Using Machine Learning to Predict the Number of Latent Skills in Online Learning Environments

Changsheng Chen[1, 2, *], Robbe D'hondt[2, 3], Celine Vens[2, 3] and Wim Van Den Noortgate[1, 2]

[1] *Faculty of Psychology and Educational Sciences, KU Leuven, Campus KULAK, Kortrijk, Belgium*
[2] *imec research group itec, KU Leuven, Kortrijk, Belgium*
[3] *Department of Public Health and Primary Care, KU Leuven, Campus KULAK, Kortrijk, Belgium*

Abstract

Extracting skill information for students in online learning environments has been a challenging topic across different domains. Predicting the number of skills is the first step towards estimating students' skills. In this paper, we propose prediction methods based on Machine Learning (ML) models, where we used the analysis model to generate simulation data reflecting the data features of our target scenarios and took the features from simulation data to train and test ML models. We illustrated this approach in tandem with Multidimensional Item Response Theory (MIRT) for the simple and complex structure, and further compared the trained ML models with a selection of statistical methods based on the test data. Our preliminary results show that, compared to statistical methods, ML models generally reach a noticeably higher proportion of correct estimations for both structures. Additionally, we find that an increase in the percentage of missing values and sample size leads to negative and positive effects on the methods' performance respectively. Using simulation data from the analysis model to train ML models and doing prediction can extend the current operation of skill extraction, which provides extra options for the practitioners.

Keywords

machine learning, multidimensional item response theory, latent skills, online learning

1. Introduction

Skill information is one type of fundamental quantitative evidence for building an online learning system (including adaptive lifelong learning system). With accurate users' skill estimates, such a system can personalize materials and instruction design to improve the learning experience effectively and efficiently. With monitoring the changes of users' skill information, the system can recommend further learning resources to adapt to users' situation frequently. However, what skills can be extracted and monitored and how the skill information can be estimated by which test items and relevant users' response are still a challenging topic.

Several kinds of techniques have been used to extract users' skill information based on users' response to test items, such as Multidimensional Item Response Theory (MIRT) [2], Cognitive Diagnostic Model (CDM) [3], Matrix Factorization (MF) [4,5], and so forth. The common start

ALL'24: Workshop on Adaptive Lifelong Learning, July 08–12, 2024, Recife, Brazil [1]
* Corresponding author.
changsheng.chen@kuleuven.be (C. Chen); robbe.dhondt@kuleuven.be (R. D'hondt); celine.vens@kuleuven.be (C. Vens); wim.vandennoortgate@kuleuven.be (W. V. D. Noortgate)
0000-0001-6092-6655 (C. Chen); 0000-0001-7843-2178 (R. D'hondt); 0000-0003-0983-256X (C. Vens); 0000-0003-4011-219X (W. V. D. Noortgate)

for conducting these techniques is to decide the number of skills and clarify the relationship between items and skills (or knowledge components). In other words, the number of skills and which items can be used to measure which skills are clearly defined before skill estimation and tracing algorithms are performed. For example, in the MIRT, the item-dimension relationship needs to be explored, which serves as the basis for estimating user's skill values, after the predetermination of the number of latent dimensions. In the CDM, the item-attributes relationship depicted by the Q-matrix functions in a similar way and the number of attributes should also be confirmed beforehand. In the MF, the number of ranks for shaping two decomposed matrices (i.e., a user-factor matrix and an item-factor matrix) is required initially before the technique is performed. Traditionally, the number of skills and the item-skill relationship are theoretically defined by domain experts. However, human examination is too inefficient to satisfy the needs of online learning system because of the large number of items, which calls for the data-driven approach (i.e., extracting the number of skills and exploring and confirming the item-skill structure based on the response matrix).

