Report on the results of DigiEfekt study: Opiq interaction

Yaroslav Opanasenko, Margus Pedaste, Leo Aleksander Siiman

Introduction

E-Books have quickly turned from a trendy and curious novelty into an effective educational tool that is widely used in Estonian basic schools. All of Estonia's 500+ schools had used Opiq until the study year 2021/2022– «an interactive digital learning materials platform that replaces all the old-school workbooks». Given the popularity of the E-Book as a pedagogical tool, several questions immediately arise: are E-Books really that effective in education? How much do modern E-Books differ from their printed analogues, and what is the difference? What behavioral patterns of interaction with E-Books exist?

The topic of the effectiveness of the use of electronic workbooks in education has remained relevant in the scientific literature over the past decade. E-Books can ease obtaining educational materials (Yaya, 2015); also help students understand learning materials «more systematic» (Noor, Embong & Ridhuan, 2012); E-Books can enhance students' access to information, and also help «revolutionize the processes of reading, analyzing information» (Blummer & Kenton, 2020). A study by MacNish et al. (2017) confirms the effectiveness of the use of E-Books in conditions of problem-based learning; significant increase in student curiosity in Science courses when using interactive multimedia E-Books (Herianto et al., 2022).

However, as Ogata (2015) mentioned, most studies dedicated to E-book-based learning pay little attention to analysing the e-book data logs, although it is imperative to investigate how these logs can be used to improve E-Books contents and the quality of learning and education. The use of classical methods of data analysis in this field may not be sufficient (Cerezo et al., 2020). Therefore, sequence analysis and process mining were used to study clusters of students based on their behavioral patterns of interaction with Opiq and identify effective strategies for interacting with the E-Book.

Methodology

Sequence analysis is a categorical longitudinal modelling procedure with a holistic approach that allows the inclusion of the whole trajectories of a set of objects, stating all states of interest experienced within a specific timeframe, to create categorical sequences and usually produce a product of a typology or clustering classification (Tan & Savedham, 2022). In the last two decades, sequence analysis has seen many successful applications in the social sciences (Liao et al., 2022). Despite the growing popularity of this method in the social sciences, sequence analysis is just beginning to be actively used in educational research. So, Tan & Savedham (2022) conducted a study on the analysis of student habits when interacting with MOOCs. Brzinsky-Fay (2022) analyse individual labour market entry processes after vocational education and training (VET) in Germany.

Tan (2022) describes a possible application algorithm and the main goals of sequence analysis in educational research:

1. Identify what are the main keys/categories to use to form the sequence.
2. Create sequences of students' action during the learning process.
3. Examine similarity of the students' learning activity sequences
4. Classify students into clusters using a classification tool.

5. Relate the classification results to learning outcomes or other variables.

Process mining is an area of research which uses log-data obtained by software systems that support a big variety of different processes to build models based on specific patterns (van der Aalst & Weuters, 2004). Emond and Buffett (2015) applied process mining and sequence classification mining techniques to model and support SRL in heterogeneous environments of learning content, activities, and social networks. Like sequence analyses, process mining can be used in conjunction with cluster analysis: e.g., Bogarin et al. (2014) used cluster analyses to able to generalize Process Models of students' behaviour and learning patterns.

Descriptive results

The log data was gathered during the 2021/2022 study year from DLE Opiq. Opiq usage log data provides information about every action that was made by the student during interaction with E-Books including school, class and grade of the student; student and session IDs; subject, book title, chapter title, topic; the number of exercise; media-file ID; type, date and time-stamp of interaction; type of the exercise, information about entering, checking, fixing and saving result of tests.

The first stage of the study was to establish the average amount of time spent in Opiq for the 3rd, 6th and 9th grades and exclude from further research data of those students who spent in Opiq less than 1 hour. The filtered log data of 133 students from the 3rd grade, 204 students from the 6th grade and 158 students from the 9th grade (495 in total) was gathered.

It was estimated that average time in Opiq for the 3rd grade students is 226 minutes and for the 9th grade - 427 minutes.

To generalize the data for students of all grades, it was decided to analyse the average test result, after which the students stopped correcting/improving it. Since it's impossible to compare different tests in different subjects, our «index» could show not the specific subject competence, but student engagement and motivation.

It was found that the average final result of optional tests for 3rd grade was 94.73%, for 6th grade 89.01%, and for 9th grade 79.34%. Decreasing trend could be a signal of learning motivation decreasing from year to year.

Process sequence analysis were applied to describe different levels of the interaction process with Opiq. Dotted algorithm chart allows to visualize statistics of student interactions with Opiq throughout educational year. Dotting chart can display not only quantity of sessions (each figure on the graph) throughout the academic year, but also to characterize them by duration (colour of the figure) and the day of the week (shape of the figure). This technology allows not only to determine the most active students (each line on y-axis is one student interaction with Opiq during the year), but also track down periods, session duration statistics, and, for example, the number of interaction sessions over the weekend (Fig. 1.).

