Statistica Applicata - Italian Journal of Applied Statistics Vol. 34 (2)
doi.org/10.26398/IJAS.0034-008

HETEROGENEITY IN CLASS: CLUSTERING STUDENTS' ATTITUDES TOWARDS STATISTICS

Aschenbruck[1], Rabea, Szepannek, Gero
Stralsund - University of Applied Sciences, Stralsund, Germany

Lübke, Karsten
ifes Institute for Empirical Research & Statistics, FOM University of Applied Sciences, Dortmund, Germany

Wilhelm, Adalbert F.X.
Jacobs University Bremen, Bremen, Germany

Abstract *Following the growing need for statistical knowledge and the ability to handle data, an increasing number of degree programs include statistics courses and lectures in their curricula. Especially in non-STEM programmes, but not exclusively, it is possible to find heterogeneous groups of students in terms of attitudes towards these topics, and therefore, deeper insights into these attitudes can help to improve curricula and teaching to better fit with the respective student cohort. In a case study, we analyse the results of a survey in which students were asked about their attitude towards statistics, based on the well-known and widely-used SATS-36 questionnaire. Our aim is to identify different attitudinal profiles and to make the heterogeneity of student groups visible. By conducting several cluster analyses, we are able to find separable student groups that can be differentiated, e.g., by their interest towards statistics, by their self-confidence towards their own abilities, or by their willingness to invest more or less effort into learning for the respective class. It turns out that using the individual items instead of the proposed SATS-36 attitudinal constructs lead to a better separation of the student clusters, as does taking gender into account in a cluster analysis of mixed-type data.*

Keywords: *Attitudinal Profiles, Statistics Anxiety, Attitude Constructs, Cluster Analysis, Data Literacy, SATS-36 Survey.*

1. INTRODUCTION

Statistical and Data Literacy have become more and more important topics in a data-oriented and digital world. Many university curricula are responding to this growing demand for data literacy by including statistics and data science courses into the mandatory components of their degree programs, and not only for the

[1]Corresponding author: Rabea Aschenbruck, rabea.aschenbruck@hochschule-stralsund.de

pure STEM programs. Despite the importance and recognition of statistics as an essential component of data science and thus data use in all disciplines, statistics remains a challenging field that is often associated with a large portion of anxiety, especially among students who have bad memories of their mathematics education in high school. *Statistics Anxiety* (Onwuegbuzie and Wilson, 2003) has been identified as a major reason for poor performance in these courses and in the last decades quite some research has been conducted to alleviate this problem, e.g., the ERASMUS+ Project *Adaptive Learning in Statistics (ALEAS)*[2] in which a mobile application was implemented to provide individualised support for students' learning needs (Davino et al., 2020).

In this paper, we identify different attitudinal profiles of students towards statistical content and thus make visible the heterogeneity of students' attitudinal components towards statistics in classroom (originated from the traditional data analysis competition from the annual meeting of the working group *AG DANK* of the *German Society for Classification (GfKl)* held in 2021). Knowing and understanding these profiles is an important and necessary basis for responding appropriately to students' individual needs.

Therefore, in Section 2 we first present the data underlying our study in a very brief exploratory analysis. In Section 3, we cluster the data by using the attitudinal constructs of a widely used statistical survey to find differences between the various domains of students' attitudes. Going further, in Section 4 we cluster the survey items themselves rather than the constructs suggested by the survey design, as there may be a loss of information in determining the constructs of the attitude domains. In order not to neglect the effect of redundant variables of the latent components, an additional cluster analysis with variable selection is performed. Finally, in Section 5, the information on the gender of the students is included with the aim of increasing the quality of the clusters. A short summary of the results with a conclusion and an outlook completes the paper.