Many techniques have been proposed to estimate the number of skills based on data-driven evidence. For example, in the MIRT, the number of latent dimensions is estimated by certain statistical methods, such as Kaiser Criterion (KC) [6], Empirical Kaiser Criterion (EKC) [7], Parallel Analysis (PA) [8], non-graphical Scree Plot with Optimal Coordinates (OC) or Acceleration Factor (AF) [9], Very Simple Structure (VSS) with two variants (i.e., C1 & C2) [10], and so forth. In the CDM, the number of attributes and related Q-matrix are estimated and evaluated by the designed algorithms or statistics, such as the G-DINA Discrimination Index (GDI) method [11], the stepwise method [12], and so on. In the MF, the number of ranks is usually seen as a hyperparameter, which is predicted based on the evaluation of defined loss [13]. Additionally, some researchers have explored using Machine Learning (ML) methods to estimate the number of skills, and they found that it can increase the proportion of correct predictions. For example, Goretzko & Bühner [14] used eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Adaptive Boosting to predict the number of factors for continuous response simulation data, and found that these methods performed better than other traditional statistical methods in terms of prediction accuracy (i.e., the proportion of correct estimation). However, their study did not explore the possibilities of using ML methods to predict the number of skills for the dichotomous response with considering the features of online or adaptive learning data (e.g., the sparsity and the large number of items) and properties of different multidimensional structures.

In this study, we aim to fill this research gap by proposing ML prediction methods inspired by Goretzko & Bühner [14] and comparing their performance with other selected statistical methods. The general operation is that we use the analysis model (such as the MIRT, CDM, or MF) to generate simulation data reflecting the data features of target scenarios in online learning environments. The simulation data includes two parts, i.e., the training data (including validation data) for training and tuning ML models and the test data for evaluating the performance of ML models and selected statistical methods. In detail, the selected methods included: 1) ML models: the regression variant of XGBoost and RF whose results were rounded to the integer; 2) statistical methods: KC, PA, EKC, Scree Plot (OC), Scree Plot (AF), VSS (C1), and VSS (C2). For the sake of parsimony, the explanation of methods' mechanism is skipped, and relevant details can be consult by provided references.

In the following sections, we illustrate this operation in tandem with the MIRT for the simple and complex structure. MIRT is the prevailing statistical model for analyzing students' binary response (0: wrong; 1: right) to estimate students' ability and relevant item parameters in the field of psychological and educational assessments. The principle of MIRT is that it models the probability of giving a correct answer based on the interaction between students' ability and item parameters. For example, a 2-parameter MIRT model can be expressed by:

$$P\left(x_{ij}=1\middle|\boldsymbol{\theta_i};\boldsymbol{a_j},d_j\right)=\frac{exp(\boldsymbol{\alpha_j\theta_i'}+d_j)}{1+exp(\boldsymbol{\alpha_j\theta_i'}+d_j)}$$

In the above formula, $x_{ij}=1$ refers to the correct response of user *i* for item *j* and the $\boldsymbol{\theta_i}=(\theta_{i1},\theta_{i2},\ldots,\theta_{ik})$, $\boldsymbol{\alpha_j}=(\alpha_{j1},\alpha_{j2},\ldots,\alpha_{jk})$, and d_j indicate the ability of user *i* for skill *k*, the item discrimination of item *j* for skill *k*, and the item intercept for item *j* respectively [2]. As for the two multidimensional structures, under the simple structure, each item is solely related to one latent skill and the latent skills are correlated with each other. Under the complex structure, each item is related to more than one latent skill and the latent skills are correlated with each other as well.

2. Method

2.1. Data

Table 1 presents the settings for generating the training data and test data by a 2-parameter MIRT model for the simple and complex structures based on R function "simdata" of R package "mirt" [15] in R 4.3.2 [16]. These simulation features contained the number of items, the number of latent skills, the sample size, and the proportion of missing values in the response matrix, and the correlation between latent skills. The relevant settings mimicked the possible features of online learning and assessments [17,18]. The settings for generating the training data were randomly selected from the designed range for each simulation feature, except for the number of latent skills. In detail, we randomly selected 20 values from the specified range for the number of items. For the sample size and missingness, we randomly selected 10 values, and for the correlation, we randomly selected 5 values. The setting for generating the test data were based on fixed values for detecting their effects on methods' performance. In total, there were 80,000 and 7200 scenarios for the training and test data respectively. Considering the constraints on computation power, we randomly selected 1000 scenarios for both and generated one dataset for each scenario as the preliminary results for the subsequent analysis. The simulation codes will be publicly available by contacting the corresponding author when the paper with final results is published.

