Automated Identification and Validation of the Optimal Number of Knowledge Profiles in Student Response Data

Brad Din
Durham University
Durham, United Kingdom
bradley.p.din@durham.ac.uk

Tanya Nazaretsky
Weizmann Institute of Science
Rehovot, Israel
tanya.nazaretsky@weizmann.ac.il

Yael Feldman–Maggor
Weizmann Institute of Science
Rehovot, Israel
yael.feldman-maggor@weizmann.ac.il

Giora Alexandron
Weizmann Institute of Science
Rehovot, Israel
giora.alexandron@weizmann.ac.il

ABSTRACT

It is well–known that personalized instruction can enhance student learning. AI–based education tools can be used to incorporate blended learning in the science classroom, and have been shown to enhance teachers' ability to prescribe this personalization. We utilise cluster analysis to reveal student knowledge profiles from their response data. However, clustering algorithms typically require the number of clusters as a hyperparameter, yet there is no clear method for choosing the optimal number. Motivated by a practical instance of this foundational problem for a group–based personalization tool, this paper discusses several variations of the gap statistic to identify the optimal number of clusters in student response data. We begin with a simulation study where the ground truth is known to evaluate the quality of the identified methods. We then assess their behaviour on real student data and suggest a stability–based approach to validate our predictions. We identify an empirical threshold for the number of observations required for a prediction to be stable. We found that if a dataset had cluster structure, very small subsamples also showed cluster structure – large datasets were only required to discern the number of clusters accurately. Finally, we discuss how the method enables teachers to tailor their personalization according to their class environment or teaching goals.

Keywords

Clustering; Gap Statistic; Personalized Instruction

1. INTRODUCTION

In recent years, the increased usage of digital learning environments has led to the mass collection of student data [3]. The task of translating these data into tangible insights for understanding and improving student learning remains an active challenge. Blending technology into student learning and providing actionable analytics has massive potential to support teachers in adopting personalized pedagogy [4, 24, 32, 45]. Personalized instruction has been shown to significantly enhance learning outcomes by adapting various attributes of the learning procedure, such as the pace and the contents, to the specific needs of the individual students [6, 56]. The recent development of GrouPer, a learning analytics tool, has assisted teachers in implementing more personalized instruction [39]. The tool was co–designed with teachers and separates students into competency-based knowledge profiles. Whilst participating teachers acknowledged the power of personalization, they suggested that individual tailoring would be impractical in real K–12 classrooms, and that 'group–based personalization' would be a viable compromise between individual adaptation and frontal instruction, whilst also supporting social learning. In addition to competency–based profiling of the students, the teachers also requested semantic information explaining the knowledge profile that each cluster represents; providing this information has been shown to enhance teachers' ability to prescribe personalized learning sequences [39]. GrouPer with its group–based personalization strategy is currently being integrated into the PeTeL (**Pe**rsonalized **Te**aching and **L**earning) environment[1], allowing teachers to blend digital learning resources into their teaching and provide personalized pedagogy. Over 1000 physics, chemistry, and biology teachers have chosen to make the environment accessible to more than 12,000 students in real classrooms since 2018. In order to perform a sound analysis, GrouPer must first identify *how many* unique knowledge profiles a given activity contains. This is an instance of a fundamental problem – deciding on the number of clusters in a dataset. This is relevant for many applications in education [44, 46], such as discovering knowledge profiles, adaptive learning and student modelling [12, 13, 21, 22, 25, 29, 33, 34, 40, 50]. Despite this vast use, the issue of investigating ways to decide on the number of clusters in student response data was not studied in a systematic manner. This is the focus of the current work, which is motivated, as described above, by an actual EDM application.

B. Din, T. Nazaretsky, Y. Feldman-Maggor, and G. Alexandron. Automated identification and validation of the optimal number of knowledge profiles in student response data. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 458–465, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115744

[1] https://stwww1.weizmann.ac.il/petel/en/home-en/

2. BACKGROUND

In our application, the student responses to each activity are binary. The number of responses for each activity may vary from a few hundred responses to many thousands. The datasets are highly dimensional, where the number of questions ranges between 5 to 30. Combined with the inherent noise in human–based data [40], identifying cluster structure, if it exists, is significantly non–trivial. Unsupervised clustering learns the natural groups in a dataset from the raw data alone [20, 26]. This can be difficult, since there is no rigorous definition of a cluster [19]. Cluster analysis is used in a wide range of applications. Outside of education, it has found usage in image recognition [17], healthcare [30] and finance research [14], amongst many others. There are many algorithms in the literature, such as density–based clustering (e.g. DBSCAN [11]), distribution clustering (e.g. Gaussian mixture modelling [8]) and hierarchical clustering [37]. To avoid placing strict assumptions on our data structure, we choose the simple yet robust k–means algorithm [20, 31, 49].

