Online Appendix

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

Peida Zhan[1] Xin Qiao[2]

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

[1] Peida Zhan, Zhejiang Normal University, E-mail: pdzhan@gmail.com
[2] Xin Qiao, University of Maryland, College Park, E-mail: xin.qiao56@gmail.com

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

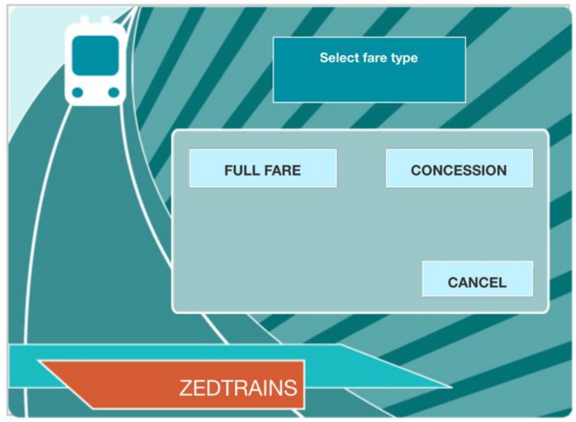
Screenshot for the TICKETS Task2 (CP038Q01) in PISA 2012 after Clicking on A Train Network in Figure 2.

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

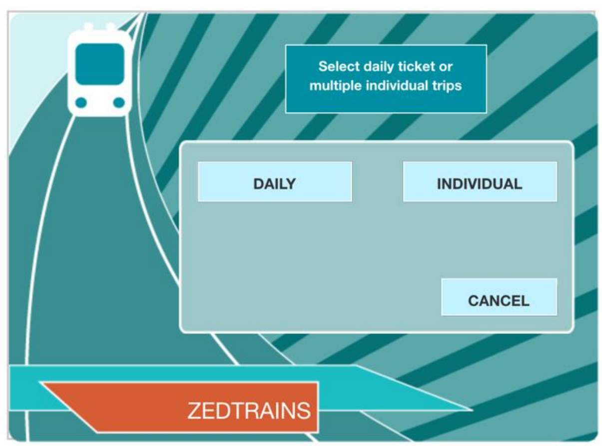
Screenshot for the TICKETS Task2 (CP038Q01) in PISA 2012 after Clicking on A Fare Type in Figure A3.

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.