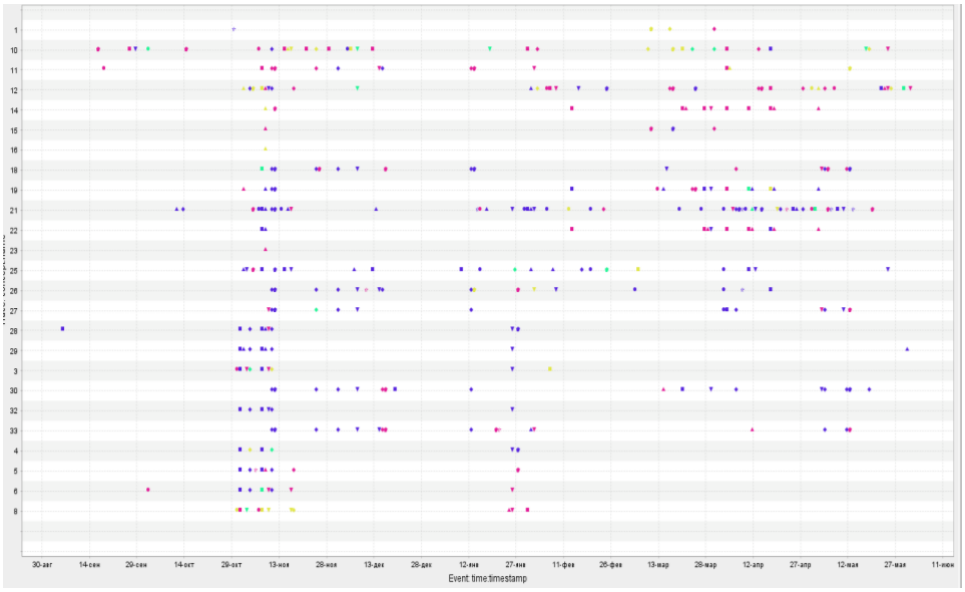

Fig. 1. Dotted chart graph example (Process Mining) .
X- axis–date; Y-axis–student ID

Sequence analysis has great analytical potential in the field of studying the process of passing tests. So, sequence analyses could help the features of both the order of passing tests (Fig. 2), and the effectiveness of their corrections and the final result (Fig. 3). Test sequences are analysed using cluster analysis, which makes it possible to cluster students based on their behavioural patterns.

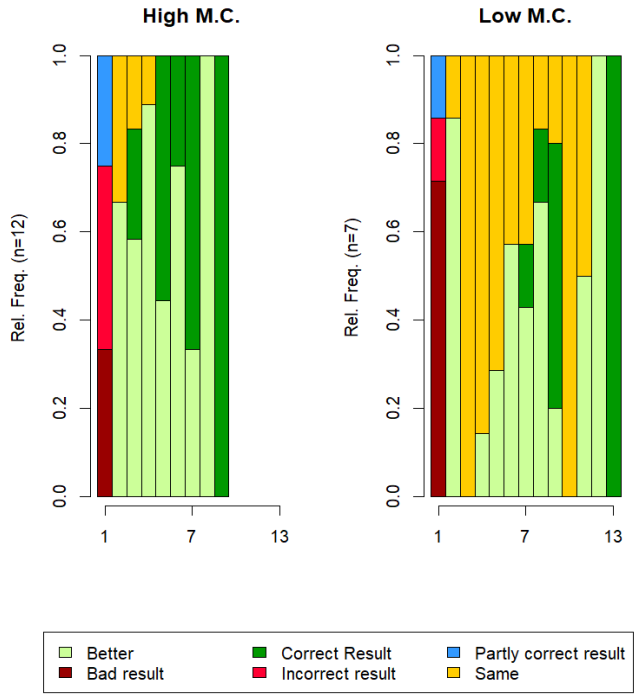

Fig. 2. Sequences of fixing tests (Sequence analysis)
X- axis–number of actions; Y-axis = frequency