2. DATA AND EXPLORATORY ANALYSIS

For a long time, statistics educators have been interested in their students' attitudes toward statistics (Nolan et al., 2012). To evaluate these attitudes and understand their influence on teaching and learning, the widely used *Survey of Attitudes Toward Statistics (SATS)* was developed (Schau et al., 1995). To assess students' attitudes, the SATS-36 instrument consists of 36 items measuring the six constructs briefly described below (Schau, 2019):

[2] `https://aleas-project.eu/`

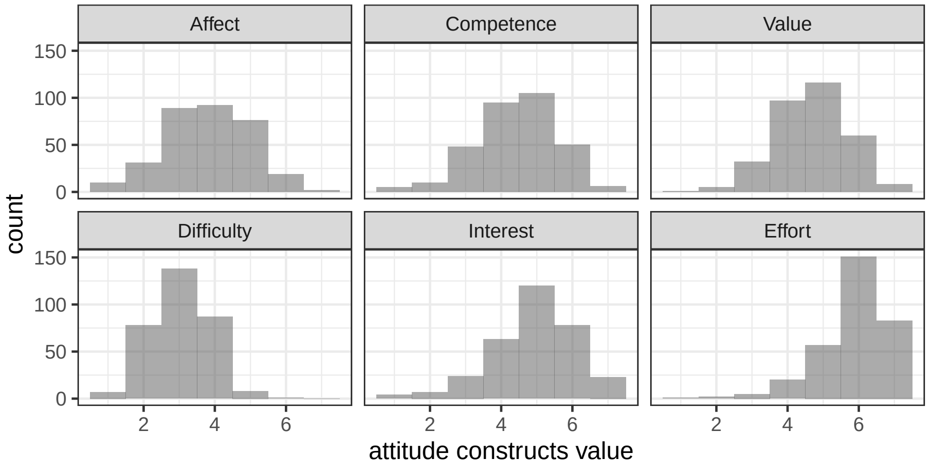

Figure 1: Overview on the attitudinal components as latent variables received by the SATS-36 questionnaire, where higher values represent a more positive attitude towards statistics.

- **Affect**, represents students' feelings concerning statistics
- **Cognitive Competence**, describes students' attitudes about their own intellectual knowledge and skills when applied to statistics
- **Value**, represents students' attitudes about the usefulness, relevance, and worth of statistics in personal and professional life
- **Difficulty**, depicts students' attitudes about the difficulty of statistics as a subject
- **Interest**, shows students' level of individual interest in statistics
- **Effort**, means the amount of work the students expend to learn statistics

The answers are given on a 7-point response scale („do not agree" to „totally agree"). The scores for the six attitude constructs are calculated as the mean over the different attitude domains, so that the higher the scores at the individual constructs, the more positive the attitude towards statistics. A discussion of validity and reliability of this measurement can be found in Nolan et al. (2012). Therefore, as other research on SATS-36, we treat these Likert scale data here as real-valued data (compare e.g., Joshi et al., 2015; Michalopoulou, 2017).

Together with other educators, the present SATS-36 data and the gender of the students were collected by the third author on the basis of an anonymous, voluntary online survey before the statistic-centered course began. The FOM university of applied sciences offers professionals the opportunity to obtain academic qualifications concurrent with their work. More information was not available for the AG DANK data analysis competition. The collected data consist of 153 male and 166 female respondents, that is the students that participated in the survey, thus we have a roughly balanced gender ratio among the total of 319 participants with a major in different business-related subjects (Bachelor or Master degree).

As a preliminary conclusion, when looking at the distributions of the constructs of the students' attitudes towards statistics in Figure 1, it can be generally stated that high values were observed for *interest* and *value*. The latent variable *competence* also tends to have higher values, whereas *affect* is neutral for all students surveyed. The right-skewed distribution for the construct *difficulty* as well as the left-skewed distribution for the construct *effort* are remarkable results, which could be a result of social desirability, even though the students participated anonymously in the study. Together with the very high values for the component *effort* compared to the other components, the low values for *difficulty* show that the students apparently expect problems in the statistics lecture.

The aim of conducting the survey was to find out students' attitudes towards statistics in order to identify factors that may influence their learning behaviour directly at the beginning of the course. By recognising and being aware of the different attitudes, the methodology of teaching can be adapted. Thus, in the following, these attitude components should form the basis for a cluster analysis in order to identify student groups that are as separable as possible.

