Information Extraction from Binary Skill Assessment Data with Machine Learning

https://doi.org/10.3991/ijai.v3i1.24295

Susanne Jauhiainen (✉)
University of Jyväskylä, Jyväskylä, Finland
susanne.m.jauhiainen@jyu.fi

Tron Krosshaug
The Norwegian School of Sport Science, Oslo, Norway
Oslo Sports Trauma Research Center, Oslo, Norway

Erich Petushek
Michigan Technological University, Houghton, MI, US

Jukka-Pekka Kauppi,
University of Jyväskylä, Jyväskylä, Finland

Sami Äyrämö
University of Jyväskylä, Jyväskylä, Finland

Abstract—Strength training exercises are essential for rehabilitation, improving our health as well as in sports. For optimal and safe training, educators and trainers in the industry should comprehend exercise form or technique. Currently, there is a lack of tools measuring in-depth skills of strength training experts. In this study, we investigate how data mining methods can be used to identify novel and useful skill patterns from a binary multiple choice questionnaire test designed to measure the knowledge level of strength training experts. A skill test assessing exercise technique expertise and comprehension was answered by 507 fitness professionals with varying backgrounds. A triangulated approach of clustering and non-negative matrix factorization (NMF) was used to discover skill patterns among participants and patterns in test questions. Four distinct participant subgroups were identified in data with clustering and further question patterns with NMF. The results can be used to, for example, identify missing skills and knowledge in participants and subgroups of participants and form general and personalized or background specific guidelines for future education. In addition, the test can be optimized based on, for example, if some questions can be answered correct even without the required skill or if they seem to be measuring overlapping skills. Finally, this approach can be utilized with other multiple choice test data in future educational research.

Keywords—Data mining, Clustering, Non-negative matrix factorization, Strength training skill test, Binary data

1 Introduction

Data collections in educational settings provide unique types of data that offer many opportunities for data mining to extract useful information [1]. This information can be used, for example, to better understand person's skills, how and in which settings they learn, and to recognize training and development needs. While large data in its raw form rarely offers easily interpretable information, data mining can be used to detect novel or unsuspected patterns and relationships in the data as well as to summarize data in novel and understandable ways [2].

Due to the abilities of data mining, it has received increasing amounts of attention in many domains, including educational research, where the term educational data mining (EDM) has risen [3]. For example, previous studies have focused on predicting the performance of a student [4], [5], analyzing student motivation and attitudes [6], [7], classifying and clustering different learning styles and behaviours [8]–[10], and recommending best courses for students [11], [12].

Some of the common objective ways to measure person's skills in educational research are testing, and curriculum-based measurement (CBM). Testing is the most widespread way of assessing person's skills [13]. A widely known example in the educational domain is the Programme of International Students Assessment (PISA) that is a standardized test developed to measure 15-year-olds' skills in domains such as reading, mathematics, and science. CBM is a simple yet very effective way for educators to track student's skill levels and development both on individual and group levels [14]. Multiple choice questions (MCQs) can be used in both objective and subjective skill assessments and are common in many fields. For example, in medicine MCQs are considered as the most widely applicable and useful form of testing [15] and are increasingly used in physics education as well [16]. Well-constructed MCQs by themselves can assess higher cognitive skills, such as interpretation, concept understanding, and knowledge application, and are reliable and cost-effective [15], [17], [18]. Most often MCQ comprise of a set of options with one correct answer that matches the question and other options, called distractors [15], [17]. Therefore, MCQ answers can be easily, and often are, transformed to binary format for data analysis.

Essentially, binary data is categorical data with two possible values, most often labeled as zeros and ones. Many phenomena can be directly measured with binary variables, such as the presence/absence of something or success/failure in something [19]. Moreover, categorical as well as continuous data can be easily simplified to a binary format. For example, above mentioned MCQs can be transformed to correct/incorrect, questions measured in nominal scale can be one-hot-encoded, and continuous data can be categorized into classes (e.g., categorizing continuous test score as high-performing or low-performing) [20].

Educational data is commonly utilized in the field of sports and physical education as well. For example, the relationship between physical activity and school performance has been an excessively discussed topic [21] and the learning process of sport coaches has been of interest to researchers [22]. Strength training exercises are essential in sports, rehabilitation and for improving our health. It is therefore of high importance

that people working in the strength training and conditioning related industries comprehend and master exercise techniques in order for them to safely and effectively teach and train other people. Currently, there are no available tools for measuring in-depth skills of strength training experts, and there is a severe lack of biomechanical educational material. We therefore developed the current test to investigate the knowledge level of strength training experts.

The test included typical skill domains that are commonly believed to be of important for strength training experts, i.e. anatomy, muscle activation, modification of exercises, and last but not least, biomechanics (forces and moment arms) [23]. We included multiple questions that represented the different skill domains, for a robust identification of the skill of each participant. Thus, the main aim of this study was to investigate how data mining methods can be used to identify novel skill patterns and what type of knowledge that might be missing among strength training experts, based on a MCQ binary skill test. The secondary aims were to investigate whether the discovered knowledge could be used to recognize specific education needs among the participants or optimize the test.

2 Data

The initial test was developed by two experienced practitioners and researchers in the fields of biomechanics and human factors. These initial items were further carefully examined by domain experts from a group of highly skilled researchers and practitioners in the field of strength and conditioning. Questions were modified based on this feedback and a final test of thirty items was included for testing. Items include both MCQs and true/false questions, with both images as well as text-only questions. In two of the questions, participants were supposed to select two correct answers instead of just one. The skill test included thirty questions and was completed online. The questions can be seen in Appendix A.