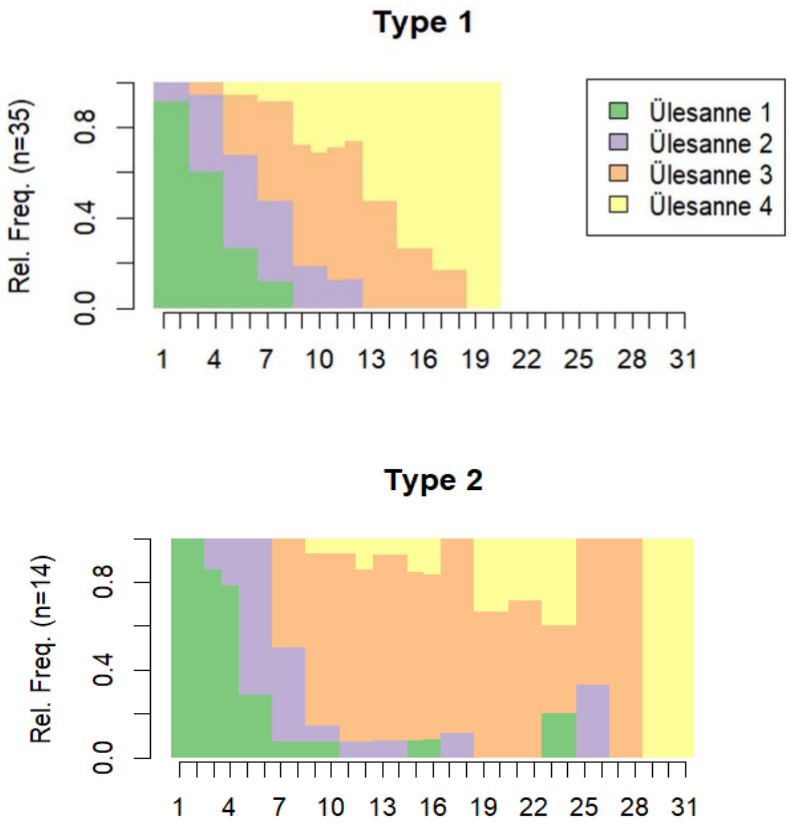

Fig. 3. Sequences of taking exercises in different order (Sequence analysis) .
X- axis–number of actions; Y-axis = frequency

Results of the main analysis

In order to explore the features of students' interaction with Opiq, the mining package ProMLite 3.3 was used for running process mining as it is one of the most popular software for running this technology and has been successfully used in educational studies. Heuristic miner was used due to the data requirements. IN result was received model interactions students With Opiq. The lines in Fig. 4 show transitions between different actions; bold lines show actions that were performed repeatedly; numbers inside the blue boxes show the number of times the actions were performed.

Given the heterogeneity of the current dataset, the creation of generalized models of student interaction with Opiq using process mining is currently not possible. Each chapter has a unique structure that includes different types of media content, text blocks, practice assignments, etc. In addition, now we have no information about how teachers included Opiq in educational process – in the case of some classes, the study of the chapter could be mandatory, other students could independently study the material. One of the requirements of process mining is the homogeneity of the data, which in this case will include the same conditions for students (for example, the whole group must study the chapter in class or at home).

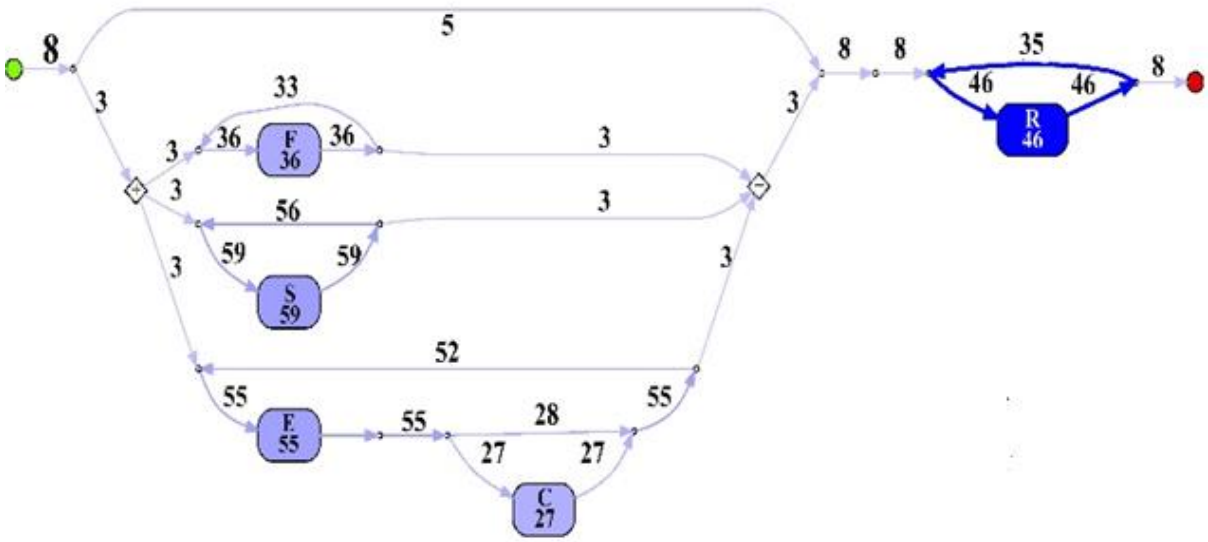

Fig. 4. Heuristic mining graph example (Process Mining).
F – fixing, S – saving, E – entering, C – checking, G – gallery opening and closing, R – reading

However, studying the results obtained using process mining prompted us to generalize the main types of interaction with Opiq based on the available log-data. Thus, 4 main groups of activities were identified, each of which included several separate types of interactions:

- Reading – including reading chapters and «text-to-audio» function.
- Media-content interaction – including playing video and audio tracks and watching gallery.
- Practice – entering practice tests, checking, fixing and saving the result.
- Formal tests – tests outside the chapter, which are given directly by the teacher.