The k–means algorithm takes a predefined number of clusters as a hyperparameter, k. One can initialise the cluster centroids randomly, or choose them strategically to avoid finding a local minima [54]. Each point is assigned to its nearest centroid; each centroid is then updated by taking the mean of all cluster members. This procedure is repeated until convergence. An alternative framework is the k–modes algorithm [7, 18], which updates the centroids by taking the mode of all members, retaining their binary nature. For our application, since there is no inherent meaning to the centroid, we use the more robust k–means algorithm, which we found to provide more reliable clustering than k–modes.

3. METHODOLOGY

A handful of methods exist in the literature to identify the optimum number of clusters within a dataset, denoted k^*. Classical statistical approaches (e.g. silhouette index [47]) have been used for many decades. X–means works alongside k–means to estimate k^* using information criteria [41]. Cluster prediction and validation methods have also been exploited [9]. Information theoretic approaches [51] and eigenvalue decomposition methods [16] have recently been implemented with success. However, the simple gap statistic has remained a consistent contender, and importantly does not require stringent assumptions to be made on the dataset. We follow the approach from Tibshirani [53], measuring the quality of clustering at each value of k. We use the Euclidean metric as a measure for the distance between two observations. For each cluster, we calculate the total distance between all members:

$$D_r = \sum_{i,i' \in C_r} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 = 2n_r \sum_{i \in C_r} \|\boldsymbol{x}_i - \boldsymbol{\mu}_r\|^2. \quad (1)$$

We reduce the complexity to $\mathcal{O}(n_r)$ by comparing each point to the cluster centroid, $\boldsymbol{\mu}_r$. Taking the sum over all clusters, we obtain the total within–cluster sum of squares (WSS):

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \quad (2)$$

As we increase the number of clusters, this quantity will monotonically decrease. After the optimal number, since all points are already close to a centroid, the total WSS plateaus, creating a sharp 'kink' at the optimum k. Methods of detecting this bend have been developed [48], but can be subjective, particularly for noisy data. To alleviate this, we utilise the gap statistic [53]; a comparison between the true sample data and its expectation under an appropriate null reference distribution, (W_k^*):

$$\mathrm{Gap}(k) = E\left[\log\left(W_k^*\right)\right] - \log\left(W_k\right). \quad (3)$$

We obtain $E\left[\log\left(W_k^*\right)\right]$ by taking the average of many binary bootstrapped samples. Finally, k^* is selected by considering adjacent values of the gap plot with the selection criterion:

$$k^* = \min_k \left\{\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}\right\}, \quad (4)$$

where $s_k = \mathrm{sd}_k \sqrt{1 + 1/B}$ and sd_k is the standard deviation of the bootstrap samples. In our work, we took $B = 280$, but we observed no significant difference with varying B. The gap statistic performs well when clusters are well-separated and uniform, but fails when the dataset becomes noisy. Prior work removed the logarithms in Eq. (3) [36]; we observed no benefit in doing so. Finally, the criterion in Eq. (4) is not robust; even if the plot has a clear optimum, the criterion fails to identify it correctly. We identify two methods to successfully overcome both of these issues: the weighted gap and DD–stopping criterion. The weighted gap approach [57] is identical to Tibshirani's approach, but modifies Eq. (2):

$$W_k^* = \sum_{r=1}^{k} D_r^* = \sum_{r=1}^{k} \frac{1}{2n_r(n_r - 1)} D_r. \quad (5)$$

This robust quantity D_r^* represents the averaged sum of the pairwise distances between all points in cluster r; this averaging reduces sensitivity to outliers. These statistics are interpreted as a comparison between a dataset and a truly unclustered distribution, which is crucial for identifying datasets with no cluster structure. However, the weighted gap statistic is also prone to overestimate the numbers of clusters, even if there is a clear optimum in the curve. We consider the alternative 'DD–stopping criterion' [57], which compares *adjacent* neighbours in the gap curve:

$$k^* = \max\left\{2\mathrm{Gap}(k) - \mathrm{Gap}(k-1) - \mathrm{Gap}(k+1)\right\}. \quad (6)$$

We have also used this criterion with the Tibshirani gap statistic. We therefore consider four methods: the gap statistic, the weighted gap statistic, and their DD–stopping criterion variants. Their typical outputs are shown in Fig. 1. We note that the DD–comparisons not only estimates the 'dominant' cluster structure, but also suggests multiple local maxima. The gap statistic can also produce local maxima [53]; we only obtained a single maximum in our applications.

