

Methodology

- Data from the *Descriptive Statistics* course at the University Duisburg-Essen, Germany
- Exams consist of arithmetical problems, programming tasks in [R](#), and a short essay task
- Both exams are conducted digitally with the e-assessment system [JACK](#)

	Comparison	Test
Year	18/19	20/21
N	109	151
Style	proctored	unproctored
Total points	60	60
Sub tasks	19	17
Duration	60	60

, with the intent to compare the outcomes of an unproctored exam ($N = 151$) during the COVID-19 pandemic with a proctored exam ($N = 109$) held in-person before the pandemic. * The two groups' exams involved solving arithmetical problems, completing R programming tasks, and writing a short essay. During the exam process, student activities and time stamps were tracked in event logs, and the points achieved for each task were recorded. * The dataset was cleaned to ensure comparability by removing students with minimal participation or achievement, as well as those who encountered internet issues. * The study utilized an agglomerative (bottom-up) hierarchical clustering algorithm that can be described by following equation:

$$D(x_i, x_{i'}) = \frac{1}{h} \sum_{j=1}^h w_j \cdot d_j(x_{ij}, x_{i'j})$$

- $D(x_i, x_{i'})$ is the global pairwise dissimilarity, while $d_j(x_{ij}, x_{i'j})$ denotes the pairwise attribute dissimilarity. The weights w_j sum up to 1. Index i denotes the number of Students $i = 1, \dots, N$ with $N = 151$ students, while j is the index for each of the h attributes.
- We compared two different kinds of attributes, namely the dissimilarities in the student's event patters (time of submission) defined as $d_j^T(v_{ij}, v_{i'j})$ and the dissimilarities in points achieved $d_j^P(s_{ij}, s_{i'j})$.

$$D(s_i, s_{i'}, v_i, v_{i'}) = \frac{1}{h} \sum_{j=1}^h (w_j^P \cdot d_j^P(s_{ij}, s_{i'j}) + w_j^T \cdot d_j^T(v_{ij}, v_{i'j}))$$

is the combined model, where all weights w_j^P and w_j^T add up to one.