The percentage of the frequency of each category in the student's overall interaction with Opiq was measured. After that, cluster analyses, using the cluster package version 2.1.4 for R and Agglomerative Nesting (Hierarchical Clustering), was applied. Results are shown in the Fig. 5.

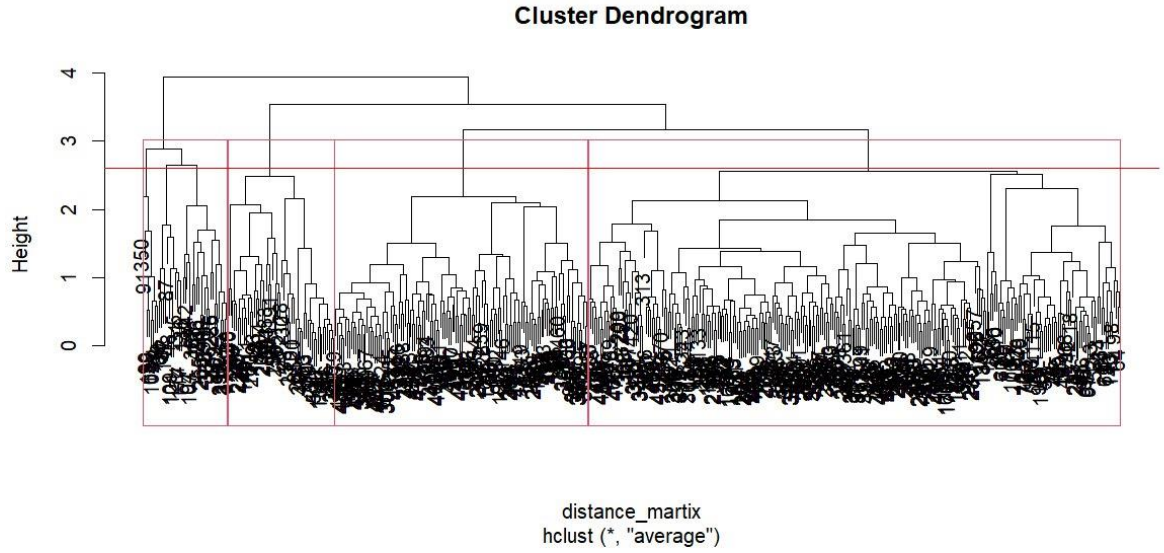

Fig. 5. Cluster Dendrogram of students preferred different activities in Opiq

Four meaningful distinguished clusters, that describe students' behaviour in Opiq were allocated. Each cluster corresponds to the predominance of one preferred activity (reading, media-content interaction, taking practice tests and taking formal tests over the others) reading, media-content interaction, taking practice tests and taking formal tests. Clusters were named based on the preferred activities:

- Readers – the ones who prefer reading text or listening to it in an audio format using a corresponding function.
- Improvers – the ones who prefer taking practice tests, which includes entering answers, result checking, fixing if it's needed and saving the final result.
- Media-content interactors – the ones who prefer watching video, listening to the additional audio-tracks, watching gallery and switch slides within Opiq chapters.
- Formal tests passers – one who prefer passing formal tests, that are obligatory and given by the teacher; these students usually avoid other types of interaction with Opiq.

Statistics of the frequency of each cluster are presented on the Table 1.

Table 1. Percentage of interaction clusters in each grade

Cluster	3rd grade	6th grade	9th grade
Improvers	50.76%	63.73%	45.57%
Readers	15.91%	6.86%	5.06%
Media- interactors	22.73%	9.80%	2.53%
Formal tests passengers	10.61%	19.61%	46.84%

For a better understanding, a graph was also built, on which one can trace the trends in the predominance of a particular cluster (Fig. 6). As could be seen from the presented figure, the most common cluster of interactions with Opiq turned out to be Improvers. So, in the 3rd grade, this cluster occurs in 51% of students, 64% in the 6th grade and 46% in the 9th grade. It is worth noting the upward trend in the number of Formal tests passers: if in the 3rd grade they accounted for only 11% of the total number of students, then in the 6th grade their number has almost doubled-up to 20%, and in the 9th grade they already account for almost half of all students – 47%. Along with this, the percentage of students who were assigned to the media-interactors cluster is falling. In the third classroom such students were almost a quarter of all sample, in the 6th grade 10%, and in the 9th grade the number fell to 3 percent.