On real student data, we do not know the ground truth. We begin with a simple study on five different structures of binary synthetic data. In all cases, the dataset will be a matrix of dimensions $n_\mathrm{s} \times n_\mathrm{f}$, where n_s is the number of student responses and n_f is the number of items within the activity. In the context of this study, we refer to the items as features of the model. The simplest structure, but perhaps most fundamental, is the case when the data has no inherent clustering (Model N). Here, the data is simply noise: we generate a matrix where each entry is uniformly chosen to be either 0 or 1. Well–defined cluster structure (Model WC) is generated by defining a matrix of correct responses and overlaying blocks

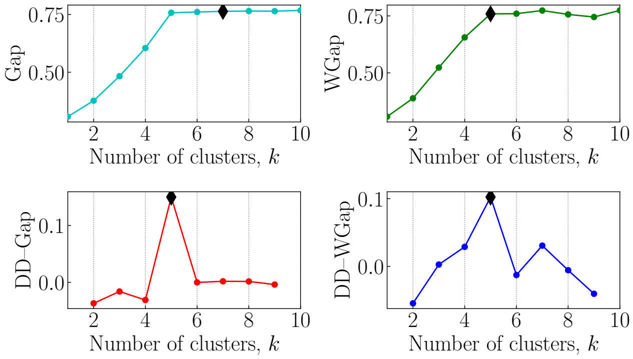

Figure 1: Outputs of the gap statistic, weighted gap statistic, and their DD–variants as a function of the hyperparameter k, shown for synthetic dataset R1 (Table 1).

Table 1: Predicted k^* for a selection of synthetic models. Predictions denoted 'F' failed to satisfy the selection criterion. DD–local maxima are in brackets. Here, $n_s^* = n_s/1000$.

Synthetic Model				k^* Prediction			
Model	n_f	n_s^*	k	Gap	WGap	DD–Gap	DD–WGap
N1	20	1	1	1	1	–	–
WC1	20	1	5	F	8	5	5
WC2	20	1	8	F	F	8	8
UWC1	20	1	5	6	F	5	5
UWC2	5	1	3	3	9	3	3
R1	20	1	5	7	5	5 (3, 5)	5 (5, 7)
R2	5	1	3	F	7	3 (3, 6)	3 (3, 7)
UR1	15	1	3	F	8	3	3 (3, 6, 8)
UR2	15	1	5	F	6	3 (3, 5)	2 (2, 4, 6)
UR3	15	10	5	F	F	5	2 (2, 4, 7)
UR4	15	1	8	F	F	6 (3, 6)	7 (4, 7)
UR5	20	1	5	F	F	5 (5, 8)	2 (2, 5, 7)
UR6	32	1	8	F	F	7 (3, 7)	2 (2, 5, 8)

of incorrect responses along the diagonal. We assume that different clusters have students who are weak in particular skills – a specific block of questions are assumed to measure a particular skill. For k evenly sized clusters, each block has dimensions of $n_\text{s}/k \times n_\text{f}/k$. To generate psuedo-realistic datasets with noise (Model R), we allow for the probability of students slipping ($P_\text{slip} = 0.1$) and guessing ($P_\text{guess} = 0.2$) [40]. We generate the background matrix where each entry has a probability of $1 - P_\text{slip}$ to be correct. We again overlay incorrect diagonal blocks but allow for the chance of guessing; each entry has a probability of $1 - P_\text{guess}$ to being incorrect. Finally, we impose uneven population distributions by defining the k^th triangle number, $k_\text{t} = k(k+1)/2$. Each cluster population has an increasing fraction of k_t; e.g. cluster n has n/k_t of the total population. This is utilised in the well clustered and realistic synthetic datasets, Models UWC and UR respectively. It is worth noting here that the number of features assigned to each cluster remains constant.