3. CLUSTER ANALYSIS OF ATTITUDE CONSTRUCTS

The analysis is conducted using the popular and well-known cluster algorithm k-Means (Jain, 2010; MacQueen, 1967). At this point, it should be briefly mentioned that the gender of the students is not taken into account in this part of the analysis and therefore a gender-neutral consideration is initially made. As always for partitioning clustering methods, the number of clusters has to be determined in advance (Hennig et al., 2015). Taking into account various so-called internal validation indices to determine the index-optimal number of clusters (Halkidi et al., 2015), the consideration of a set of 24 validation indices proposed by Charrad et al. (2014) shows that according to the majority rule the best number of clusters is three (an overview of the distribution of the resulting number of clusters is given in Table 1).

The cluster partition determined by the k-Means algorithm is rated with an internal validation index gamma value of 0.593. The biplot in Figure 2 on the left, based on a principal component analysis of the six constructs, shows that the

Table 1: Overview of the distribution of the resulting index-optimal number of clusters, determined by 24 internal validation indices. According to the majority rule, the best number of clusters is three.

Number of Cluster	2	**3**	4	11	14	15	17	18	20
Number of Index-optimal choices	7	**8**	1	1	1	1	2	1	2

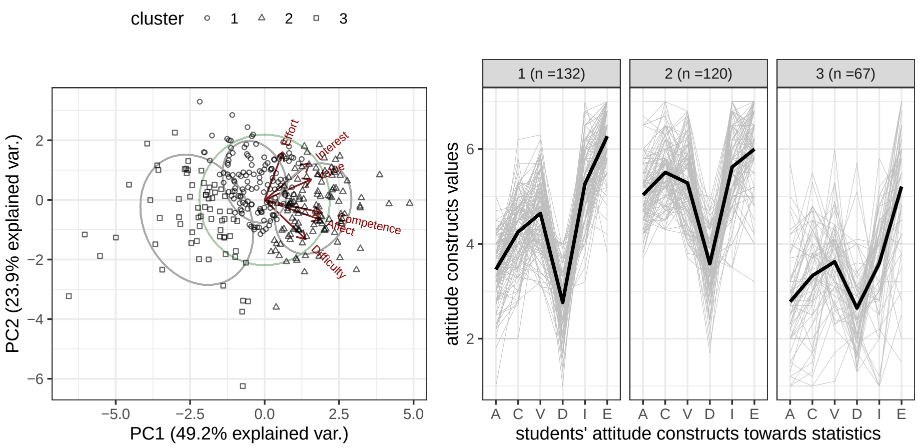

Figure 2: Visualisation of the three clusters of the students' attitude components towards statistics determined by the k-Means algorithm (gamma = 0.593). ***Left:*** **Biplot of the principal components.** ***Right:*** **Students' attitudinal profiles (light grey) and the mean for each cluster (black) for the cluster partition.**

clusters mainly differ in the first principal component. The loadings reveal correlation between the *competence* and *affect* scores as well as between *interest* and *value*. The rather low correlation between *difficulty* and *effort* is also striking. On the right-hand side in Figure 2, the three clusters are visualised by the attitudinal profiles of the students (light grey). For each cluster, a characterising profile was formed by averaging the values per attitude construct (black). First of all, it can be observed that clusters 1 and 3, in contrast to cluster 2 (37.7% of the respondents), have rather negative attitudes towards statistics education (overall lower scores for the six construct values). Cluster 1 (41.4%) differs from cluster 3 (21.0%) in that these students indicate a greater interest and want to invest more working time.

