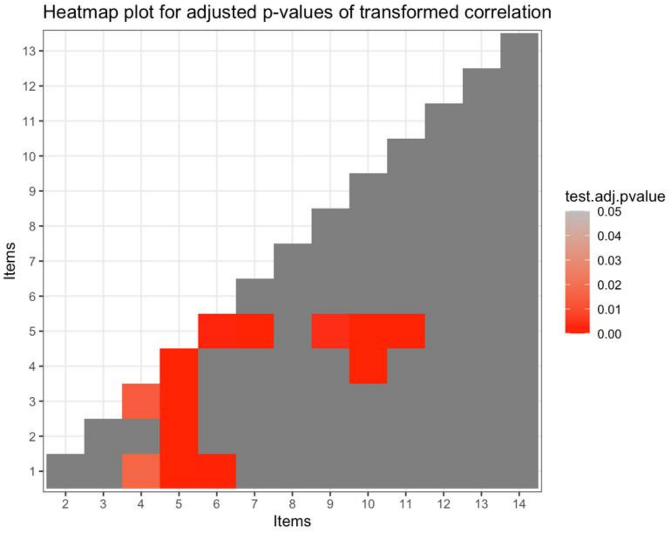
Question TICKETS

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY.

Once you have pressed BUY, you cannot return to the question.

Figure A7

Item-level Absolute Model-data Fit.



The attribute mastery proportions, attribute pattern mixing proportions, and attribute correlations obtained from the HO-DINA model fitted to the formatted process data with 3,760 respondents and 14 phantom items are presented in Tables S1, S2, and S3, respectively. Based on Table S1, α_4 (i.e., comparing the two ticket prices to find the cheapest) was estimated to be mastered by the least respondents (39.9%), which indicates that α_4 was the most difficult problem-solving skill given the current PISA item. Further, the attribute pattern mixing proportions shown in Figure A8 indicate the proportions of respondents in each estimated latent attribute pattern. It can be seen that respondents were classified into more categories than their observed score categories. Such fine-grained diagnostic classifications provide valuable remedial information to the respondents. Lastly, the maximum likelihood polychoric correlation estimates among the attributes were obtained using the *polychor* function in the "polycor" R package (Fox, 2019), as shown in Table S2. All correlations were positive and statistically significant except for the correlation between α_4 and α_5, which was negative and nonsignificant. This indicates that respondents who were able to make a decision to buy tickets did not necessarily compared the tickets to find the cheapest one to buy.

Table S1
Attribute Mastery Proportions.

Attribute	Attribute Mastery Proportions
α_1	0.903
α_2	0.914
α_3	0.882
α_4	0.399
α_5	0.883

Note. α_1 = understanding the city subway and the correct train network, α_2 = understanding that concession fares were available, α_3 = understanding that either a daily or four individual tickets allowed them to travel four times around the city, α_4 = comparing the two ticket prices to find the cheapest, α_5 = making a decision to buy.

Figure A8

Attribute Pattern Mixing Proportions.

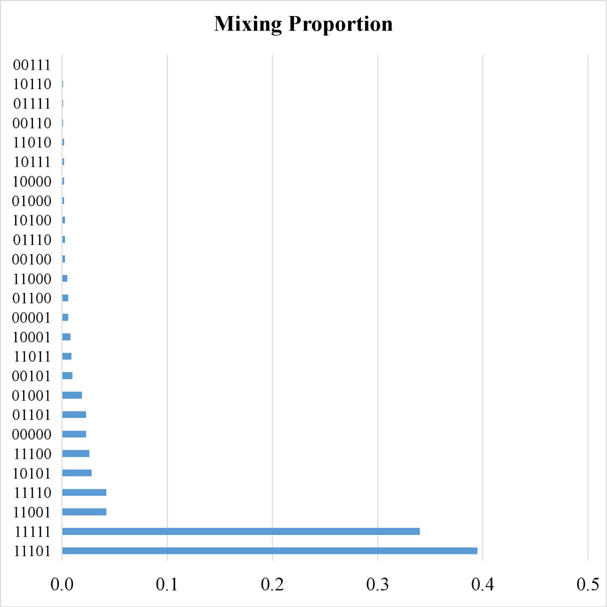


Table S2

Attribute Correlations.