Table 1
Settings of Generating Simulation Data

Features	Settings for Training Data	Settings for Test Data
The number of items	From 300 to 800	300, 400, 500, 600, 700, 800
The number of latent skills	1, 2, 3, 4, 5, 6, 7, 8	1, 2, 3, 4, 5, 6, 7, 8
Sample size	From 300 to 800	300, 400, 500, 600, 700, 800

Missingness (proportion)	From 0 to 0.9	0, 0.25, 0.5, 0.75, 0.9
Correlation (latent skills)	From 0.1 to 0.5	0.1, 0.2, 0.3, 0.4, 0.5

2.2. Methods Implementation

All methods implementation was based on R 4.3.2 [16]. The statistical methods were mainly implemented based on the tetrachoric correlation matrix corresponding to the dichotomous responses by R function "tetrachoric2" of R package "sirt" [19] with Bonett method [20]. The results of KC and EKC were estimated by manual function in R. PA and scree plot (OC & AF) were performed by relevant functions in R package "nFactors" [21], and VSS (C1 & C2) was implemented by relevant functions in R package "psych" [22].

Table 2
Hyperparameter Consideration for ML models

Random Forest		XGBoost	
Number of trees	From 10 to 500	Maximum depth of a tree	From 1 to 20
Number of considered variables at each split	From 1 to all features	Minimum sum of instance weight (hessian)	From 1 to 10
Minimum size of terminal nodes	From 1 to 10	Fraction of features for each tree	From 0.5 to 1
Maximum size of terminal nodes	From 5 to 50	Fraction of samples for each tree	From 0.5 to 1
Maximum number of iterations for tuning	100	Number of boosting rounds	From 30 to 100
Loss function	Mean Squared Error	Learning rate	From 0.01 to 0.5
		Minimum loss reduction	From 0 to 10
		Loss function	Mean Square Error

The RF and XGBoost were implemented by relevant functions in R package "mlr" [23] and "xgboost" [24]. Both ML models were trained and tested based on the features extracted from available information, such as the original response matrix, the estimated tetrachoric correlation matrix, and the estimated results of statistical methods. The features included [14]: 1) from the response matrix: the sample size, the number of items, and the proportion of missingness; 2) from the correlation matrix: the determinant, the number of entries smaller or equal to 0.1, the number of eigenvalues larger than 0.7, the relative proportion of eigenvalues, the standard deviation of all eigenvalues, the number of eigenvalues accounting for over 50% or 75% of the variance, the matrix norms (i.e., the L1-norm, Frobenius-norm, maximum-norm, and spectral-norm), the average of off-diagonal entries and the communality estimates, the sampling adequacy [25], the Gini-coefficient [26], the Kolm inequality [27], the top 50

eigenvalue estimates; 3) from the results of statistical methods: KC, PA, EKC, scree plot (OC), scree plot (AF), VSS (C1), and VSS(C2).

As ML models can be trained by integrating the results of statistical methods, which may lead to a fairness concern regarding the method comparison, we trained RF and XGBoost in two ways, i.e., one without including results of statistical methods in the features and another with including them. Additionally, all ML models were trained by 10-fold cross-validation based on the training data. Table 2 provides the partial hyperparameter settings for the RF and XGBoost with or without extra features (i.e., the results of statistical methods). The settings of other possible hyperparameters followed the default settings of two R packages. The relevant codes will be publicly available by contacting the corresponding author when the paper with final results is published.