Empirical results

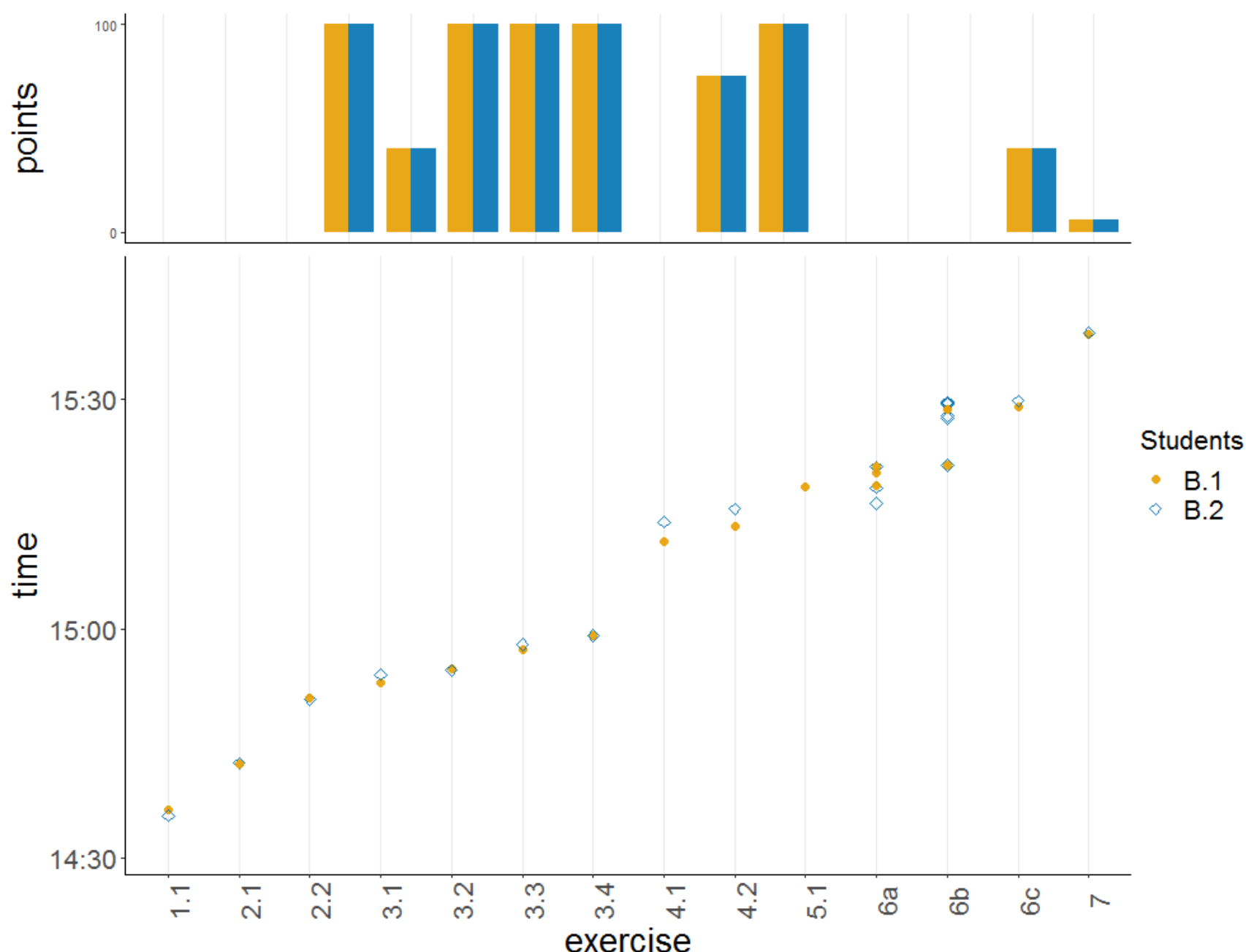


Figure 1: Event logs and achieved points of the cluster **B** from the test group (2020/21). Above the scatter plot, a bar chart is added to compare the points per subtask.

- Figure 1 illustrates the individual comparison of achieved points and event logs of the student cluster with the highest similarity.
- Figure 2 represents the dendrogram of the test group where students took the exam from home.
- The dendrogram of the test group displays a lower overall level of dissimilarity compared to the comparison group, suggesting possible collusion.
- Six clusters within the test group, labeled **A-F**, show significantly lower dissimilarity, standing out noticeably from the rest of the cohort.

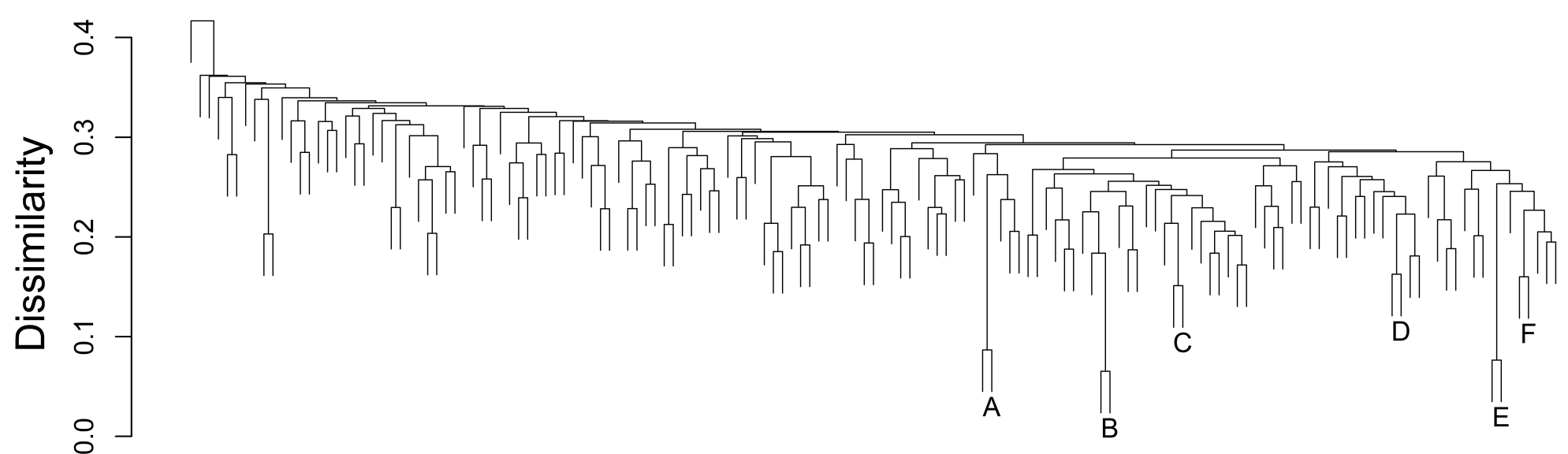


Figure 2: Dendrogram produced by average linkage clustering of the unproctored test group (2020/21). **A-F** mark the clusters with the lowest dissimilarity.

- Figure 3 compares the normalized distributions of the dissimilarity measures between the comparison and test groups.
- Three data points from the test group are markedly distinct from the rest of the data points.