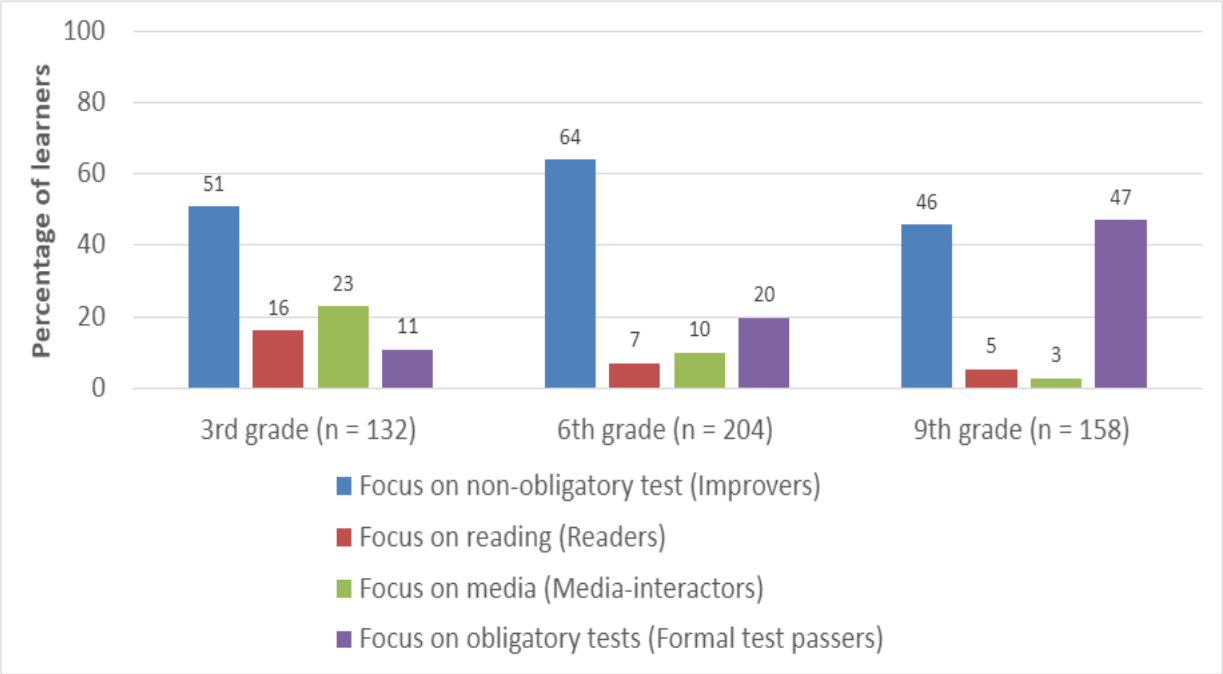

Fig. 6. Frequency of strategies for using E-Books.

Next stage was studying features of students' interaction with Science, Estonian language and Math workbooks in Opiq. The task of this stage was to identify the main strategies for interacting with E-Books in different disciplines.

In order to effectively explore behavioural patterns of interacting with Opiq in different grades (3rd, 6th and 9th grade) and different subject workbooks (Science, Math and Estonian language) sequence analysis and cluster analyses were used. TraMineR 2.2.7 package for R was used for sequence analyses: computing distances between sequences with optimal matching, identifying the most discriminating ones among them and visualizing results. For cluster analysis, the cluster package version 2.1.4 was used. Due to developers' recommendations and data requirements, Agglomerative Nesting (Hierarchical Clustering) algorithm was used.

The sequence analysis was applied to find most frequent sequences in 3rd, 6th and 9th grade students' interaction with Math, Science and Estonian Language workbooks.