2.3. Evaluation Metrics

To evaluate and compare the performance of all candidate methods, the deviation score and several metrics based on the deviation score were used. The deviation score is defined as the estimated number of latent skills minus the true number of latent skills. The correct-estimation proportion is the number of deviation scores equal to zero divided by the total number of estimates (i.e., 1000). The under-estimation proportion is the number of deviation scores lower than zero divided by the total number of estimates. The over-estimation proportion is the number of deviation scores higher than zero divided by the total number of estimates. The bias is the average of deviation scores. The precision is the average absolute deviation score.

3. Results

Table 3 shows the results of all selected methods based on the test data. For the simple structure, KC, PA, EKA, and scree plot (OC) performed worse than other methods. Their correct-estimation proportions were nearly equal to zero. For scree plot (AF), VSS (C1), and VSS (C2), the correct-estimation proportions ranged from 0.5 to 0.3, which was obviously better than other statistical methods. Regarding the performance of ML models, RF and XGBoost without extra features reached even higher correct-estimation proportions (more than 0.7) than the variants with extra features (less than 0.7). The correct-estimation proportions of all ML models were higher than the statistical methods, with the minimum difference between them equal to 0.1954. In terms of under and over estimation, KC, PA, EKC, and scree plot (OC) tended to overly estimate the number of latent skills, which was further confirmed by the result of bias and precision. ML models tended to estimate a higher number of latent skills as well, although their over-estimation proportions were relatively lower. In contrast, scree plot (AF), VSS (C1), and VSS (C2) estimated a smaller number of latent skills than the true number of skills.

For the complex structure, the general pattern was similar to the simple structure. KC, PA, and scree plot (OC) had the lowest proportions of correct estimations, again close to zero. EKC performed poorly as well, even though its correct proportion was around 0.1. The correct proportion of scree plot (AF), VSS (C1), and VSS (C2) ranged from 0.2490 to 0.3740, which was better than other statistical methods. In terms of the performance of ML models, their proportions of correct estimation were higher than 0.74, which was substantially better than statistical methods. Regarding the under and over estimation, KC, PA, scree plot (OC), and EKC overly estimated the number of latent skills (their over-estimation proportions above 0.9), while

scree plot (AF), VSS (C1), and VSS (C2) tended to estimate a smaller number of latent skills (their under-estimation proportions ranging from around 0.4 to 0.5). ML models also estimated a smaller number of latent skills, but their under-estimation proportions (around 0.14) were noticeably lower than statistical methods. The patterns of under and over estimations were further supported by the results of bias and precision.

Table 3
Results of Test Data

	Correct-estimation Proportion	Under-estimation Proportion	Over-estimation Proportion	Bias	Precision
Simple Structure					
KC	0	0	1	168.0779	168.0779
PA	0.0065	0	0.9935	94.0455	94.0455
EKC	0.0195	0	0.9805	71.4870	71.4870
Scree Plot (OC)	0.0130	0.0065	0.9805	30.3312	30.3442
Scree Plot (AF)	0.3571	0.6169	0.0260	-2.7273	2.7792
VSS (C1)	0.3636	0.4481	0.1883	-1.6039	2.3571
VSS (C2)	0.4545	0.2792	0.2662	-0.6818	1.3312
RF	0.7143	0.1364	0.1494	0.0519	0.4935
RF (extra)	0.6883	0.0519	0.2597	0.5260	0.6818
XGBoost	0.7078	0.1104	0.1818	0.1169	0.4545
XGBoost (extra)	0.6494	0.0909	0.2597	0.4935	0.7143
Complex Structure					
KC	0	0	1	160.7990	160.7990
PA	0.0550	0	0.9450	86.2640	86.2640
EKC	0.1010	0	0.8990	67.2240	67.2240
Scree Plot (OC)	0.0580	0.0020	0.9400	27.1280	27.1420
Scree Plot (AF)	0.3740	0.5800	0.0460	-2.5580	2.6500
VSS (C1)	0.2490	0.6710	0.0800	-2.5470	2.7810
VSS (C2)	0.3130	0.3930	0.2940	-0.3890	1.7310
RF	0.7600	0.1280	0.1120	-0.1920	0.5240
RF (extra)	0.7490	0.1380	0.1130	-0.2280	0.5200
XGBoost	0.7940	0.1420	0.0640	-0.3040	0.4840
XGBoost (extra)	0.7820	0.1490	0.0690	-0.3210	0.5050