The participants consisted of 507 fitness professionals, including, for example, physical therapists, academics and sport science students. They were recruited online through personal and social media networks. The participants ranged between 18 and 61 years (mean 31.25±7.71) with 78% of males and 22% of females. The background information is summarised in Table 1. They have a variety of different backgrounds in education and experience and come from forty different countries and six different continents with 32.3% coming from Norway. Large majority of participants have a graduate degree.

2.1 Data processing

The answers were encoded as correct (=1) and incorrect (=0), resulting into a binary matrix of size 507x30. Altogether 2,1% of the values were missing in the data. A missing answer was interpreted as incorrect and these were imputed with zero values. In addition, we also investigated how dimension reduction with Principal Component Analysis (PCA) before clustering affected the clustering results. In addition to cluster-

Table 1. The proportion of participants with certain education level, work experience, and nationality.

Education level		**Work experience**		**Nationality**	
High school	9.1%	0-1yr	15.8%	Norway	32.4%
Bachelor	35.1%	2-5yr	42.4%	UK	16.0%
Masters	38.9%	6-10yr	24.1%	Other Europe	14.6%
Phd	7.1%	11-20yr	10.9%	USA + Canada	21.3%
Other	6.1%	20+yr	4.5%	Asia	2.8%
No info	3.8%	No info	2.4%	Other	7.9%

ing participants based on the original data, a number of PCs were chosen to represent the questions and results compared. The number of PCs was chosen so that at least 90% of the variance was explained [24]. All data preprocessing and analysis were done in MATLAB 2018b (MathWorks Inc).

3 Methods

In this study we used methodological triangulation [25], [26] to search for meaningful patterns in the data. Triangulation means that multiple approaches are used to assess the research objective and their results are combined for more reliable results [26], [27]. It can be divided into data, investigator, theoretical, and methodological triangulation where, respectively, either multiple different data sets, researchers, theoretical positions, or methods are utilized in analysis [25]. We approached the data from two different viewpoints for more confident and interpretable results. First, clustering was used to discover groups of participants with similar answer patterns and both participant and question patterns were assessed based on the results. Second, NMF was used to assess patterns among test questions and further interpret the discovered clusters. For the following method definitions, let us define the data matrix as $X \in \mathbb{R}^{n \times \mathrm{d}}$, where n is the number of rows (i.e., participants) and d is the number of columns (i.e., questions).

3.1 Clustering

Clustering is an unsupervised method to divide data observations into distinct groups [28]. In this study, prototype-based clustering methods, namely k-means and k-medoids, were used [29]. In prototype-based clustering, the goal is to partition the data directly into a given number of k clusters, which are represented by prototypes [30]. First, k cluster prototypes are initialized and each observation is assigned to the closest prototype. After the assignment, the prototypes are recomputed and then the assignment and recomputation steps are repeated until the prototypes do not change anymore or a user-defined stopping criterion is reached. In general, this can be formalized as a minimization problem [31]:

$$\min_{\{b_k\}_{k=1}^{K}} J\left(\{\boldsymbol{b}_k\}\right), \tag{1}$$

where

$$J(\{\boldsymbol{b}_k\}) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left|\boldsymbol{x}_i - \boldsymbol{b}_k\right|\right|_p^q = \sum_{k=1}^{K} J_k \tag{2}$$

Here k is the number of clusters, x_i is an observation assigned to cluster k, b_k is the kth cluster prototype, and $\left|\left|x\right|\right|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ is the L^p-norm. J_k is the clustering error of cluster k and J is the error sum over k clusters, so the total clustering error being minimized. By choosing $p = 2$ and $q = 2$ for equation (2) we obtain the k-means optimization problem, where the cluster prototype is the mean of the observations in each cluster.

K-medoid algorithm [32], in contrast to k-means, does not use a summary metric as cluster prototype but chooses one of the observations in the cluster to represent the prototype instead. It has been suggested as a better alternative for clustering binary data [33]. The cluster prototype minimizes the within cluster error and the minimization problem can be formalized as

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left|\boldsymbol{x}_i - \boldsymbol{m}_k\right|\right|_p^q , \tag{3}$$

where $\boldsymbol{m}_k \in \boldsymbol{x}_i$ is the prototype (medoid) of the kth cluster and with $p = 2$ and $q = 2$ the squared Euclidean distance is used to measure dissimilarity of point x_i to the medoid of its cluster. K-medoids is more robust than k-means which can be useful when there is substantial noise or outliers in the data. The most common approach to k-medoids is Partitioning Around Medoids (PAM) [34], that is a greedy search algorithm trying to minimize the dissimilarities between the cluster prototype and other observations in the same cluster. The *kmeans*- and *kmedoids*-functions in the *Statistics and Machine Learning Toolbox* in MATLAB were used. They solve the clustering problem using an iterative method and use the k-means++ algorithms for cluster prototype initialization.

Clustering stability. Different cluster initializations can lead to different clustering solutions, causing instability of the results [35]. Assessment of clustering stability can be used to select the best number of clusters [35] and the clustering model. Stability of clustering with k-means and k-medoids were assessed with different values of k between one and twenty with the Rand index [36] to choose the number of clusters and suitable clusterig method. Rand index is an objective measure for similarity of two data partitions, $P = \{P_1, P_2, \ldots, P_{K_1}\}$ and $P' = \{P'_1, P'_2, \ldots, P'_{K_2}\}$, for the set of data observations $X = \{x_1, x_2, \ldots, x_n\}$, where K_1 and K_2 are the number of subgroups in partitions P and P' respectively. It looks at all possible pairs of observations in the data matrix X. If we define s as the number of pairs that are clustered to the same subgroup in both P and P', and d as the number of pairs that are not clustered to the same subgroup in either P or P', the Rand Index is calculated as $R = s + \frac{d}{\binom{n}{2}}$), where the denominator is the total number of pairs in X. In practice, Rand index measures the proportion of similar

pairings, over all possible pairs of observations. The index receives a value between 1 and 0, with 1 indicating the clusterings are exactly the same while 0 indicates that clustering do not agree on any parts.