4. RESULTS ON SYNTHETIC DATA

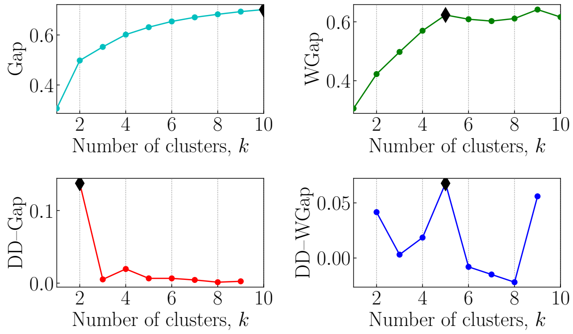

Figure 2: Outputs of the gap statistic, weighted gap statistic, and their DD–variants as a function of the hyperparameter k, shown for the real dataset P2 (Table 4).

A selection of results on synthetic data is shown in Table 1. On unclustered data (Model N), both the gap method and the weighted gap method are able to successfully identify unclustered data. However, on well–clustered data (Model WC), both methods predicted poorly; as can be seen in Fig. 1, the kink of the plot commonly occurs at the correct k, yet the stopping criterion proposed by Tibshirani is unsatisfactory. Both the gap and weighted gap methods were found to suffer from this problem, typically overestimating the number of clusters within the system. The DD–comparison methods were found to solve this issue, performing excellently for data with well–separated and compact clusters. The same results are found with uneven well–clustered data (Model UWC). On more realistic data (Model R), we see similar results. Again, the gap and weighted gap methods are unable to identify the correct number of clusters; however, they were able to identify that some cluster structure exists. The DD–comparison methods again performed well.

Finally, on the uneven realistic data (model UR), we see some interesting results. Model UR1, where each cluster had 5 features, provided the correct prediction. In Models UR2–4, the predicted value from the DD–models was not correct; we can gain some insight by interpreting the 'strength' of a cluster. Models UR2 and UR3 have 3 questions per cluster, whilst Model UR4 has between 1 and 2 questions per cluster. By comparing the labelling of students from the synthetic generation to the labels generated from the clustering, we found that the smallest clusters are prone to being mislabelled and 'absorbed' into the noise of others. Increasing the number of students within this smallest clusters has no effect, as seen in comparing Models UR2 and UR3. We conclude that the strength of a cluster with binary data is determined by the number of questions associated with each cluster – in Models UR5 and UR6, each cluster has 4 features within it and the method is now able to predict correctly. The DD–gap and the DD–weighted gap performed similarly. We therefore adopt a two–step approach: we first apply the gap or weighted gap method to discern if $k > 1$, and then use the DD–comparison method for determining the optimal number of clusters.

5. RESULTS ON STUDENT DATA

The student data considered here was collected from PeTeL activities in a mixture of subjects (Physics, Chemistry) and

Table 2: Predicted k^* for a variety of real student datasets.

Student Dataset			k^* Prediction	
ID Number	n_f	n_s	WGap	DD–WGap
P1	17	1572	4	2 (2, 4)
P2	18	726	5	5 (2, 5, 9)
C1	23	943	4	4 (2, 4, 7, 10)
C2	13	216	4	4 (2, 4, 6, 9)

subtopics (magnetism, forces). An example output on real student data is shown in Fig. 2. Since the signal to noise ratio is now lower, the original gap statistic curve is much shallower. Correspondingly, the DD–Gap method does not provide significantly meaningful predictions, typically finding the optimum number of clusters to be 2; we attribute this to the algorithm identifying the simple splitting of the students into strong/weak groups, which does not represent a meaningful pedagogical contribution. We therefore consider only the weighted gap and DD–weighted gap methods for the remainder of the paper. In Table 4, we show the results on real student datasets, with varying numbers of student responses and items in each learning activity.