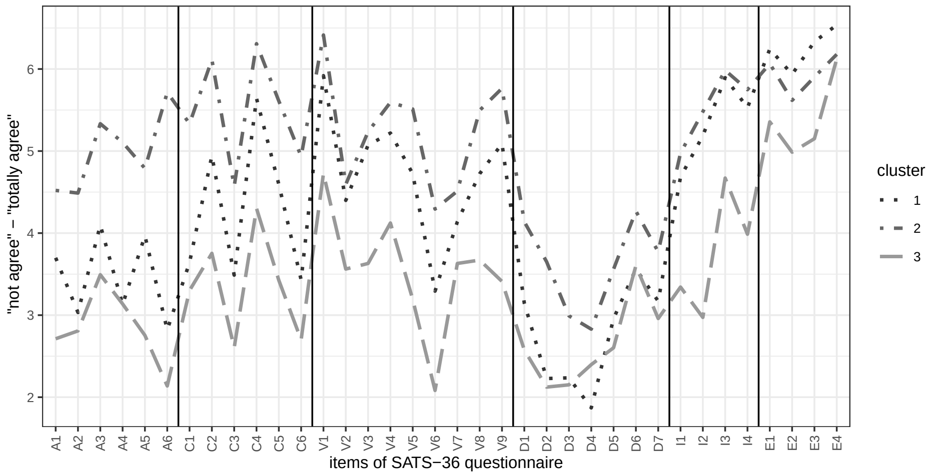

Figure 3: Visualisation of the cluster attitudinal profiles based on the items of the SATS-36 obtained by the application of the k-Means algorithm (gamma = 0.475).

Students in the second cluster have greater self-confidence in their own abilities with regard to statistics.

In summary, for the different attitudes of each cluster group, students from cluster 1 expect difficulties in studying statistics, but are interested and motivated, which is reflected in their high effort. Students in cluster 2 have a significantly more positive attitude towards statistics and are more confident in their own statistics skills. These students also show interest in statistics and a high effort. Cluster 3 contains students who have a negative attitude towards statistics and show neither interest nor motivation in learning statistics.

4. CLUSTER ANALYSIS OF THE 36 ITEMS OF SATS-36

The cluster groups derived so far are based exclusively on the attitudinal constructs aggregated from the SATS-36 questionnaire. The resulting loss of information by considering only the constructs possibly reduces the quality of the clusters. Therefore, in the following, in an alternative approach student groups will be created based on the 36 items themselves. The goal remains to identify characteristics that differentiate as well as possible in order to be able to optimally adapt teaching and learning methods. It should be noted that interpretation of the results can only be based on the premise that the data were collected using a 7-point Likert scale and are considered numerical due to the scale construction.

Figure 3 shows the attitudinal profiles of the three clusters based on the application of k-Means to the 36 items of the questionnaire. For better illustration, vertical lines were used to highlight the items' different attitude domains in the diagram. Overall, it can be seen that the attitudinal profile of cluster 3 (23% of the respondents), viewed across all items, remains in most cases (noticeably) below the other two learning profiles. The course of the profiles in the area of *affect* is very different between the three clusters, whereas the profile course within the items on *competence* is nearly the same for all three clusters. For the item V7 „Statistics conclusions are not rarely presented in everyday life", cluster 3 shows higher agreement compared to the other two clusters. Clusters 2 (38.5%) and 3 have similar profiles in the areas of *value*, *difficulty*, *interest*, and *effort*, while cluster 1 (38.5%) is only clearly more positive about the effort for statistics, which means these students are willing to increase their workload for learning statistics. Item D4 „Learning statistics does not require a great deal of discipline" is particularly interesting: Here, cluster 3 is more approving compared to the other clusters.

Due to the many different items, it is difficult and uncomfortable to identify the most striking differences. Furthermore, the gamma index score of 0.475 is worse than the cluster partition generated on the basis of the attitude constructs. It is very plausible that some variables of the same attitudinal domain may be redundant as these items form one of the six constructs of *affect*, *cognitive competence*, *value*, *difficulty*, *interest*, and *effort*. Therefore, in the following we perform a variable clustering over latent components where variables are clustered to groups such that the correlation to cluster's first principal component is maximized Vigneau and Qannari (2003). Afterwards, for each cluster one representative variable will be retained for further analysis.