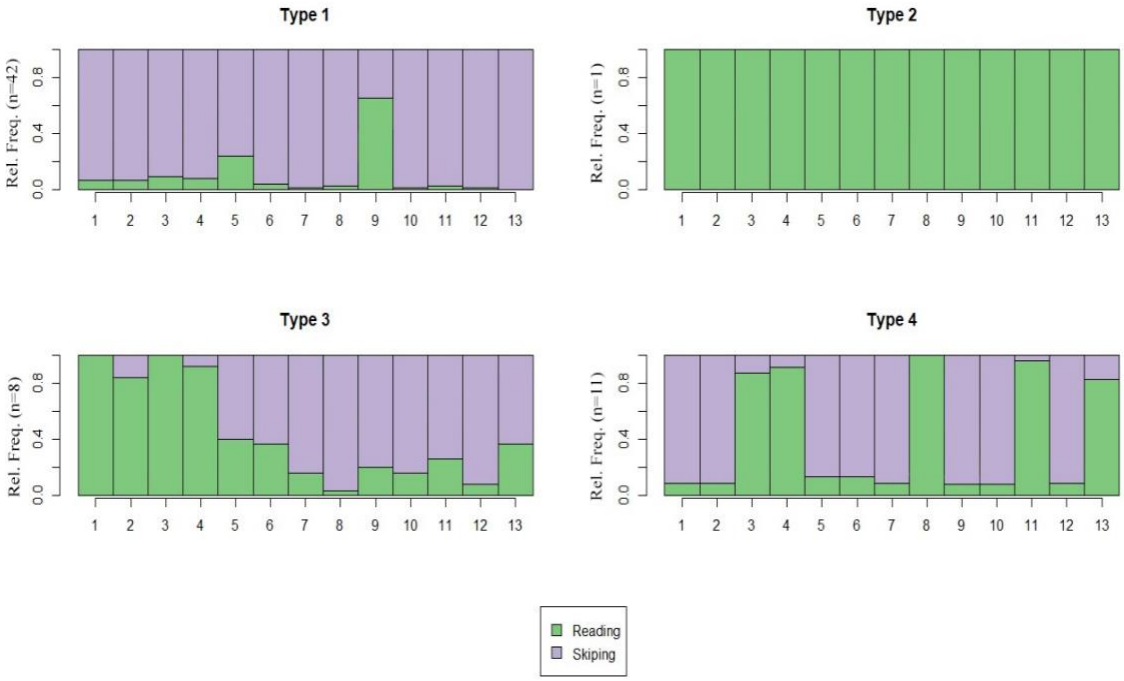

Fig. 7. 3rd grade Math workbook interaction clusters

Four types of clusters were identified based on the students' interaction with a workbook on Math in the 3rd grade. The first type is characterized by low chapters related frequency (only the 9th chapter was studied by the majority of students). Related frequency means the ratio of students who interacted with the chapter to the total number of students in the cluster and ranges from 0 (none of the students interacted with the chapter) to 1.0 (all students interacted with the chapter). The second cluster consists of one student who interacted with all chapters of the workbook. The third cluster is characterized by a high frequency of interaction with the first chapters and drops after about the first third. In the fourth cluster, almost all students interacted with chapters 3, 4, 8, 11 and 13, however, since there are large gaps between them, such interaction can be called non-systematic.

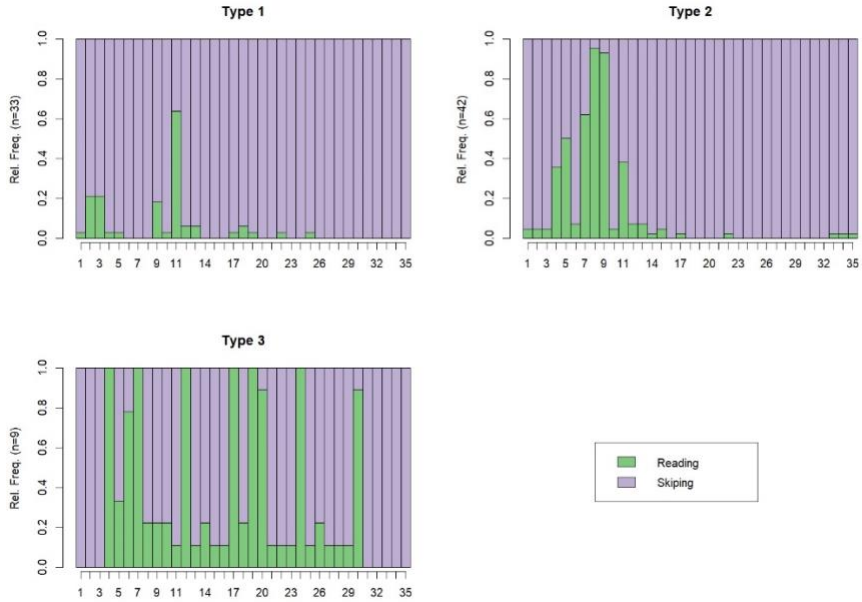

Fig. 8. 3rd grade Estonian Language workbook interaction clusters

Three types of clusters were identified based on the students' interaction with a workbook on Estonian language in the 3rd grade. The first cluster includes students who interacted with the workbook during only one time; also the last third of the workbook was not studied by any student. The second cluster is characterized by a high frequency of studying chapters in the first third of the workbook, but a complete lack of interaction after it. The third cluster is characterized by a frequency of 1.0 with 7 chapters at once (all students studied 4, 7, 12, 17, 19, 24 and 30 chapters), but the remaining chapters were studied much less frequently.