Since we do not have a ground truth for these real datasets, we need to assess the validity of these predictions. If we receive a prediction that $k^* > 1$, how do we know that this k^* is correct (true positive)? Conversely, if we receive a prediction that $k^* = 1$, do we require more data (false negative), or does the activity have an inherent unclustered structure (true negative)? Both questions are addressed by considering the *stability* of our prediction. There are many methods of validating the stability of a cluster [9, 27, 52]; we utilise a resampling method used in similar approaches [28]. A stable cluster prediction is one that is similar under a small perturbation to the data (e.g. taking a subsample) [5, 55]. Many methods of cluster stability introduce some figure of merit, typically measuring the similarity between clusterings. We choose a simpler (but more practically–oriented) approach, and compare the predictions of the optimum number of clusters in the resampled dataset. In particular, since we are focusing on the DD–weighted gap method, we consider the predictions for the first 2 local maxima. This has a practical motivation; we do not want to provide teachers with a number of profiles that is too large to manage. We measure the validity of our clustering predictions by repeatedly taking fractional subsamples of our dataset and comparing the prediction results to those of the complete dataset. In order to address the second issue of true/false negatives, since we cannot collect more data, we instead take a dataset which has previously exhibited clustering (e.g. P1) and take subsamples of it. By taking successively smaller fractions, we attempt to identify some quantitative threshold for a 'sufficient' number of student responses.

In Fig. 3, we compare the predictions of the complete dataset to the predictions on three different fractional subsamples of the P1 dataset. Unsurprisingly, the positions of the first two maxima are identical for the largest fraction (90%, corresponding to 1415 students), indicating that the prediction we found was a stable one. We see that there is an increase in variance of the DD–weighted gap plots as we decrease the fraction to 50% (786 students), but the local maxima are again identical. Finally, when we take very small fractions, such as 10% (157 students), we observe significant variance in the DD–weighted gap curve itself, and the position of the local maxima now begin to vary. In Fig. 4, we present the predictions of the first and second local maxima from the DD–weighted gap as a function of the number of students for dataset P1. We infer the stability of each fractional subsample by indicating the frequency of anomalous observations from the complete dataset.

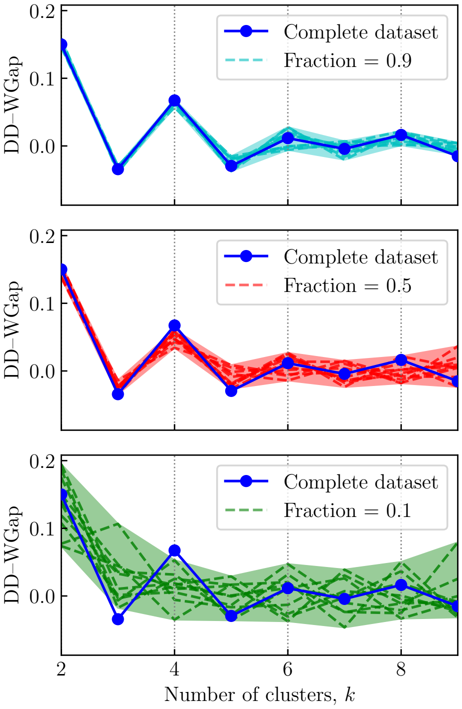

Figure 3: DD–weighted gap plots for three fractions of the P1 dataset: 90% (top), 50% (middle) and 10% (bottom), compared to the full dataset. Each fraction is sampled 10 times.

Our notion of stability allows a prediction on a smaller subsample to be considered stable if the difference is within ±1 of the prediction on the complete dataset, since we expect only a small change after making a small perturbation to the dataset. For the dataset shown in Fig. 4, we find that P1 has a threshold of 550 students. It is worth noting here that similar numerical thresholds were observed in the other clusterable datasets; C2 had a threshold of 660 students, P2 had a threshold of 653, and C3 was found to be unstable immediately. This latter result is not surprising given the small number of observations in the dataset, which is far below the threshold observed in other datasets. Perhaps the most interesting result we found is that the identification of cluster structure required only a remarkably small number of students. Explicitly, when taking 5% of the P1, P2, C2 or C3 datasets (with as few as 30 students), an overwhelming majority of the the weighted gap predictions were still that $k^* > 1$. Although the prediction of k^* in these small fractions was prone to extreme variation, the method was still

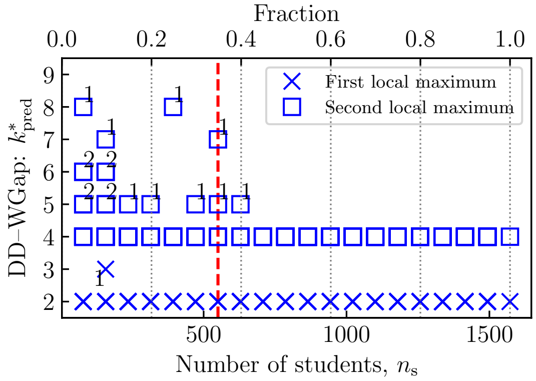

Figure 4: Predictions of the first (crosses) and second (squares) local maxima from the DD–weighted gap method, as a function of number of student observations (sample fraction), for the P1 dataset. If the observation of a fraction was different to that of the complete dataset, then the frequency of each anomalous observation is indicated.