For each k, a reference clustering partition P was first calculated and then the clustering was repeated a hundred times for hundred partitions P'. Rand index values were calculated between P and each P' and finally averaged over the hundred values for final value measuring the stability of clustering.

Cluster comparisons. The background information of the participants (e.g., nationality, education level, work experience, and sex) and correct answers in each cluster were compared with chi-squared test with significance limit of $\alpha = 0.05$. When significant differences were observed, pair-wise post-hoc tests were performed between all clusters with chi-squared tests and the significance adjusted with Bonferroni's correction.

3.2 Non-negative matrix factorization

Non-negative matrix factorization (NMF) [37] is an intuitive method for non-negative data that has become widely used in the machine learning and data mining fields [38]. It can be applied in many use cases, including, feature extraction, signal processing, dimension reduction as well as text and image analysis [37], [39]–[41]. NMF factorizes a non-negative data matrix $X \in R^{n \times d}$ into two non-negative lower rank matrices $W \in \mathbb{R}^{n \times \mathrm{r}}$ and $H \in \mathbb{R}^{r \times \mathrm{d}}$, so that $X \approx WH$ and $r < \min\{n, d\}$. The matrices W and H are approximated by solving the following problem:

$$\min_{W,H} \; ||X - WH||_F, \qquad W, H \geq 0, \tag{4}$$

where $||\cdot||_F$ is the Frobenius norm. The elements in W are the basis elements of the latent lower rank space, while H includes the corresponding coefficients. With $r < \min\{n, d\}$, NMF reduces the dimension of data, working very similarly to PCA with the most distinctive feature being non-negativity [37]. NMF also has an inherent clustering property [42] and H can be used to cluster the columns of the input data X. If we would further constraint the orthogonality of H in equation (4), the minimization problem would be equivalent to that of k-means clustering [39]. In our study, NMF was run with the *nnmf*-function in the *Statistics and Machine Learning Toolbox* 11.4 in MATLAB 2018b (MathWorks Inc). It solves the problem using an iterative method, starting with random initializations for W and H.

To further interpret the NMF results, components were visualized with a biplot that allows representing information about both the variables and observations in the same figure. For this visualization $r = 2$ was chosen for NMF and clustering was also utilized for visualization of participants. An additional biplot was plotted with grouping based on pure performance to assess the relationship between test performance and the cluster division.

4 Results

4.1 Clustering

For k-means, smaller numbers of clusters were more stable (i.e., least variation in the mean Rand index across repetitions) and dimension reduction with PCA beforehand did not seem to affect the stability of clustering results (see Figure 1). For k-medoids, on the other hand, the stability varied largely between different numbers of clusters and results with PCA were slightly different from those without. In general, for larger numbers of k, clustering was far more stable for k-medians than k-means and these clustering results are slightly more stable without PCA (see Figure 1). Based on the aforementioned observations, we chose to cluster with k-medoids, without PCA dimension reduction, and with number of clusters k = 4. We also investigated the case of k = 2 clusters, which was similarly stable but omit the results as they do not bring any added information. In this case, cluster one consisted roughly of participants in clusters one (C1) and two (C2) from the k = 4 case and similarly cluster two consisted of participants in clusters three (C3) and four (C4) from the k = 4 case. The high Rand index value for k = 4 suggests that this further division pattern is present in our data.

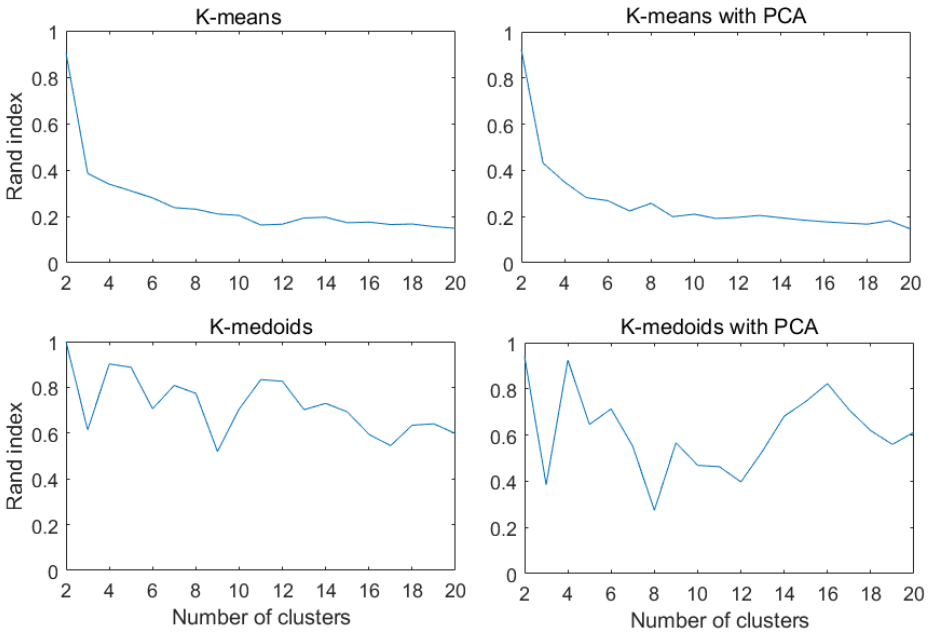

Fig. 1. Mean Rand index between the reference clustering and hundred repetitions. Higher Rand index values indicate more stable clustering as the results of two separate clusterings agree better.