	α_1	α_2	α_3	α_4
α_2	0.715 (0.025)			
α_3	0.691 (0.025)	0.600 (0.030)		
α_4	0.619 (0.032)	0.654 (0.033)	0.536 (0.030)	
α_5	0.526 (0.032)	0.492 (0.034)	0.398 (0.035)	-0.015 (0.034)

Note. α_1 = understanding the city subway and the correct train network, α_2 = understanding that concession fares were available, α_3 = understanding that either a daily or four individual tickets allowed them to travel four times around the city, α_4 = comparing the two ticket prices to find the cheapest, α_5 = making a decision to buy. Numbers in parenthesis are standard errors.

Section S1. Sensitivity Analysis

A sensitivity analysis was conducted to examine the impact of different choices of phantom items in the construction of Q-matrix and formatted process data. Specifically, Q-matrices and formatted process data that included less or more items than that used in the current study were constructed. To be consistent with the main study, HO-DINA model was fitted to the modified datasets. Then, the diagnostic classification results obtained from the analyses were compared to that from the main study.

Table S3 shows the reduced Q-matrix consisting of 11 phantom items, which satisfied the minimum requirements for the identifiability of the DINA model, as mentioned in the section "Item Expansion". Table S4 shows the expanded Q-matrix with 3 more items (i.e., con_ind_trip4, ind_trip4_cancel, ind_trip4_buy), which were considered as either not reflecting the latent construct or duplicated from existing items. Table S4 shows the diagnostic classification results from the two modified Q-matrices and formatted process data. We can see that the classification results remained similar to the original analysis in terms of the number of classes and the number of respondents in each class. It is worth noting that reduced Q-matrix led to slight misclassifications among the respondents. Therefore, we recommend to keep all phantom items that are believed to reflect the latent construct.

Table S3

Reduced Q-matrix Created for PISA 2012 Problem-solving Item TICKETS Task 2.

Item Numbe r	Phantom Items	Problem-solving Skills				
		α_1	α_2	α_3	α_4	α_5
1	city	1	0	0	0	0
2	con	0	1	0	0	0
3	daily/trip4	0	0	1	0	0
4	cancel	0	0	0	1	0
5	buy	0	0	0	0	1
6	city→con	1	1	0	0	0
7	city→con→daily/trip4	1	1	1	0	0
8	daily→cancel	0	0	1	1	0
9	con→daily→cancel	0	1	1	1	0
10	daily/trip4→buy	0	0	1	0	1
11	con→daily/trip4→buy	0	1	1	0	1

Note: city = city subway, con = concession, ind = individual, other = number of individual trips other than four, trip4 = four individual trips, α_1 = understanding the city subway and the correct train network, α_2 = understanding that concession fares were available, α_3 = understanding that either a daily or four individual tickets allowed them to travel four times around the city, α_4 = comparing the two ticket prices to find the cheapest, α_5 = making a decision to buy.

Table S4

Expanded Q-matrix Created for PISA 2012 Problem-solving Item TICKETS Task 2.

Item Number	Phantom Items	Problem-solving Skills				
		α_1	α_2	α_3	α_4	α_5
1	city	1	0	0	0	0
2	con	0	1	0	0	0
3	daily/trip4	0	0	1	0	0
4	cancel	0	0	0	1	0
5	buy	0	0	0	0	1
6	city→con	1	1	0	0	0
7	con→daily/trip4	0	1	1	0	0
8	city→con→daily/trip4	1	1	1	0	0
9	city→con→daily→cancel	1	1	1	1	0
10	daily→cancel	0	0	1	1	0
11	con→daily→cancel	0	1	1	1	0
12	daily/trip4→buy	0	0	1	0	1
13	con→daily/trip4→buy	0	1	1	0	1
14	city→con→daily/trip4→buy	1	1	1	0	1
15	con→ ind→trip4	0	1	1	0	0
16	ind→trip4→cancel	0	0	1	1	0
17	ind→trip4→buy	0	0	1	0	1

Note: city = city subway, con = concession, ind = individual, other = number of individual trips other than four, trip4 = four individual trips, α_1 = understanding the city subway and the correct train network, α_2 = understanding that concession fares were available, α_3 = understanding that either a daily or four individual tickets allowed them to travel four times around the city, α_4 = comparing the two ticket prices to find the cheapest, α_5 = making a decision to buy.