able to confirm that *some* cluster structure existed; very few student responses were needed to discern if a dataset is clusterable. Specifically, it suggests that if an activity (with a reasonable number of responses) is predicted to have $k = 1$, then that particular activity likely *will not* have cluster structure. In this case, one should investigate the specific activity more closely, checking for any issues within the dataset and the data collection procedures itself.

6. DISCUSSION AND CONCLUSIONS

The results on real student data have demonstrated that our approach is able to provide reasonable predictions, proving to be robust even in the presence of noise. We have found that our approach is applicable for a wide range of learning activities. In particular, our stability verification results suggest that the number of responses is not a limiting factor in identifying if cluster structure exists. The method proposed is also completely generic in that it does not rely on any subject–specific knowledge. Although the interpretation of the clusters (e.g. as knowledge profiles) may vary between applications, we expect that this approach should be applicable as a generic tool for identifying cluster structure in a wide range of educational contexts.

A usability-oriented aspect that may influence our decision for the number of clusters is that, in reality, teachers may be constrained in the number of clusters that they are capable of treating simultaneously. This consideration provides a further secondary justification for why the DD–weighted gap method was selected. Providing teachers with multiple good clustering solutions allows them to choose how many clusters they want to work with enables the tool to be useful in a variety of situations; if there are additional teaching assistants in the classroom, or the activities require addtional care and attention, then the teacher may choose to split the class into more/fewer groups as required. Predictions on datasets with an insufficient number of responses will be inaccurate, but may only deviate by a couple of clusters. For our application, it could be argued that it is acceptable to provide teachers with a non-optimal recommendation. Moreoever, the tool is intended to be a recommendation, allowing teachers to override the suggestions if they deem it to be necessary – this is crucial for maintaining trust in the tool [38].

Applying this method in real environments requires careful data collection; it is very easy for a dataset to become very noisy. Some environments allow activities to be customized by teachers, enabling them to remove, modify or rearrange items, inserting inconsistencies into the data. Noise may also result from cheating, making responses unrepresentative of authentic student performance [1, 2]. Such sources of noise (amongst others) are typical for real educational applications [10, 15, 43, 58], and our process handled them in various ways (e.g., excluding activities modified by teachers). We note that the *theoretical* basis for an activity to be suitable for clustering is yet to be established and a better understanding of the types of assessment for which cluster analysis is theoretically justified is an interesting direction for future research. We expect clusters to exist in multi–dimensional activities that involve several binary skills (or skills with very steep learning curve) with some interconnections among them. However, in assessments that make the assumptions of IRT (normally distributed uni/multi–dimensional data), clusters may simply not exist.

In this work, we have evaluated common options for deciding on the optimum number of clusters within a dataset, and discussed their application on binary student data. We have compared these methods on synthetic data where the ground truth is known. We also found some insights into the factors determining the strength of a cluster; the number of features that comprise a cluster is important. This synthetic study formed the basis of our method applied to real student data; we discern if cluster structure exists by using the weighted gap method, and then subsequently determine the precise number of clusters using the DD–weighted gap method, as in [57]. We described an approach to validate the predictions from our method based on fractional resampling [28], and found an empirical threshold for the number of responses to have a stable prediction, typically around 500–600 student observations. Interestingly, we also found that if a data had cluster structure, then the existence of structure was observable with only a small handful of responses. This suggests that large datasets are only important in identifying the precise number of clusters. Our final contribution is the flexibility to the teachers, providing them with options of 'good' clustering solutions that they can apply according to the class environment and pedagogical goals. However, the challenge of providing pedagogically meaningful information about the strengths/weaknesses of each cluster is still outstanding. Methods of providing explanations of the knowledge profiles have already been studied in the literature, automatically building pedagogically meaningful explanations from item-level metadata [23, 35, 42].