In the chosen k-medoid model, C1 (n=149) had average of 60.56%, C2 (n=161) had an average of 49.05%, C3 (n=104) had an average of 41.15%, and C4 (n=93) had an

average of 39.71% correct answers. For reference, random guessing in the test would result in approximately 30% correct answers. Thus 40% is only slightly better than pure guessing, whereas 50% and 60% are two and three times better, respectively. Background information for these clusters is summarized in Figure 2. The distribution of participants from different countries is different in C1 compared to all three other clusters (C1 vs C2 p = 0.008, X^2 = 20.00, C1 vs C3 p = 0.004, X^2 = 20.49, C1 vs C4 p = 0.002, X^2 = 24.73). Based on Figure 2, there are clearly more participants from Norway than any other country in cluster C1 that represents the participants with the highest level of total scores. Participants in C1 had a higher education level compared with C4 (p = 0.040, X^2 = 16.05) (see Figure 2). Cluster C1 had participants with longer experience than C4 (p = 0.008, X^2 = 19.96) and C3 (p = 0.037, X^2 = 16.26) (see Figure 2). C1 had a significantly higher proportion of males compared to cluster C4 (p = 0.01, X^2 = 9.92) (see Table 2). There were no differences in the number of working days per week between the clusters.

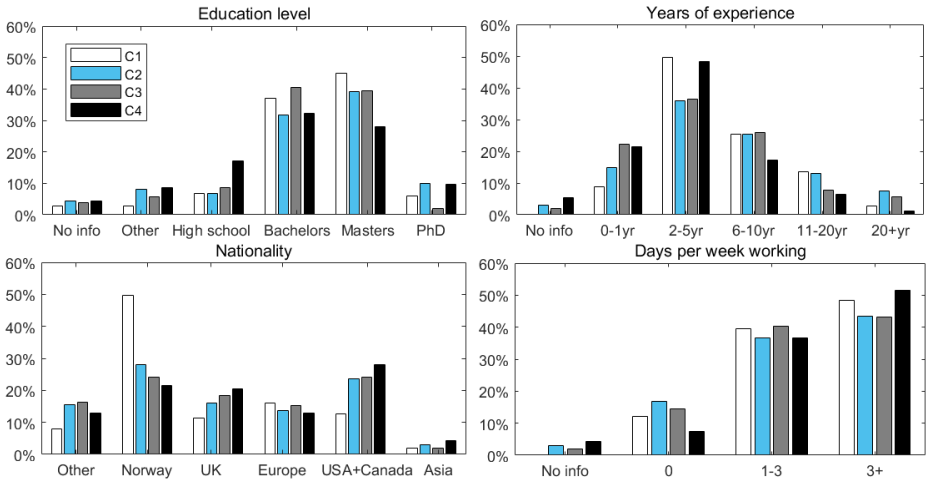

Fig. 2. Distribution of various background factors in the four clusters.

Table 2. Proportion of males and females in each cluster.

	Male	**Female**
C1	84.46%	15.54%
C2	81.41%	18.59%
C3	73.53%	26.47%
C4	67.03%	32.97%
Total population	78.07%	21.93%

From the distribution of correct answers among clusters (Figure 3), a number of patterns can be detected. It can be seen that some questions, i.e. those where a complete biomechanical analysis was necessary, were difficult for all participants (e.g., 3, 6, 25, 26, and 27) while others, e.g. those that required understanding of the mechanics of an

elastic band, or experience from how different variations of an exercise will load a muscle, were easier (e.g., 12 and 13). Some questions also have clear differences between the clusters, for example, the participants in the best performing cluster C1 performed clearly better than the rest in questions 5, 8, and 16. These questions require good biomechanical understanding. In addition, while the participants in C4 performed less good than the others overall, they performed better than all the others in question 19 and relatively well in questions 18, 20, and 22.These questions were all related to having a long thigh bone, but concerned several different aspects, e.g. anatomy, geometry and biomechanics.

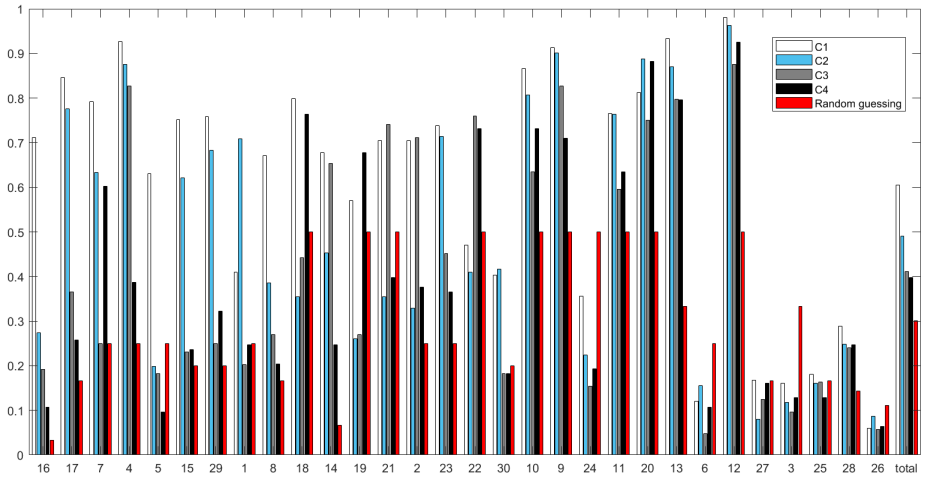

Fig. 3. Percentage of participants that answered a question correct, separately for each cluster. Question numbers on the x-axis and percentage of correct answers in cluster in the y-axis. Questions are sorted descending based on the difference between best and worst cluster performance. The red bar corresponds to probability of guessing the correct answer.