Figure 4 shows the variation of the clustering criterion when passing from a partition with K clusters to a partition with $K-1$ clusters on the left-hand side. Based on the elbow at the transition from three to two clusters, the determined number of variable clusters should be three. For each cluster, the variable with the largest correlation with its first principal component is selected (cf. Vigneau et al., 2015). The resulting data is used to cluster the students with the R function `hclust` (R Core Team, 2022). On the right-hand side of the figure, the results for three clusters are shown by boxplots. This clustering is rated with a gamma index value of 0.635, which denotes an improvement compared to the clustering of the previous section. It can be seen that the observations are mainly split between clusters 1 and 2 and the third cluster contains only 10 observations (3.1%). These few observations are characterised by a „don't care" attitude: The students do not

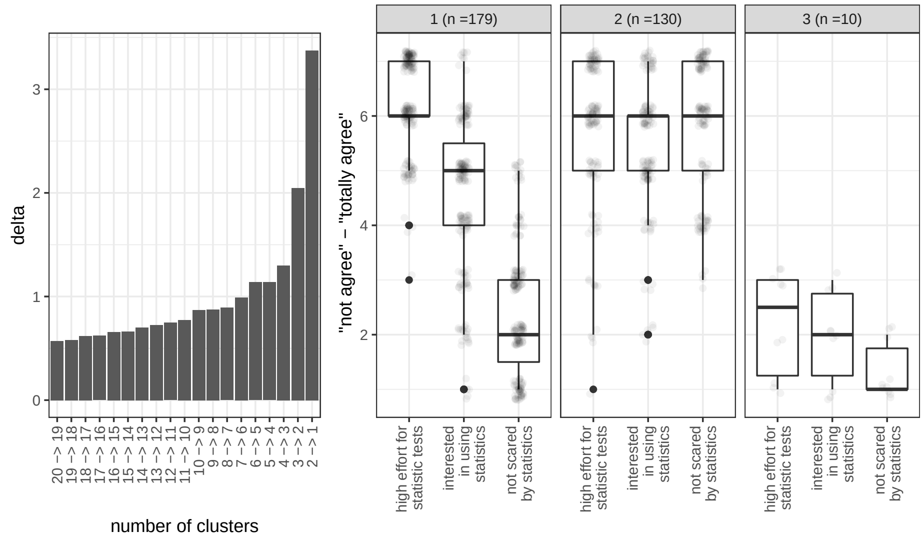

Figure 4: ***Left:*** **Variation of the clustering criterion when passing from a partition into K clusters to a partition with $K-1$ clusters. Based on this, an appropriate number of variable clusters is three.** ***Right:*** **Visualisation of the three clusters of the selected students' attitude items towards statistics determined by hierarchical clustering (gamma = 0.635).**

plan to spend a lot of time studying for the statistical tests and are not particularly interested in using statistics, while they are afraid of the subject of statistics in general. Students in clusters 1 and 2, on the other hand, plan to spend a lot of time learning and are more interested in applying statistics. The difference between these two cluster groups is that students in cluster 2 (40.8%) are not afraid of the subject of statistics in general, in contrast to students in cluster 1 (56.1%).

5. CLUSTER ANALYSIS OF THE GENDER-EXTENDED SATS-36 ITEMS

In some studies on learning behaviour, differences between genders have already been demonstrated, see for instance Pinto et al. (2018). This aspect has deliberately not been taken into account so far in order to be able to formulate gender-neutral statements. Figure 5 shows the distribution of the values of the attitude constructs per gender. It can be seen that there are differences between the genders in the tendencies of the values. Due to this fact, gender is also included in the following analysis for the sake of completeness, so that possible gender-specific