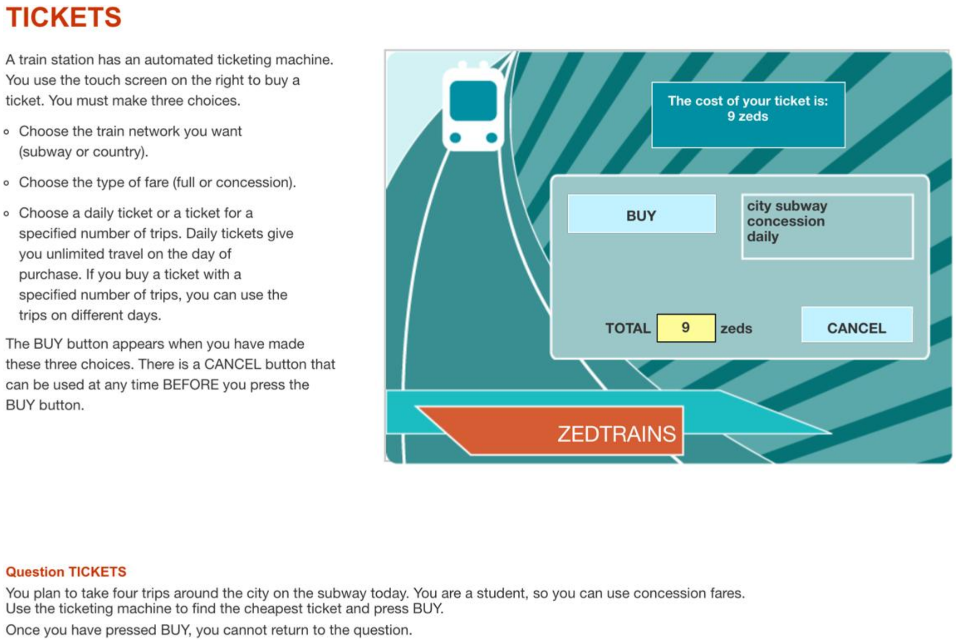


Figure A6

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

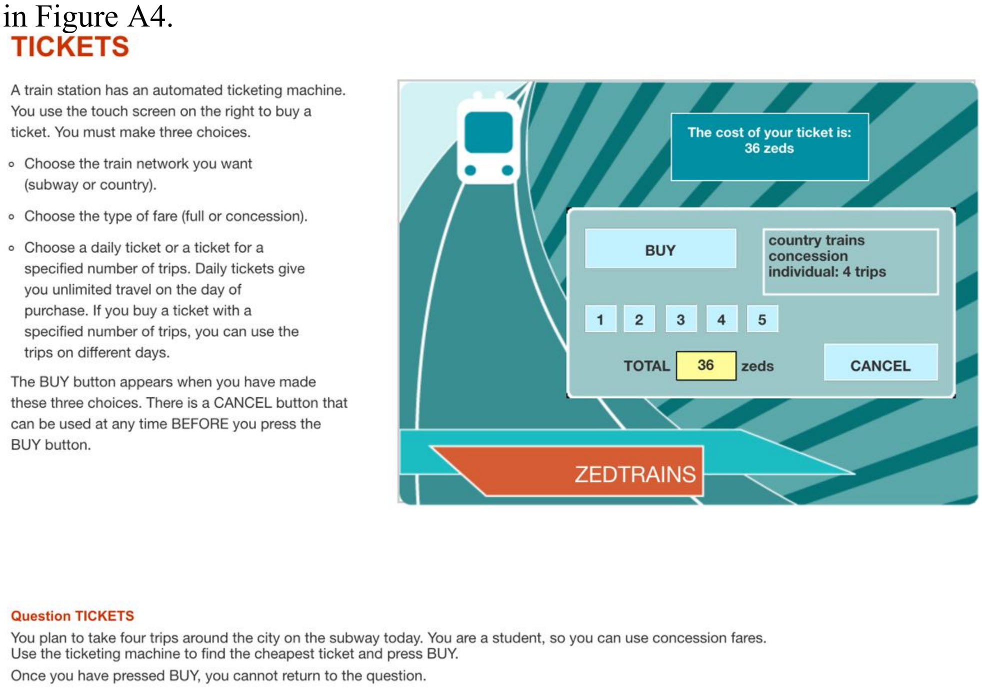
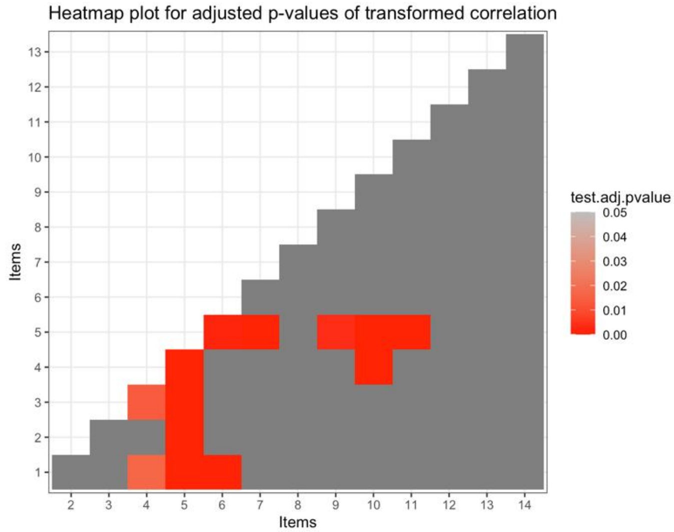


Figure A7

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 amount of 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.

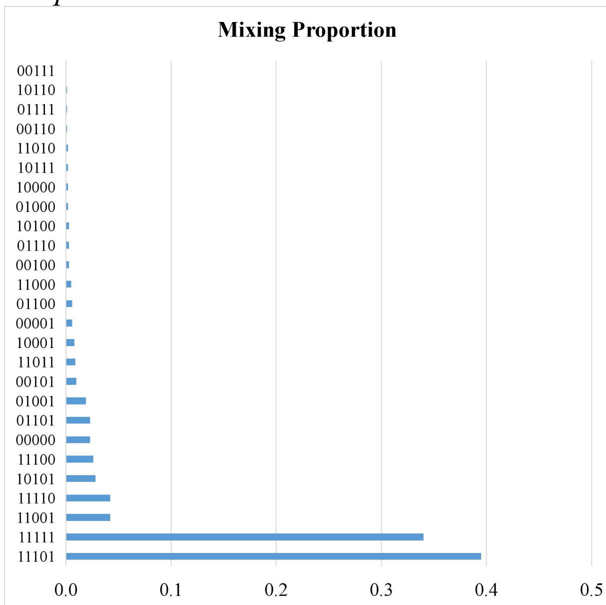


Table S2

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 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	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.