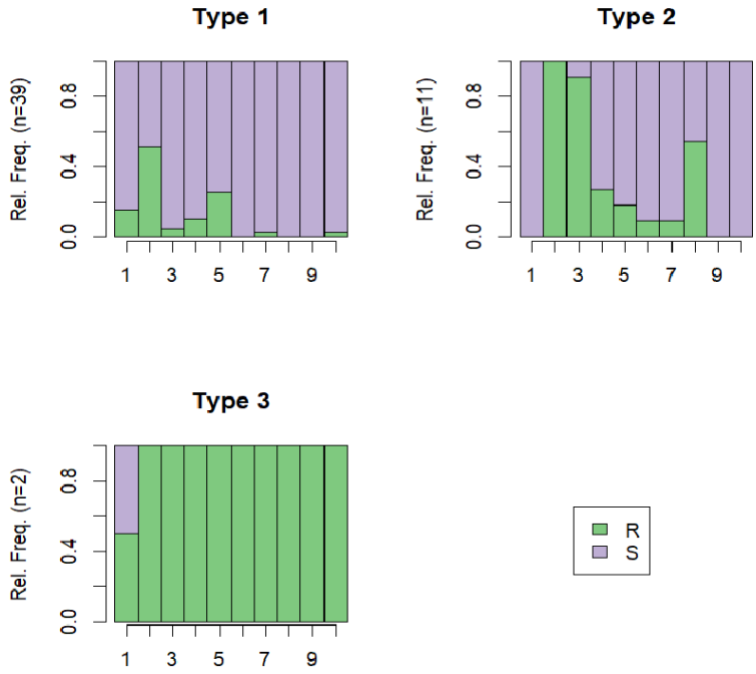

Fig. 9. 3rd grade Science workbook interaction clusters

Three types of clusters were identified based on the students' interaction with a workbook on Science in the 3rd grade. As can be seen in Fig. 6, the first is represented by students who interacted very little with the workbook – the second chapter was the most popular, while all the others had an extremely low level of relative frequency, and 6th, 8th and 9th chapters were not studied by any of the students within this cluster. The second type has a high rate of the first sections (chapters 2 and 3 were studied by almost all students of this cluster) relative frequency, after which the trend begins to fall. It is worth noting that the first chapter is introductory and, apparently, was not used by the teacher in the educational process. The third cluster consists of only two students who, however, have studied the entire content of the workbook.

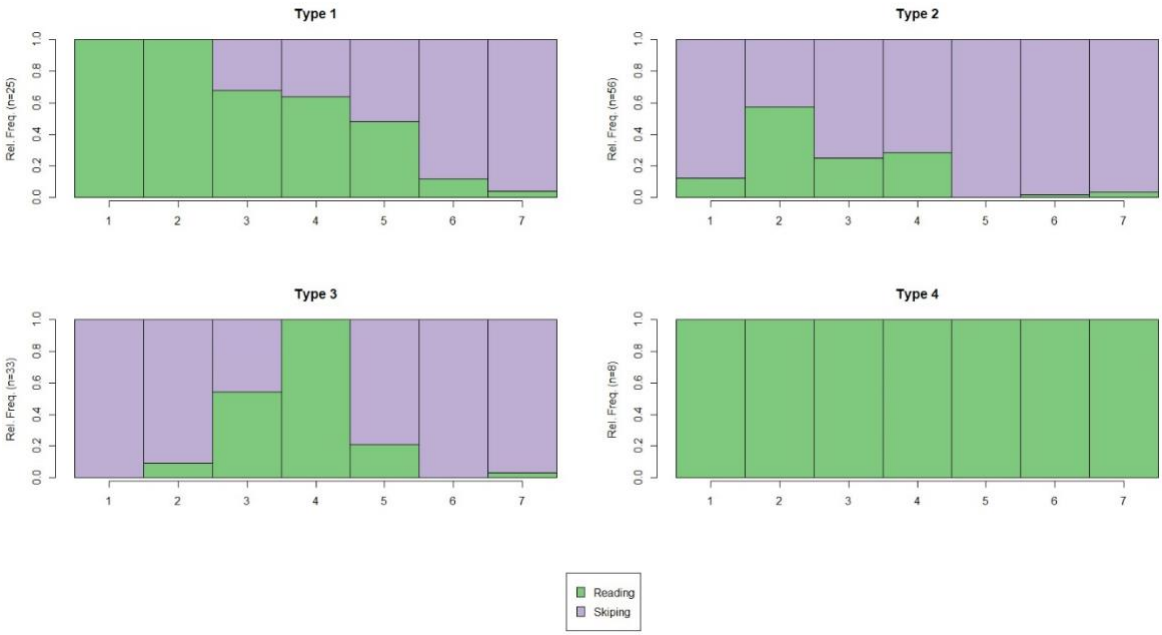

Fig. 10. 6th grade Estonian Language workbook interaction clusters

Four types of clusters were identified based on the students' interaction with a workbook on Estonian language in the 6th grade. All students of the first cluster interacted with the first two chapters of the workbook, after which the relative frequency of interaction gradually decreased. The second and third clusters are similar with the low relative frequency of studying chapters, the main difference is that if students of the second cluster interacted with chapters in the first half of the workbook, then in the third cluster the main interaction was in chapter 4. 8 students of the fourth cluster studied all 7 chapters of the workbook.