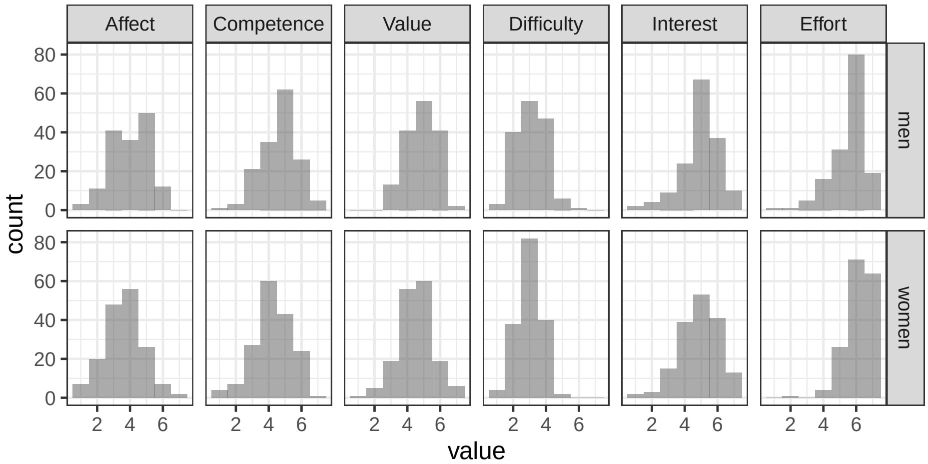

Figure 5: Constructs recieved by the SATS-36 questionnaire with respect to the students' gender, where higher values represent a more positive attitude towards statistics.

aspects can be considered.

To apply a clustering method on the numeric SATS-36 data and the categorical variable *gender*, the k-Prototypes algorithm is used, which is an extension of the k-Means algorithm for mixed-type data (Huang, 1998; Szepannek, 2018). In the light of the experience gained in Section 4, first of all the most informative variable sets should be determined in the following. Therefore, variable selection is derived using a genetic algorithm with the aim of maximising the dimension-independent internal validation index gamma (Aschenbruck and Szepannek, 2022). The auxiliary condition was set such that, in addition to the information on the gender of the surveyed students, at least one question from each of the domains of the six attitudinal constructs must be included in the variable set.

The index-optimal cluster partition resulting from the applying of a genetic algorithm consists of five cluster groups and is evaluated with a gamma index value of 0.627. The variables determined as the basis of this cluster partition are *gender* and one item each from the domains of *affect*, *competence* and *interest*. Furthermore, there are two items from the area of *value* and *difficulty* and three more from the attitude component *effort* (for more details on the different items see Table 2). Figure 6 shows the visualisation of the partition with five clusters.

Table 2: Selected items of the SATS-36 questionnaire and the gender of the students obtained by applying the genetic algorithm.

Acronym	Component	Questionnaire Item
G	–	Students' gender (male, female).
A6	Affect	I am not scared by statistics.
C1	Competence	I will not have trouble understanding statistics because of how I think.
V1	Value	Statistics is not worthless.
V6	Value	I use statistics in my everyday life.
D2	Difficulty	Statistics is not a complicated subject.
D3	Difficulty	Statistics is a subject quickly learned by most people.
I2	Interest	I am interested in using statistics.
E1	Effort	I plan to complete all of my statistics assignments.
E3	Effort	I plan to study hard for every statistics test.
E4	Effort	I plan to attend every statistics class session.

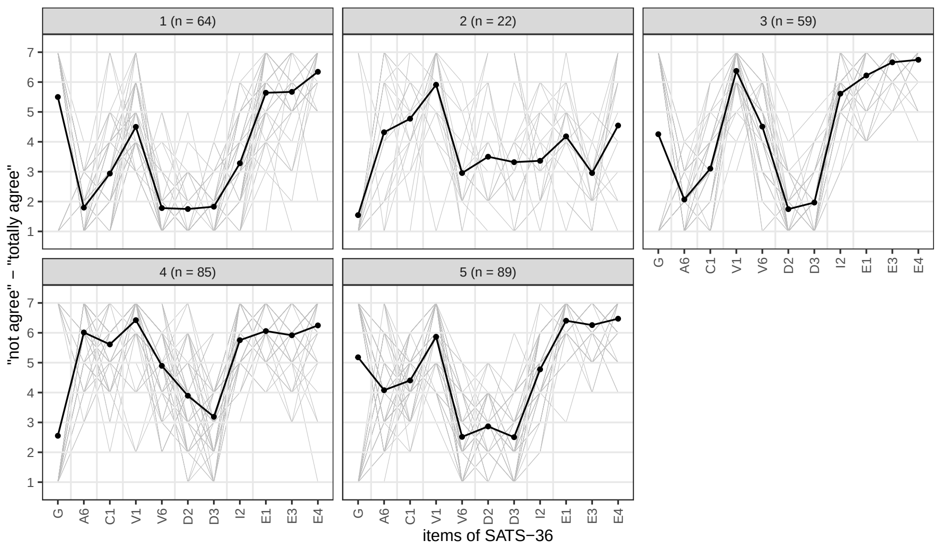

Figure 6: Overview on the resulting cluster partition with five cluster groups of the genetic algorithm application for variable selection. Visualisation of the students' attitudinal profiles (light grey) and the mean for each cluster (black) for the cluster partition (gamma = 0.627). Note: Male is represented by 1 and female by 7 without affecting the cluster analysis.