7. ACKNOWLEDGEMENTS

BD's work was supported by the Engineering and Physical Sciences Research Council [EP/L01663X/1]. GA's research was supported by The Willner Family Leadership Institute for the Weizmann Institute of Science, Iancovici- Fallmann Memorial Fund, established by Ruth and Henry Yancovich, and by Ullmann Family Foundation. TN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

8. REFERENCES

[1] G. Alexandron, J. A. Ruipérez-Valiente, Z. Chen, P. J. Muñoz-Merino, and D. E. Pritchard. Copying@ Scale: Using harvesting accounts for collecting correct answers in a MOOC. *Computers & Education*, 108:96–114, 2017.

[2] G. Alexandron, L. Y. Yoo, J. A. Ruipérez-Valiente, S. Lee, and D. E. Pritchard. Are MOOC Learning Analytics Results Trustworthy? With Fake Learners, They Might Not Be! *International Journal of Artificial Intelligence in Education*, 29:484506, 2019.

[3] R. Baker and G. Siemens. Educational Data Mining and Learning Analytics. In *The Cambridge Handbook of the Learning Sciences*, pages 253–272. Cambridge University Press, Cambridge, UK, 2014.

[4] R. S. Baker. Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26(2):600–614, 2016.

[5] A. Ben-Hur and I. Guyon. Detecting Stable Clusters Using Principal Component Analysis. In *Functional Genomics: Methods and Protocols*, pages 159–182. Humana Press, Totowa, NJ, 2003.

[6] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16, 1984.

[7] F. Cao, J. Liang, and L. Bai. A new initialization method for categorical data clustering. *Expert Systems with Applications*, 36(7):10223–10228, 2009.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[9] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7):1–21, 2002.

[10] A. Dutt, M. A. Ismail, and T. Herawan. A systematic review on educational data mining. *IEEE Access*, 5:15991–16005, 2017.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[12] H. Gabbay and A. Cohen. Exploring the Connections Between the Use of an Automated Feedback System and Learning Behavior in a MOOC for Programming. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022*, pages 116–130, 2022.

[13] H. Gabbay and A. Cohen. Investigating the effect of automated feedback on learning behavior in moocs for programming. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, pages 376–383, 2022.

[14] M. C. Gupta and R. J. Huefner. A cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research*, 10:77–95, 1972.

[15] S. Gupta and A. S. Sabitha. Deciphering the attributes of student retention in massive open online courses using data mining techniques. *Education and Information Technologies*, 24(3):1973–1994, 2019.

[16] Z. He, A. Cichocki, S. Xie, and K. Choi. Detecting the Number of Clusters in n-Way Probabilistic Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2006–2021, 2010.

[17] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons, 1999.

[18] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.

[19] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[20] A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[21] T. Kabudi, I. Pappas, and D. H. Olsen. AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, 2:100017, 2021.

[22] T. Käser, A. G. Busetto, B. Solenthaler, J. Kohn, M. v. Aster, and M. Gross. Cluster-based prediction of mathematical learning patterns. In *International conference on artificial intelligence in education*, pages 389–399. Springer, 2013.

[23] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gaevi. Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.

[24] J. King and J. South. Reimagining the role of technology in higher education: A supplement to the national education technology plan. *US Department of Education, Office of Educational Technology*, 2017.

[25] S. Klingler, T. Käser, B. Solenthaler, and M. Gross. Temporally coherent clustering of student data. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)*, pages 102–109, 2016.

[26] S. Križanić. Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management*, 12:1847979020908675, 2020.

[27] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323, 2004.

[28] E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

[29] C. Li and J. Yoo. Modeling student online learning using clustering. In *Proceedings of the 44th Annual Southeast Regional Conference*, ACM-SE 44, page 186191, New York, NY, USA, 2006. Association for Computing Machinery.

[30] M. Liao, Y. Li, F. Kianifard, E. Obi, and S. Arcona. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC nephrology*, 17(1):1–14, 2016.

[31] J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the*

5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967.

[32] R. Martinez-Maldonado, A. Clayphan, K. Yacef, and J. Kay. MTFeedback: Providing Notifications to Enhance Teacher Awareness of Small Group Work in the Classroom. *IEEE Transactions on Learning Technologies*, 8(2):187–200, 2015.

[33] P. Mejia-Domenzain, M. Marras, C. Giang, and T. Käser. Identifying and Comparing Multi-dimensional Student Profiles Across Flipped Classrooms. In *Artificial Intelligence in Education: 23rd International Conference, AIED 2022*, pages 90–102, 2022.