4.2 Nonnegative matrix factorization

Discovered NMF patterns are visualized together with the discovered clusters in Figure 4. From the cluster scatters, we can see that the better performing clusters C1 and C2 load higher on NMF component one, while clusters C3 and C4 load higher on component two. This means that questions that load high on component one, were better answered by participants in C1 and C2 and on the other hand those loading higher on component two were better answered in C3 and C4. As can be seen in the figure, clusters C1 and C2 as well as clusters C3 and C4 cannot really be separated from each other in the two-dimensional case.

From the biplot of NMF coefficients, W, it is easy to recognize groups of similar questions in data (Figure 4). For example, questions 26, 27, 6, 3, 25 (require thorough biomechanical understanding) load the first two NMF components similarly, both component one and two with very low values. As can be seen in Figure 3 these are all

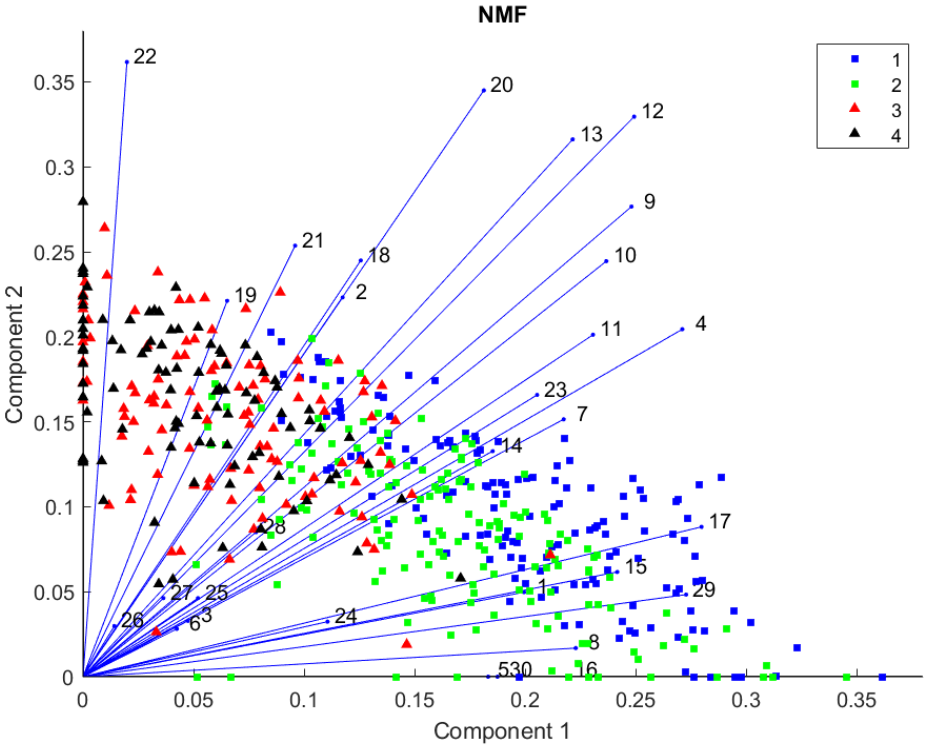

Fig. 4. Biplot of the first two NMF components. The participants who answered are scattered with colors indicating the cluster number. The numbers correspond to the thirty questions.

questions that were mostly answered incorrectly, i.e., receiving zero values on data and not able to separate the clusters. In addition, questions 28 (similar to questions 25, 26, and 27) and 24 (requires similar skills to question 8) are very close to the above-mentioned group of questions. Another recognizable group of questions in Figure 4 is 18, 19, and 21 (related to having a long thigh bone, but require versatile expertise) and 2 (requires basic biomechanical understanding). These are questions where C1 and either C3 (2, 21) or C4 (18, 19) did clearly better than the other two. Question 22, on the other hand, seems to be different from all the others, loading high on component two and low on component one. This is a question that is more often answered correctly by the participants in lower performing clusters C3 and C4 (Figure 4).

On the bottom right part of Figure 4, another group with questions 1, 5, 8, 15, 16, 17, 29, and 30 can be detected. These are questions that require basic biomechanical understanding, i.e. how external load will induce muscle activation, and were in general answered better in the higher performing clusters C1 and C2. Questions 4, 7, 11, 14, and 23 are related to different concepts; anatomy, stability, basic understanding of moments and more advanced biomechanical analysis. The questions were mostly related to squat movements (7, 11, and 14) and were answered relatively poorly in either C3 or C4 or both. The rest of the questions were mostly related to squatting exercise (9, 10,

12, 13, 20) with mostly true/false options and were in general answered correctly across the whole participant sample.

Groups based on test performance. To compare the formed clusters to the general performance of participants, the data was also divided into two groups based on participant's performance level. A biplot for these two groups can be seen in Figure 5. While the scatter of these two performance groups resembles the group division in clustering (Figure 4), clear differences can also be detected. In general, both performance groups load on both NMF components and cannot be separated that well in this 2D case. This indicates that the clustering results are not only based on the general performance, but other patterns are detected in the data as well.