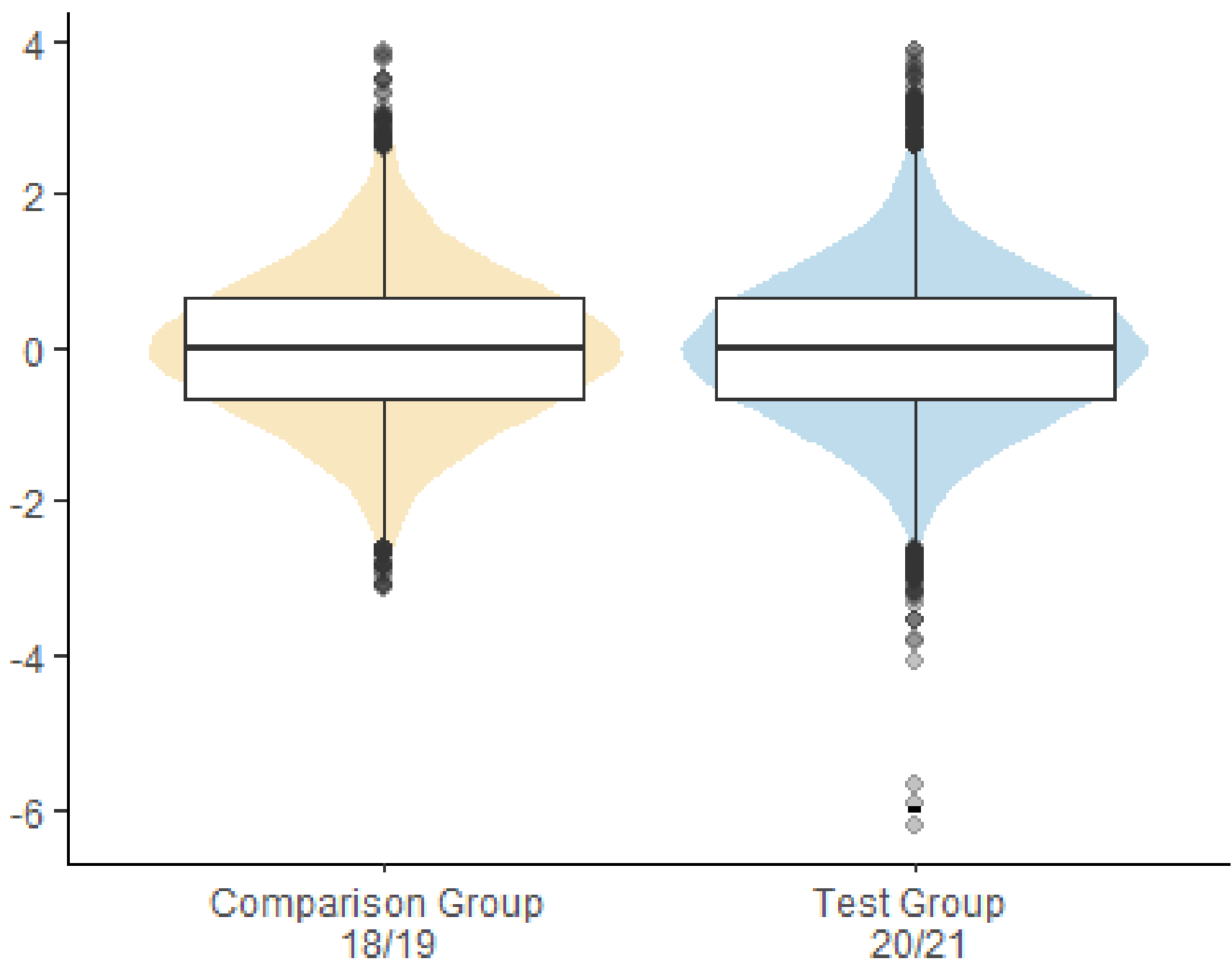


Figure 3: Comparison of the normalised distance measures.

Discussion

- The study discusses the results of hierarchical clustering algorithms, visually represented via a dendrogram, a tree-like structure.
- Various clustering algorithms were compared, and average linkage clustering was found to be the most suitable for the analysis.
- The use of average linkage clustering helped identify compact clusters (specifically clusters **A**, **B**, and **E**), suggesting a lack of large group collusion.
- Additional visual tools like scatterplots and bar charts were employed to examine similarities among students within these clusters.
- The study used a reference group for comparison, validating the method's effectiveness in detecting collusion, though limitations exist due to unknown ground truth.
- The approach not only aids in deterring cheating in unproctored exams but also contributes to the broader digital transformation of education, equipping us to handle unforeseen future challenges similar to the COVID-19 pandemic.

Further research

- Future research could assess the long-term efficacy of the collusion detection method during exams and its impact on academic integrity and student behavior.
- Additional studies might focus on refining methods for gathering and analyzing supplementary evidence, with the ultimate goal of improving collusion detection rates. These efforts aim to provide a better understanding of the prevalence and extent of student collusion.