Figure 1 and Figure 2 present the effects of simulation features on the correct-estimation proportions of selected methods. As these proportions were extremely low for KC, PA, EKC, and scree plot (OC), they were omitted in the effects analysis. For the simple structure, when the percentage of missing values in the response matrix increased from 0 to 90%, the respective proportions of all methods decreased, especially for ML models (falling from above 0.8 to below 0.2). Raising the sample size from 300 to 800 generally led to an increase in the respective proportions of ML methods by 0.2, while the effects of sample size on statistical methods were not detectable due to the fluctuations. Regarding the effects of the number of latent skills, changing the settings from 1 to 8 was related to the tremendous decrease in the proportions of

scree plot (AF) and VSS (C1) by around 0.7. For the effects of the number of items, when it rose from 400 to 600, the proportion of most methods went down by around 0.2.

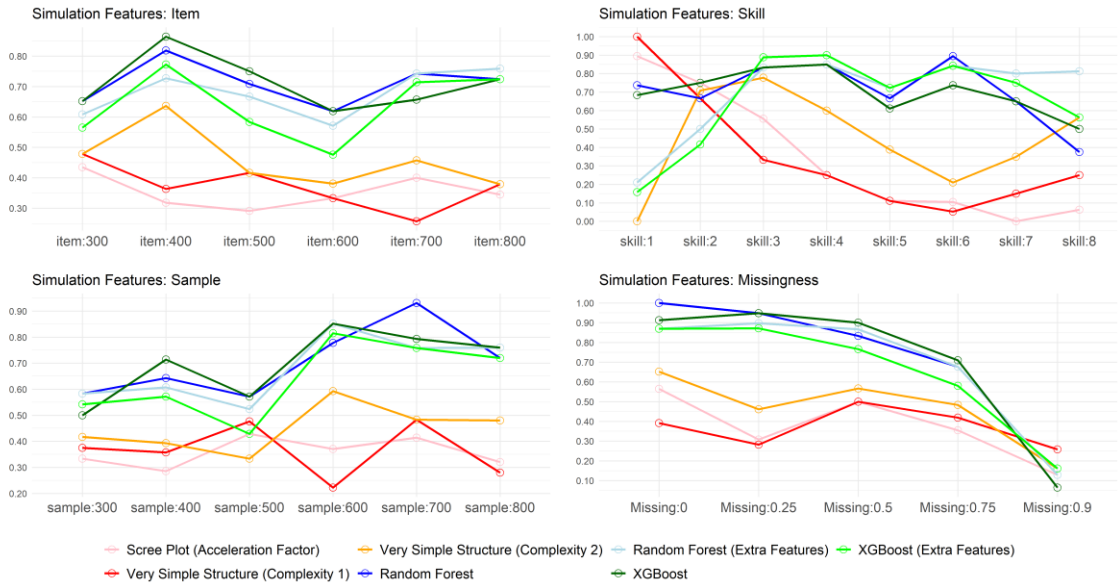

Figure 1: Effects of Simulation Features (x-axis) on the Correct-estimation Proportions (y-axis) for the Simple Structure