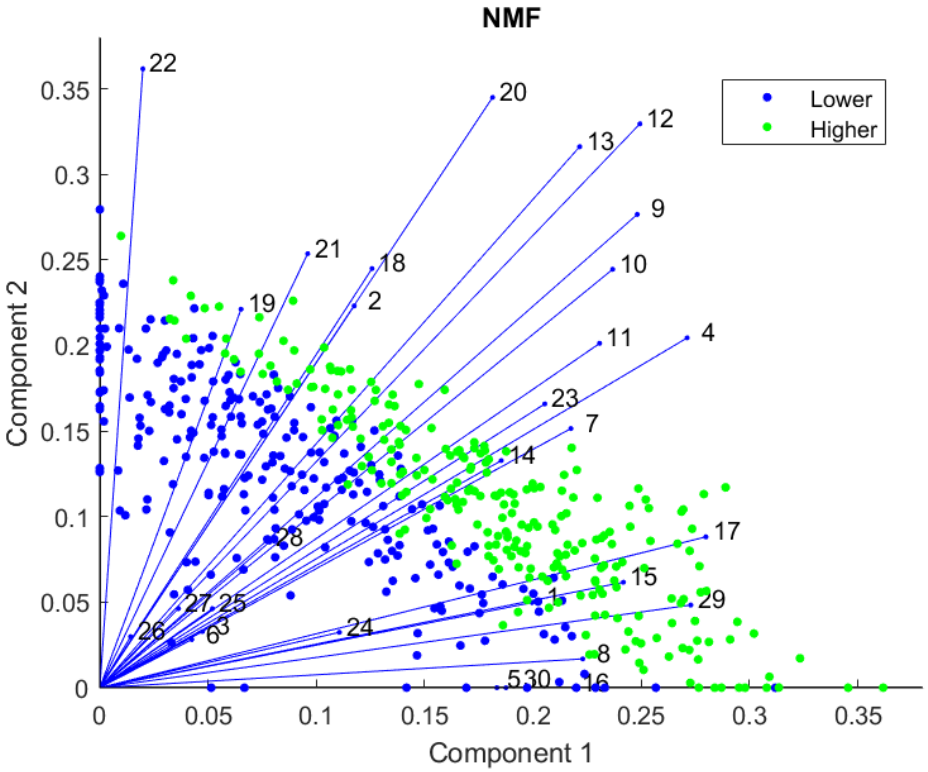

Fig. 5. Biplot of the first two NMF components. The people who answered are scattered with colors indicating the performance group. Numbers correspond to the thirty questions.

5 Discussion

This paper investigated how data mining methods can be used to identify novel skill and question patterns from educational skill test data. We utilized two methods in a triangulated manner to extract and summarize information from binary MCQ data

measuring strength training and conditioning related skills. Clustering was used to discover groups of participants with similar answer and skill patterns and NMF was used to assess patterns among test questions and to further interpret the discovered clusters.

Four distinct participant clusters were identified from the data. These groups were partly but not completely based on the general test performance, meaning that unique skill and answering patterns were present in data as well. C1 was the highest performing cluster with 60.56% correct answers on average and had a prominent proportion of Norwegian participants as well as a people with 2-5 years of experience. The higher performance of Norwegian participants might party be explained by them being exposed to similar material as what was tested previously. In addition, 2-5 years of experience might be related to better performance than longer experience due to more recent and updated education if those with longer experience have not been actively updating their knowledge and have more outdated or myth-based information. They performed clearly better than the rest in questions 5, 8, and 16, that require a good biomechanical understanding. C2 was the largest cluster with on average 49.05% of correct answers. These participants performed notably better than the others, or the level of random guessing, in question 1. In general, they performed worse than others in questions requiring basic understanding of moment arms (2, 19, and 21) and in question 18 requiring understanding of body configuration and anatomy. C3 and C4 both performed poorly overall with an average of 41.15%, and 39.71% correct answers, respectively. The participants in C3 performed poorly at almost all questions related to biomechanics but better in questions 2, 14, 21, and 22, related to understanding of anatomy and movement/joint configuration. C4 had poor basic anatomy knowledge but performed well in questions 18, 19, 20, and 22 related to different aspects associated with having a long thigh bone, e.g. anatomy, geometry and biomechanics. The cluster had the largest proportion of people working more than 3+ days a week evaluating exercise technique and thus they might be exposed to these applied skills in their daily work but then lack some more profound biomechanics understanding required to reason well in questions overall.

NMF was able to discover multiple interesting question groups and patterns. Questions 3, 6, 24-28 were answered very poorly among all participants. The clear tendency to answer question 3 incorrectly is likely related to a misconception/myth that has been established in the training industry, where only muscle activation related to hand separation has been measured, without considering if they are placed high/low or whether or not the elbows point in or out from the body. Question 6 shows that people are not able to do biomechanical analysis of the bench press. This is not surprising as all the published biomechanical analyses in the literature has been flawed [43]. In general, the participants did not seem to be aware of the fact that forces may have both medial and lateral components, which makes them choose other alternatives. Questions 9, 10, 12, 13, and 20 on the other hand were relatively easy for all participants and were all related to squat movements. Other questions groups discovered with NMF were used to further interpret the discovered clusters as they can capture skills present in certain subgroups of participants. For example, participants in C1 and C2 performed better in questions requiring biomechanical understanding, including how exercises can be effectively modified by external forces (1, 5, 8, 15, 16, 17, 29, and 30). Question 22, on the other

hand included a very different concept from all the others and required a geometrical understanding of segment movements in a squat (Figure 4). The clustering analysis revealed that the lower performing subgroups performed better in this question, perhaps due to more practical experience or by wrong reasoning as it was a true/false question.