[34] A. Merceron and K. Yacef. Clustering students to help evaluate learning. In *IFIP World Computer Congress, TC 3*, pages 31–42, 2004.

[35] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[36] M. Mohajer, K.-H. Englmeier, and V. J. Schmid. A comparison of gap statistic definitions with and without logarithm function, 2011.

[37] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[38] T. Nazaretsky, M. Ariely, M. Cukurova, and G. Alexandron. Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4):914–931, 2022.

[39] T. Nazaretsky, C. Bar, M. Walter, and G. Alexandron. Empowering Teachers with AI: Co-Designing a Learning Analytics Tool for Personalized Instruction in the Science Classroom. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 1–12, 2022.

[40] T. Nazaretsky, S. Hershkovitz, and G. Alexandron. Kappa learning: A new item-similarity method for clustering educational items from response data. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 129–138, 2019.

[41] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734, 2000.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 11351144, New York, NY, USA, 2016. Association for Computing Machinery.

[43] C. Romero, J. R. Romero, and S. Ventura. A survey on pre-processing educational data. In *Educational data mining*, pages 29–64. Springer, 2014.

[44] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[45] C. Romero and S. Ventura. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[46] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355, 2020.

[47] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[48] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171, 2011.

[49] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178, 2010.

[50] R. P. Springuel, M. C. Wittmann, and J. R. Thompson. Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics. *Phys. Rev. ST Phys. Educ. Res.*, 3:020107, Dec 2007.

[51] C. A. Sugar and G. M. James. Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463):750–763, 2003.

[52] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[53] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[54] S. Vassilvitskii and D. Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

[55] U. Von Luxburg et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274, 2010.

[56] C. A. Walkington. Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of educational psychology*, 105(4):932, 2013.

[57] M. Yan and K. Ye. Determining the Number of Clusters Using the Weighted Gap Statistic. *Biometrics*, 63(4):10311037, 2007.

[58] N. Z. Zacharis. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, 27:44–53, 2015.

APPENDIX

A. ADDITIONAL SYNTHETIC DATA RESULTS

Table 3: Predicted k^* for a selection of synthetic models. Predictions denoted 'F' failed to satisfy the selection criterion. Predictions marked by † were incorrect despite having a clear optimum. DD–local maxima are in brackets. Here, $n_s^* = n_s/1000$.

Synthetic Model				**k^* Prediction**			
Model	n_f	n_s^*	k	Gap	WGap	DD–Gap	DD–WGap
N1	20	1	1	1	1	–	–
N2	20	10	1	1	1	–	–
N3	5	1	1	1	1	–	–
WC1	20	1	5	F†	8†	5	5
WC2	20	1	8	F†	F†	8	8
WC3	8	1	3	F†	3	3	3
WC4	6	1	2	F†	3	2	2
WC5	6	10	2	F†	F†	2	2
UWC1	20	1	5	6†	F†	5	5
UWC2	5	1	3	3	9†	3	3
UWC3	15	1	8	F†	F†	8	8
R1	20	1	5	7	5	5 (3, 5)	5 (5, 7)
R2	5	1	3	F	7	3 (3, 6)	3 (3, 7)
R3	15	1	8	F	F	8 (3, 5, 8)	8 (3, 5, 8)
R4	15	10	8	F	F	8 (3, 5, 8)	8 (3, 5, 8)
UR1	15	1	3	F	8	3	3 (3, 6, 8)
UR2	15	1	5	F	6	3 (3, 5)	2 (2, 4, 6)
UR3	15	10	5	F	F	5	2 (2, 4, 7)
UR4	15	1	8	F	F	6 (3, 6)	7 (4, 7)
UR5	20	1	5	F	F	5 (5, 8)	2 (2, 5, 7)
UR6	32	1	8	F	F	7 (3, 7)	2 (2, 5, 8)

B. ADDITIONAL REAL DATASET RESULTS

Table 4: Predicted k^* for a variety of real student datasets.

Student Dataset			**k^* Prediction**	
ID Number	n_f	n_s	WGap	DD–WGap
P1	17	1572	4	2 (2, 4)
P2	18	726	5	5 (2, 5, 9)
C1	23	943	4	4 (2, 4, 7, 10)
C2	13	216	4	4 (2, 4, 6, 9)
C3	14	292	1	–
C4	14	379	1	–
C5	14	300	1	–
C6	13	241	1	–