Table S5

Diagnostic Classification Results from the Sensitivity Analyses Compared to Original Results

Observed Score Category	Latent Attribute Pattern	Frequency		
		Original Q	Reduced Q	Expanded Q
2	11111	1,093	1,093	1,093
1	11111	156	176	156
	11101	1,481	1,461	1,481
0	00000	87	87	87
	00001	22	22	22
	00100	10	10	10
	00101	39	39	39
	00110	2	1	2
	00111	1	1	1
	01000	6	6	6
	01001	72	72	72
	01100	21	19	21
	01101	88	81	88
	01110	13	8	13
	01111	2	7	2
	10000	9	9	9
	10001	29	29	29
	10100	10	10	10
	10101	107	107	107
	10110	2	4	2
	10111	6	5	6
	11000	19	19	19
	11001	157	158	157
	11010	6	6	6
	11011	35	26	35
	11100	98	92	98
	11101	3	12	3
	11110	157	160	156
	11111	29	40	30

Section S2. Reliability and Validity

When new methods are used to analyze existing data, the reliability and validity of the analysis results should also be considered. In this study, the classification accuracy (P_a) index (Wang et al., 2015) was used for evaluating the reliability of classification results. In addition, validity evidence was provided in the interpretation of the problem-solving abilities and the problem-solving skill patterns obtained using the proposed method (see *Reliability and Validity.R* in the shared code).

Reliability of Classification. P_a index refers to the degree to which a respondent's classification estimate matches his/her true latent class. According to Ravand and Robitzsch (2018), values of at least 0.8 for the P_a index can be considered as acceptable classification rates. As shown in Table S6, both test- and attribute-level classification accuracies were within the acceptable range. Therefore, the results indicate adequate reliability of classification obtained from the proposed method.

Table S6

Classification Accuracy for the HO-DINA Model.

Attributes- or Test-level Accuracy	Classification Accuracy (P_a)
A_1	0.998
A_2	0.999
A_3	0.999
A_4	0.995
A_5	0.987
Test-level	0.981

Validity evidence for problem-solving ability. Validity evidence for problem-solving ability was based on its relations to other variables (AERA et al., 2014). First, item responses for five problem-solving items CP025Q01, CP025Q02, CP038Q01, CP038Q02, CP038Q03 from these 3,760 respondents were analyzed using IRT models to show the consistency between the problem-solving ability estimates obtained from the proposed method (denoted as θ_1) and those obtained from the IRT models (denoted as θ_G). Specifically, the former is the problem-solving ability regarding the targeted cognitive process (i.e., exploring and understanding) and the latter is the general problem-solving ability regarding all four measured cognitive processes in PISA 2012 (i.e., exploring and understanding, planning and executing, monitoring and reflecting, and representing and formulating). These five problem-solving items, including 3 polytomously scored items and 2 dichotomously scored items, were analyzed using the PCM and the Rasch model (Rasch, 1960), respectively. The correlation coefficient between θ_1s and θ_Gs was 0.674 ($p < 0.001$). Such a significantly positive correlation indicates that there was a high consistency between θ_1 and θ_G, but they can still be distinguished because of the different latent constructs being measured.

Second, statistically significant correlations among problem-solving, math, and reading abilities would support the validity of the problem-solving abilities estimated from the proposed method based on existing studies (e.g., Öztürk et al., 2020). In the sample of 3,760 respondents, 2,594 respondents who had both math and reading scores were retained in the correlation analysis of their problem-solving abilities and reading/math abilities (see *cogsdata.rds* in the shared code). In PISA 2012, there were 84 math items, among which 8 items were polytomously scored and 76 items were dichotomously scored. There were 44 reading items, among which 1 item was polytomously scored and 43 items were dichotomously scored.

Then, the math items and reading items were calibrated separately using the Rasch model and the PCM. Missing responses were accommodated by FIML. The math and reading ability estimates were further obtained using IRT scoring. The problem-solving ability θ_1s obtained from the proposed method were further correlated with math and reading ability estimates, respectively.

As a result, the correlation between the problem-solving ability estimates θ_1s and the math ability estimates was 0.444 ($p < 0.001$). The correlation between the problem-solving ability estimates θ_1s and the reading ability estimates was 0.326 (p < 0.001). These results are consistent with the results from existing studies (e.g., Öztürk et al., 2020). Therefore, these results supported that the problem-solving ability estimates obtained from the proposed methods were valid.