Again, the collected data of the individual students are shown in light grey and the so-called attitudinal profiles in black. Note that the gender variable in the visualisation has been coded numerically for simplicity, with male represented by the number 1 and female by the number 7.

When looking at the clusters, it is noticeable that clusters 1 (20.1% of the respondents) and 5 (27.9%) consist mainly of female students, whereas clusters 2 (6.9%) and 4 (26.6%) consist mainly of male students. The students assigned to clusters 1 and 2 are scared by the subject statistics. Cluster 2 and 5 tend to expect no problems in understanding statistics and only cluster 4 has a mean value of higher than 5, which can be interpreted as expecting no problems. The opposite holds for clusters 1 and 3 (18.5%): these students expect problems in understanding statistics. All clusters tend to agree with the proposition that statistics is not worthless, with cluster 1 having the lowest scores for that item. The usage of statistics in everyday life is not that popular, especially among clusters 1, 2 and 5. The opinion that statistics is an easy subject to learn is rather not shared by all clusters. Clusters 1 and 3 in particular are negative minded towards the difficulty of learning the subject of statistics. Cluster 3, 4 and 5, in contrast to clusters 1 and 2, are interested in the application of statistics. The planned effort is very high in nearly all clusters, with the exception of the strikingly low effort in cluster 2.

In summary, clusters 1 and 3 contain mainly anxious students with low self-confidence who show a medium interest in the subject of statistics but a high effort. Differences can be found in the assessment of the value of statistics: While students in cluster 1 tend to be neutral about the value of statistics and see no benefit of the subject in their everyday life, students in cluster 3 deny the worthlessness of the subject statistics. The characterisation of cluster 5 is similar, but the associated students are less afraid of statistics. Cluster 2 shows a neutral attitude towards statistics, but with a low effort to study the subject, which is the smallest cluster group compared to all others. Finally, students in cluster 4 are not scared by statistics, do not see any difficulties and put a lot of effort into studying statistics.

6. RESULTS AND CONCLUSION

First, differences in the attitudinal components of students were found, which allowed the formation of 3 groups: On the one hand, students who are not motivated to learn statistics content and, on the other hand, interested and motivated students who differ in whether they expect problems in learning or are confident about their own statistical abilities (see Figure 2). An improvement of almost 10% in cluster quality assessed with the Gamma Index, could be achieved by examining the questionnaire items instead of the attitude constructs and selecting variables on them. The three resulting clusters show similar group characteristics: on the one hand, a small cluster with anxious students who plan little effort and have no interest in applying statistics. On the other hand, students can be identified for the most part who plan a high effort for the statistics lecture and are more interested in the application of statistics. Again, these two clusters differ in terms of anxiety about statistics (see Fig. 4). Although significant gender differences are generally identifiable, particularly in learning behavior, the addition of student gender did not increase the quality of the clustering, see Section 5.

In summary, it seems to be important for individuals who are teaching statistics to find out how interested students are in using statistics, how much work they are willing to put in, and whether they are afraid of the statistical content. Individual curricula and teaching can then be adapted accordingly.

In addition, as noted by Schield (2018): Students often see less value in statistics after taking the course than before, which cannot be analysed using the data at hand. Chance et al. (2018) also provide data for this and show that Simulation Based Inference could have a positive impact on students' attitudes. Furthermore, including causality might also be beneficial (Lübke et al., 2020). However, since we can see that there is heterogeniety in class, simply looking at averages may not be sufficient to analyse students' attitudes towards statistics. Statistics educators should be aware of this and look for ways to stimulate learning in all different groups.