In general, the scores of participants were poor, even in the higher performing clusters, which shows the limited biomechanical understanding among professionals and students in the field. There are presently no tools to measure skills of strength training experts based on a biomechanical rationale. Our analysis approach was able to discover multiple surprising and interesting patterns in the test data. Although the questions were designed with the purpose of including varying difficulty levels to measure several skill domains, the performance of participants did not completely follow the expected patterns. For example, some of the presumably easier questions were difficult even for the generally higher performing participants (e.g., question 1 for C1 or question 2 for C2). It was also surprising how some questions were similarly easy or hard for all participants while others had huge differences between the performances of clusters. Thus, cluster analysis can possibly lead us onto the questions where thought processes differ and what type of knowledge that might be missing among strength training experts. In addition, applied skills (working more frequently with evaluating exercise technique) seemed to be very beneficial in handling some questions, as opposed to a high-level understanding of biomechanics (e.g., questions 18-22 for clusters C3 and C4). Furthermore, while some questions can be considered very similar, i.e., measuring similar biomechanics skills, large differences in the performance were discovered in data (e.g., questions 8 and 24 which required identical analyses, but were still answered above and below the random level in all clusters, respectively, but especially clear difference in C1). It is also surprising, how little participant's background (e.g., experience, education) affected the performance, raising even more questions about the general skill level, quality of education, persistent training misconceptions etc. among strength training experts.

When interpreting the results of this kind of test data, it should be taken into consideration that people may arrive at the correct answer using a variety of different thought processes, which for our data can be divided into a) reasoning of biomechanical analysis of forces and moment arms (true experts, can use their expertise to answer all questions), b) reasoning of a subset of biomechanical skills (e.g., understanding elastic bands, anatomy knowledge, or COM and base of support), c) remembering (been exposed to a particular topic before in scientific studies/text books/expert statements etc.), d) experience/"feeling" (have worked more and longer with real training situations and through exposure understand movements better), or e) pure luck.

Compared to more simple and traditional statistical analysis, our data mining approach allows us to discover novel and unexpected knowledge from data. In our case, the acquired knowledge can be used to, for example, recognize what type of knowledge might be missing among strength training experts and form guidelines to improve their education accordingly. More specifically, a cluster analysis approach can be useful to recognize subgroups of learners and make guidelines more personalized or background specific. In addition, the results can be used to optimize the developed test by recognizing what type of questions to focus on, need for more specific questions in certain skill

domains or recognizing possible overlapping questions to exclude. For example, the analysis revealed that questions considered very similar (or requiring similar skills) can be answered very differently across the population and are thus important to include despite similarity. The approach can be used in other educational domains as well to extract and summarize knowledge about learners' skill levels, to better understand learners' domain representations, and recognize optimal questions/need for more specific questions according to test purpose. Use of methodological triangulation is recommended as it increases the trust and confidence for results [27] and helps with further interpretation of results.

6 Acknowledgements

Susanne Jauhiainen was funded by the Jenny and Antti Wihuri Foundation (grant 00190110) and by the Emil Aaltonen Foundation (grant 180063 KO).

7 References

[1] L. Cen, D. Ruta, and J. Ng, "Big education: Opportunities for big data analytics," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 502–506. https://doi.org/10.1109/icdsp.2015.7251923

[2] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. MIT press, 2001.

[3] Baker, "Data mining for education," *Int. Encycl. Educ.*, vol. 7, no. 3, pp. 112–118, 2010.

[4] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," *Int. J. Comput. Sci. Manag. Res.*, vol. 1, no. 4, pp. 686–690, 2012.

[5] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory," *Comput. Human Behav.*, vol. 47, pp. 168–181, 2015. https://doi.org/10.1016/j.chb.2014.09.034

[6] I. Arroyo and B. P. Woolf, "Inferring learning and attitudes from a Bayesian Network of log file data.," in *Artificial Intelligence in Education (AIED)*, 2005, pp. 33–40.

[7] K. Kularbphettong and C. Tongsiri, "Mining educational data to analyze the student motivation behavior," *World Acad. Sci. Eng. Technol*, vol. 6, no. 8, pp. 1036–1040, 2012.

[8] P. D. Antonenko, S. Toy, and D. S. Niederhauser, "Using cluster analysis for data mining in educational technology research," *Educ. Technol. Res. Dev.*, vol. 60, no. 3, pp. 383–398, 2012. https://doi.org/10.1007/s11423-012-9235-8

[9] F. Ghorbani and G. A. Montazer, "Learners grouping improvement in e-learning environment using fuzzy inspired PSO method," in *6th National and 3rd International Conference of E-Learning and E-Teaching*, 2012, pp. 65–70. https://doi.org/10.1109/icelet.2012.6333367

[10] N. A. Rashid, M. N. Taib, S. Lias, N. Sulaiman, Z. H. Murat, and R. S. S. A. Kadir, "Learners' Learning Style Classification related to IQ and Stress based on EEG," *Procedia-Social Behav. Sci.*, vol. 29, pp. 1061–1070, 2011. https://doi.org/10.1016/j.sbspro.2011.11.339

[11] S. B. Aher and L. Lobo, "Applicability of data mining algorithms for recommendation system in e-learning," in *Proceedings of the International Conference on Advances in*