Validity evidence for problem-solving skill pattern. As aforementioned, Table 5 in the main text shows the distribution of respondents and their latent attribute patterns with respect to their observed score categories. The consistency between the latent attribute patterns and their observed score categories suggested the score validity from the proposed method. In addition, *k*-means and SOM were used to cross validate the classification results from the proposed method. Specifically, the *k*-means was carried out using the *kmeans* function in the R package *stats* (Version 4.0.3) with maximum iterations allowed equal to 10. The SOM was carried out using the R package *kohonen* (Version 3.0.10). The learning rate of the SOM declined from 0.05 to 0.01 over 2000 iterations. The phantom item response matrix was used as input data. The number of clusters was set at both the number of observed scores (i.e., 3 score categories) and the number of latent attribute patterns obtained from the DCM (i.e., 26 latent classes). It is expected that the number of latent attribute patterns is more than the number of observed scores, thus, showing more fine-grained diagnostic classification information on students' problem-solving skills. Consistency between the classification results from the DCM and the unsupervised data mining methods indicates the validity of the proposed method.

The results of the two unsupervised data mining methods (i.e., *k*-means and SOM) with 3 clusters based on 3 observed score categories (i.e., 0, 1, 2) are shown in Table S7. It can be seen that the classification results are consistent between the two data mining methods. In addition, these results are also consistent as those obtained from the proposed method shown in the Table 5 in the main text. The results of the two unsupervised data mining methods with 26 clusters (i.e., the number of latent attribute pattens from the DCM) were also obtained and are presented in Table S8. The classification results were consistent among the two data mining methods and the proposed method in general. Specifically, the numbers of students in cluster 1 in both data mining methods were the same as the number of students who mastered all attributes (i.e., latent attribute pattern = 11111) in the score category 2. Although the classification results are not exactly the same for students in score categories 1 and 0 among the methods due to their inherent differences and estimation errors, consistency can be found to some extent. In addition, the proposed method based on DCM is more advantageous than unsupervised data mining methods in that it has readily interpretable latent attribute patterns while the clusters obtained from unsupervised data mining methods require further labeling. Therefore, the proposed method can provide both valid and interpretable diagnostic classifications on students' problem-solving skills.

In sum, we have provided validity evidence related to the purpose of the proposed method based on the available data from PISA. All the evidence indicates that the proposed method has the capability to assess problem-solving competence.

Table S7
Results from k-means and SOM with 3 Clusters with Respect to Their Observed Score Category.

Observed Score Category	*k*-means Clusters			SOM Clusters		
	1	2	3	1	2	3
2	1,093	0	0	1,093	0	0
1	156	1,481	0	156	1,481	0
0	194	9	827	189	9	832

Note, results are obtained based on fixed random seed in R: set.seed(1234).

Table S8

Results from k-means and SOM with 26 Clusters with Respect to Their Observed Score Category.

Observed Score Category	Cluster	*k*-means	SOM
2	1	1,093	1,093
1	1	130	130
	2	156	156
	3	1,305	1,305
	4	46	26
	5	0	20
0	1	28	5
	2	3	3
	3	0	0
	4	1	1
	5	4	1
	6	6	0
	7	7	0
	8	8	0
	9	9	8
	10	11	11
	11	13	14
	12	16	15
	13	16	17
	14	19	19
	15	22	29
	16	25	32
	17	39	35
	18	44	45
	19	54	54
	20	73	57
	21	77	78
	22	78	80
	23	92	94
	24	100	110
	25	102	114
	26	183	208

Note, results are obtained based on fixed random seed in R: set.seed(1234).

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Fox, J. (2019). polycor: Polychoric and Polyserial Correlations. [R package version 0.7-10]. Retrieved from https://CRAN.R-project.org/package=polycor.

Kohonen, T. (1997). *Self-Organizing Maps.* Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-97966-8

Öztürk, M., Akkan, Y., & Kaplan, A. (2020). Reading comprehension, Mathematics self-efficacy perception, and Mathematics attitude as correlates of students' non-routine Mathematics problem-solving skills in Turkey. *International Journal of Mathematical Education in Science and Technology*, *51*(7), 1042-1058.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen and Lydiche, Copenhagen.

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educational Psychology, 38*, 1255-1277.