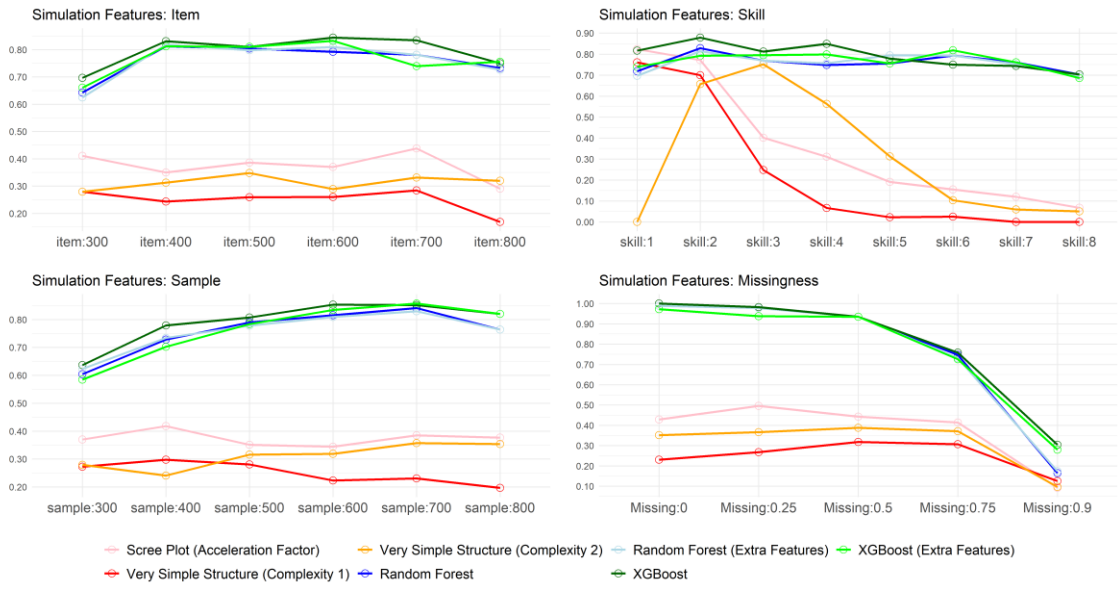

Figure 2: Effects of Simulation Features (x-axis) on the Correct-estimation Proportions (y-axis) for the Complex Structure

Compared to the patterns in the case of simple structure, the changes of proportions for the complex structure fluctuated less. When the missingness percentage went up from 0 to 90%, the

proportions of ML methods dropped down from over 0.9 to lower than 0.3 and the proportions of statistical methods went down relatively slightly by around 0.2. Raising the sample size led to the increase in proportions of ML methods by around 0.2, while the proportions of statistical methods fluctuated by a small amount. In terms of the number of latent skills, when it changed from 2 to 8, the proportion of statistical methods fell down massively from over 0.6 to below 0.1. In contrast, the proportion of ML models almost stayed the same. Regarding the number of items, the proportion of all methods fluctuated slightly without noticeable changes across different settings.

4. Discussion

In the present study, we proposed a general operation of building prediction models using ML, with simulation data to estimate the number of latent skills for online learning environments, which was illustrated based on the MIRT. The results of the performance comparison revealed that ML models had a markedly better performance than statistical methods regarding the correct-estimation proportions. This finding is generally consistent with the previous study [14]. However, the correct estimation of proportions in the previous study is higher than 0.9, which is different from the results in this study (ranging from 0.65 to 0.8). One possible explanation for this difference might be due to the different simulation models and scenarios. In the previous study, the dichotomous response generated by the MIRT was not considered. The simulation settings more reflected the features of relatively small-scale psychological tests instead of the large-scale online learning settings. For example, the number of items is usually set below 100 in the field of psychology, while it might be over hundreds and even thousands in the online learning environments. Additionally, the problem of missingness or sparsity is also less of a concern in previous research. Regarding the performance of statistical methods, our results showed that they performed surprisingly poorer than previous studies. Goretzko & Bühner [14] found that KC, EKC and PA reached over 0.75 regarding the correct estimation proportion, which is completely different from our results. Guo & Choi [28] found that the proportion of identifying the correct number of latent skills for PA with tetrachoric ranged from 0.43 to 1 across various simulation features, which is also dissimilar from our results. It may be speculated that this is because of the different settings of simulation features.