Computing, Communications and Informatics, 2012, pp. 1034–1040. https://doi.org/10.1145/2345396.2345562
[12] N. Bendakir and E. Aïmeur, "Using association rules for course recommendation," in *Proceedings of the AAAI Workshop on Educational Data Mining*, 2006, vol. 3, pp. 1–10.
[13] J. P. Allen and R. Van Der Velden, *The role of self-assessment in measuring skills*. Researchcentrum Onderwijs & Arbeidsmarkt, 2005.
[14] S. L. Deno, "Curriculum-based measurement: The emerging alternative," *Except. Child.*, vol. 52, no. 3, pp. 219–232, 1985.
[15] M. O. Al-Rukban, "Guidelines for the construction of multiple choice questions tests," *J. Family Community Med.*, vol. 13, no. 3, p. 125, 2006.
[16] L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," *Phys. Rev. Spec. Top. Educ. Res.*, vol. 5, no. 2, p. 20103, 2009.
[17] M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang, "Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review," *Rev. Educ. Res.*, vol. 87, no. 6, pp. 1082–1116, 2017. https://doi.org/10.3102/0034654317726529
[18] E. J. Palmer and P. G. Devitt, "Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper," *BMC Med. Educ.*, vol. 7, no. 1, p. 49, 2007. https://doi.org/10.1186/1472-6920-7-49
[19] D. Collett, *Modelling binary data*. Chapman and Hall/CRC, 2002.
[20] R. K. Hambleton and H. Swaminathan, *Item response theory: Principles and applications*. Springer Science & Business Media, 2013.
[21] F. Trudeau and R. J. Shephard, "Physical education, school physical activity, school sports and academic performance," *Int. J. Behav. Nutr. Phys. Act.*, vol. 5, no. 1, p. 10, 2008. https://doi.org/10.1186/1479-5868-5-10
[22] W. D. Gilbert and P. Trudel, "Learning to coach through experience: Reflection in model youth sport coaches," *J. Teach. Phys. Educ.*, vol. 21, no. 1, pp. 16–34, 2001. https://doi.org/10.1123/jtpe.21.1.16
[23] S. Dorgo, "Unfolding the practical knowledge of an expert strength and conditioning coach," *Int. J. Sports Sci. Coach.*, vol. 4, no. 1, pp. 17–30, 2009. https://doi.org/10.1260/1747-9541.4.1.17
[24] I. T. Jolliffe, *Principal component analysis*, 1st ed. Springer, 1986.
[25] N. Denzin, "Strategies of multiple triangulation," *Res. act Sociol. A Theor. Introd. to Sociol. method*, vol. 297, no. 1970, pp. 297–313, 1970.
[26] M. Saarela and T. Kärkkäinen, "Analysing student performance using sparse data of core bachelor courses," *J. Educ. data Min.*, vol. 7, no. 1, 2015.
[27] M. Lewis-Beck, A. E. Bryman, and T. F. Liao, *The Sage encyclopedia of social science research methods*. Sage Publications, 2003. https://doi.org/10.4135/9781412950589
[28] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999. https://doi.org/10.1145/331499.331504
[29] S. Äyrämö and T. Kärkkäinen, "Introduction to partitioning-based clustering methods with a robust example," *Reports Dep. Math. Inf. Technol. Ser. C, Softw. Eng. Comput. Intell.*, no. 1/2006, 2006.
[30] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014. https://doi.org/10.1017/cbo9780511810114
[31] S. Äyrämö, *Knowledge mining using robust clustering*. University of Jyväskylä, 2006.
[32] L. Kaufman and P. J. Rousseeuw, "Clustering by means of medoids," *Stat. Data Anal. based L1 Norm*, 1987.

[33] F. Alalyan, N. Zamzami, M. Amayri, and N. Bouguila, "An improved k-medoids algorithm based on binary sequences similarity measures," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2019, pp. 1723–1728. https://doi.org/10.1109/codit.2019.8820298
[34] L. Kaufmann and P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," *New York Jonh Wiley*, 1990.
[35] U. Von Luxburg and others, "Clustering stability: an overview," *Found. Trends Mach. Learn.*, vol. 2, no. 3, pp. 235–274, 2010.
[36] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.
[37] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999. https://doi.org/10.1038/44565
[38] Q. Sun, P. Wu, Y. Wu, M. Guo, and J. Lu, "Unsupervised multi-level non-negative matrix factorization model: Binary data case," *J. Inf. Secur.*, vol. 3, no. 4, p. 245, 2012. https://doi.org/10.4236/jis.2012.34031
[39] C. Ding, X. He, H. D. Simon, and R. Jin, "On the equivalence of nonnegative matrix factorization and k-means-spectral clustering," 2008.
[40] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994. https://doi.org/10.1002/env.3170050203
[41] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, 2003, pp. 177–180. https://doi.org/10.1109/aspaa.2003.1285860
[42] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 2005 SIAM international conference on data mining*, 2005, pp. 606–610. https://doi.org/10.1137/1.9781611972757.70
[43] L. Mausehund, A. Werkhausen, J. Bartsch, and T. Krosshaug, "Understanding Bench Press Biomechanics — The Necessity of Measuring Lateral Barbell Forces," *J. Strength Cond. Res.*, vol. Epub ahead, 2021. https://doi.org/10.1519/jsc.0000000000003948

8 Authors

Susanne Jauhiainen is with the Faculty of Information Technology, University of Jyväskylä, Finland (susanne.m.jauhiainen@jyu.fi).

Tron Krosshaug is with the Department of Sports Medicine, The Norwegian School of Sport Science, Norway and the Oslo Sports Trauma Research Center, Oslo, Norway.

Erich Petushek is with the Department of Cognitive and Learning Sciences, Michigan Technological University, US.

Jukka-Pekka Kauppi is with the Faculty of Information Technology, University of Jyväskylä, Finland.

Sami Äyrämö is with the Faculty of Information Technology, University of Jyväskylä, Finland.

Article submitted 2021-05-27. Resubmitted 2021-07-05. Final acceptance 2021-07-05. Final version published as submitted by the authors.