ACKNOWLEDGEMENT

This paper is based on the traditional data analysis competition of the meeting of the working group *AG DANK (Arbeitsgruppe Datenanalyse und Numerische Klassifikation)* of the *German Society for Classification (GfKl)*, which was held in Bielefeld in 2021. We appreciate presenting and outlining the different individual approaches to analysis and the lively exchange between all participants at this meeting.

References

Aschenbruck, R. and Szepannek, G. (2022). Genetic algorithm-based variable selection for optimization of cluster partitions for mixed-type data and its application on a digital teaching survey. In *Archives of Data Science, Series A*. Special Issue - Post-Conference Proceedings ECDA 2021; submitted 2021.

Chance, B., Mendoza, S., and Tintle, N.L. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. In M.A. Sorto, A. White, and L. Guyot, eds., *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics*.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. In *Journal of Statistical Software*, 61 (6). doi:10.18637/jss.v061.i06.

Davino, C., Fabbricatore, R., Pacella, D., Vistocco, D., and Palumbo, F. (2020). Aleas: a tutoring system for teaching and assessing statistical knowledge. In *PSYCHOBIT*.

Halkidi, M., Vazirgiannis, M., and Hennig, C. (2015). Method-independent indices for cluster validation and estimating the number of clusters. In *Handbook of Clusteranalysis*, 595–618. Chapman and Hall/CRC.

Hennig, C., Meila, M., Murtagh, F., and Rocci, R., eds. (2015). *Handbook of Cluster Analysis*, chap. Cluster Analysis: An Overview. CRC Press. doi:10.1201/b19706.

Huang, Z. (1998). Extension to the k-Means algorithmus for Clustering Large Data sets with Categorical Values. In *Data Mining and Knowledge Discovery*, 2 (6): 283–304. doi:10.1023/A:1009769707641.

Jain, A.K. (2010). Data clustering: 50 years beyond k-means. In *Pattern Recognition Letters*, 31 (8): 651–666. doi:10.1016/j.patrec.2009.09.011. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

Joshi, A., Kale, S., Chandel, S., and Pal, D.K. (2015). Likert scale: Explored and explained. In *British Journal of Applied Science & Technology*, 7 (4): 396.

Lübke, K., Gehrke, M., Horst, J., and Szepannek, G. (2020). Why we should teach causal inference: Examples in linear regression with simulated data. In *Journal of Statistics Education*, 28 (2): 133–139.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Michalopoulou, C. (2017). Likert scales require validation before application-another cautionary tale. In *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 134 (1): 5–23.

Nolan, M.M., Beran, T., and Hecker, K.G. (2012). Surveys assesing students' attitudes toward statistics: A systematic review of validity and reliability. In *Statistics Education Research Journal*, 11 (2): 103–123.

Onwuegbuzie, A.J. and Wilson, V.A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments–a comprehensive review of the literature. In *Teaching in Higher Education*, 8 (2): 195–209. doi:10.1080/1356251032000052447.

Pinto, G., Bigozzi, L., Vettori, G., and Vezzani, C. (2018). The relationship between conceptions of learning and academic outcomes in middle school students according to gender differences. In *Learning, Culture and Social Interaction*, 16: 45–54. doi:10.1016/j.lcsi.2017.11.001.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Schau, C. (2019). Survey of attitudes toward statistics (sats-36). `https://www.evaluationandstatistics.com/`.

Schau, C., Stevens, J., Dauphinee, T.L., and Del Vecchio, A. (1995). The development and validation of the survey of attitudes toward statistics. In *Educational and Psychological Measurement*, 55: 868–875.

Schield, M. (2018). Confounding and cornfield: Back to the future. In M.A. Sorto, A. White, and L. Guyot, eds., *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics*.

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. In *The R Journal*, 10 (2): 200–208. doi:10.32614/RJ-2018-048.

Vigneau, E., Chen, M., and Qannari, E.M. (2015). Clustvarlv: An r package for the clustering of variables around latent variables. In *The R Journal*, 7 (2): 134–148.

Vigneau, E. and Qannari, E.M. (2003). Clustering of variables around latent components. In *Communications in Statistics - Simulation and Computation*, 32 (4): 1131–1150.