Except for the results of methods comparison, the effects analysis of simulation features found that the increase in the missingness and sample size lead to a going-down and going-up trends for most of methods regarding the correct estimation proportions. It is interesting to note that raising missingness and sample size may have negative and positive impact on methods' performance respectively. As mentioned above, missingness was not considered in the previous study, and our study fills this gap. As for the positive effects of sample size, our results further confirm the findings of the previous study. For example, the correct estimation proportion of ML models increased by 0.06 when the sample size rose from 250 to 1000 in the study of Goretzko & Bühner [14].

Overall, the results of this study imply that compared to statistical methods, using simulation data generated by the analysis model (e.g., the MIRT) to train ML models and applying them to do predictions can work relatively effectively for estimating the number of latent skills in online learning environments. This kind of operation can be generalized to other kinds of analysis models. For example, when practitioners believe that their real-world data fits the assumptions

of CDM, they can choose a suitable model of CDM to simulate data reflecting the data features of expected scenarios and train ML models to predict the number of attributes in the Q-matrix. This can also be used for MF in terms of predicting the number of ranks.

Several limitations of this study need to be acknowledged. First, the trained and tuned ML models were not tested by real data. The conclusions of simulation study heavily rely on the data-generation model and the settings of simulation features, so relevant findings should be confirmed further based on real data. Second, due to the constraints of computational power, the present preliminary study only covered partial simulation scenarios, and the number of simulated data was limited to one for each scenario, which may make the relevant conclusions less stable. Third, as mentioned above, the illustration was based on the MIRT, and whether the findings remain the same for CDM or MF still needs to be tested.

5. Conclusion

In this study, we used the MIRT to generate simulation data reflecting the data features of target scenarios and took the features from simulation data to train and test two ML models (i.e., RF and XGBoost) for the simple and complex structure. These two ML models were compared with selected statistical methods regarding their performance of predicting the number of latent skills. The preliminary results show that the ML models (with or without including results of statistical methods during the training stage) generally outperform statistical methods in terms of correct estimation proportions. Additionally, regarding the effects of simulation features, we find that raising missingness level and the number of samples leads to a falling-down and going-up trend respectively in the correct estimation proportions of most methods. To conclude, our result implies that compared to statistical methods, using simulation data generated by the selected analysis model to train ML models and further doing prediction can relatively improve the prediction of the number of latent skills and extend the current operation related to users' skill extraction.

Acknowledgements

This work was funded by Research Fund Flanders (FWO fellowship 1S38023N). We also acknowledge the Flemish Government (AI Research Program).

References

[1] A. Gharahighehi, R. Van Schoors, P. Topali, J. Ooge, Adaptive Lifelong Learning (ALL), in: International Conference on Artificial Intelligence in Education, Springer Nature Switzerland, Cham, 2024: pp. 452–459.
[2] W. Bonifay, Multidimensional item response theory, Sage, 2020.
[3] M. von Davier, Y.S. Lee, Handbook of diagnostic classification models, Springer Publishing, 2019.
[4] M.C. Desmarais, Mapping question items to skills with non-negative matrix factorization, ACM SIGKDD Explorations Newsletter 13 (2012) 30–36. https://doi.org/10.1145/2207243.2207248.
[5] M.C. Desmarais, R. Naceur, A matrix factorization method for mapping items to skills and for enhancing expert-based Q-matrices, in: H.C. Lane, K. Yacef, J. Mostow, P. Pavlik (Eds.),

Artificial Intelligence in Education, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 441–450. https://doi.org/10.1007/978-3-642-39112-5_45.
[6] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151. https://doi.org/10.1177/001316446002000116.
[7] J. Braeken, M.A.L.M. Van Assen, An empirical Kaiser criterion., Psychol. Methods 22 (2017) 450–466. https://doi.org/10.1037/met0000074.
[8] J.L. Horn, A rationale and test for the number of factors in factor analysis, Psychometrika 30 (1965) 179–185. https://doi.org/10.1007/BF02289447.
[9] G. Raîche, T.A. Walls, D. Magis, M. Riopel, J.-G. Blais, Non-graphical solutions for cattell's scree test, Methodology 9 (2013) 23–29. https://doi.org/10.1027/1614-2241/a000051.
[10] W. Revelle, T. Rocklin, Very simple structure: an alternative procedure for estimating the optimal number of interpretable factors, Multivariate Behavioral Research 14 (1979) 403–414. https://doi.org/10.1207/s15327906mbr1404_2.
[11] J. De La Torre, C.-Y. Chiu, A general method of empirical Q-matrix validation, Psychometrika 81 (2016) 253–273. https://doi.org/10.1007/s11336-015-9467-8.
[12] W. Ma, J. De La Torre, An empirical Q-matrix validation method for the sequential generalized DINA model, Br. J. Math. Stat. Psychol. 73 (2020) 142–163. https://doi.org/10.1111/bmsp.12156.
[13] W.-S. Chin, Y. Zhuang, Y.-C. Juan, C.-J. Lin, A fast parallel stochastic gradient method for matrix factorization in shared memory systems, ACM Trans. Intell. Syst. Technol. 6 (2015) 2:1-2:24. https://doi.org/10.1145/2668133.
[14] D. Goretzko, M. Bühner, One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis., Psychological Methods 25 (2020) 776–786. https://doi.org/10.1037/met0000262.
[15] R.P. Chalmers, mirt: a multidimensional item response theory package for the R environment, Journal of Statistical Software 48 (2012). https://doi.org/10.18637/jss.v048.i06.
[16] R Core Team, R: A language and environment for statistical computing, (2024). https://www.R-project.org/.
[17] Y. Liu, F. Robin, H. Yoo, V. Manna, Statistical Properties of the *GRE* ® Psychology Test Subscores, ETS Research Report Series 2018 (2018) 1–13. https://doi.org/10.1002/ets2.12206.
[18] USMLE, 2024 USMLE bulletin of information, (2023). https://www.usmle.org/sites/default/files/2023-08/2024bulletin.pdf.pdf (accessed March 23, 2024).
[19] A. Robitzsch, sirt: Supplementary item response theory models, (2024). https://CRAN.R-project.org/package=sirt.
[20] D.G. Bonett, R.M. Price, Inferential methods for the tetrachoric correlation coefficient, J. Educ. Behav. Stat. 30 (2005) 213–225. https://doi.org/10.3102/10769986030002213.
[21] G. Raiche, D. Magis, nFactors: Parallel analysis and other non graphical solutions to the cattell scree test, (2022). https://CRAN.R-project.org/package=nFactors.
[22] William Revelle, psych: Procedures for psychological, psychometric, and personality research, (2024). https://CRAN.R-project.org/package=psych.
[23] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, Z.M. Jones, mlr: Machine Learning in R, Journal of Machine Learning Research 17 (2016) 1–5.
[24] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, J. Yuan, xgboost: Extreme gradient boosting, (2024). https://CRAN.R-project.org/package=xgboost.
[25] H.F. Kaiser, A second generation little jiffy, Psychometrika 35 (1970) 401–415. https://doi.org/10.1007/BF02291817.
[26] H. Dalton, The measurement of the inequality of incomes, Econ. J. 30 (1920) 348. https://doi.org/10.2307/2223525.

[27] S.-C. Kolm, The rational foundations of income inequality measurement, in: Handbook of Income Inequality Measurement, Springer, 1999: pp. 19–100.
[28] W. Guo, Y.-J. Choi, Assessing dimensionality of IRT models using traditional and revised parallel analyses, Educ. Psychol. Meas. 83 (2023) 609–629. https://doi.org/10.1177/00131644221111838.