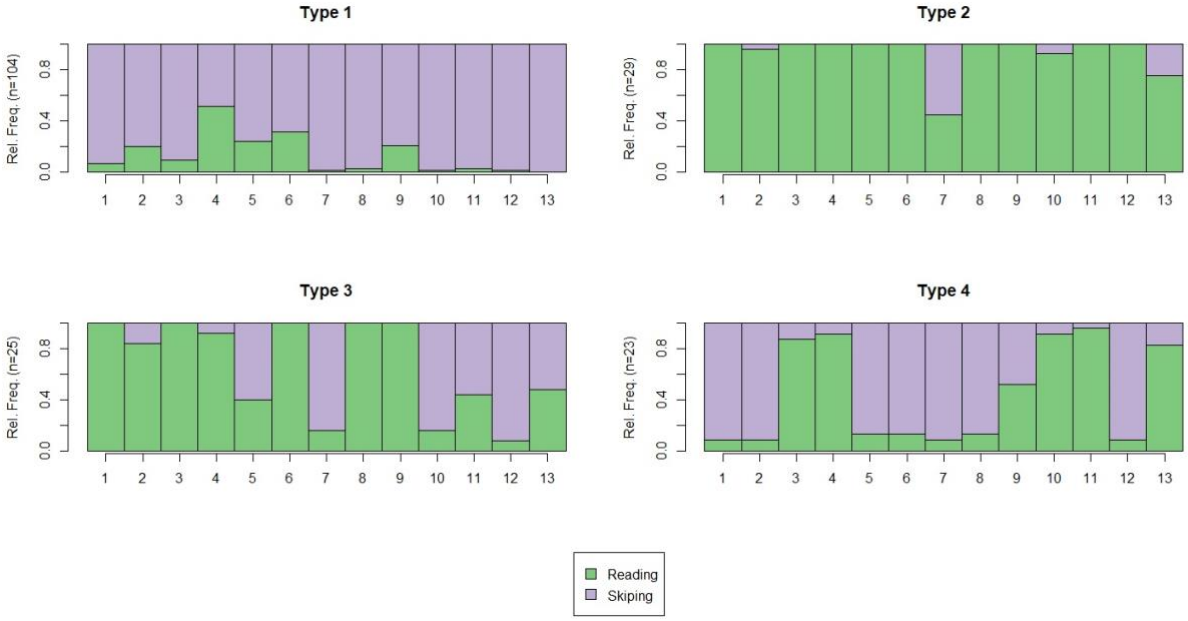

Fig. 11. 6th grade Mathematics workbook interaction clusters

Four types of clusters were identified based on the students' interaction with a workbook on Math in the 6th grade. The first type is characterized by a low chapter related frequency for all chapters, which indicates a one-time use of Opiq. The second cluster included students who

interacted with almost all chapters of the workbook (with the exception of the 7th chapter). The third cluster is characterized by a high relative frequency of interaction with chapters in the first two thirds of the workbook and a decrease in the last third. The fourth cluster is characterized by non-systemic interaction with chapters at the beginning and end of the workbook.

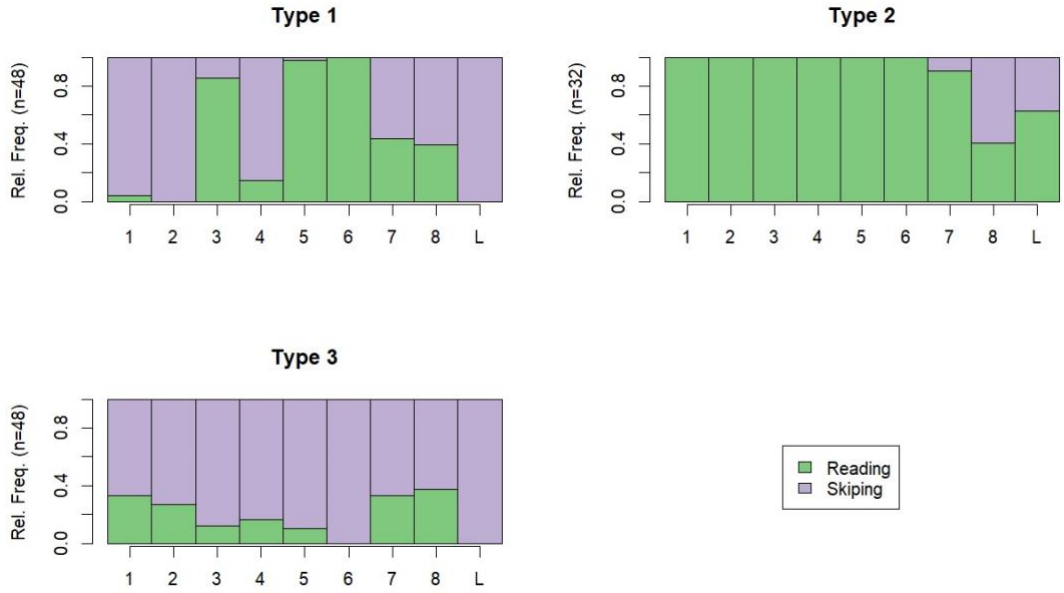

Fig. 12. 6th grade Science workbook interaction clusters

Three types of clusters were identified based on the students' interaction with a workbook on Science in the 6th grade. The first cluster is characterized by a high relative frequency of students who interacted with chapters in the middle of the workbook (chapters 3, 5 and 6), while not interacting with the first and last chapters. The second cluster is characterized by interaction with almost all chapters of the workbook (except for the chapter 8 and additions). The third cluster consisted of students with a low relative frequency of interaction with workbook chapters.

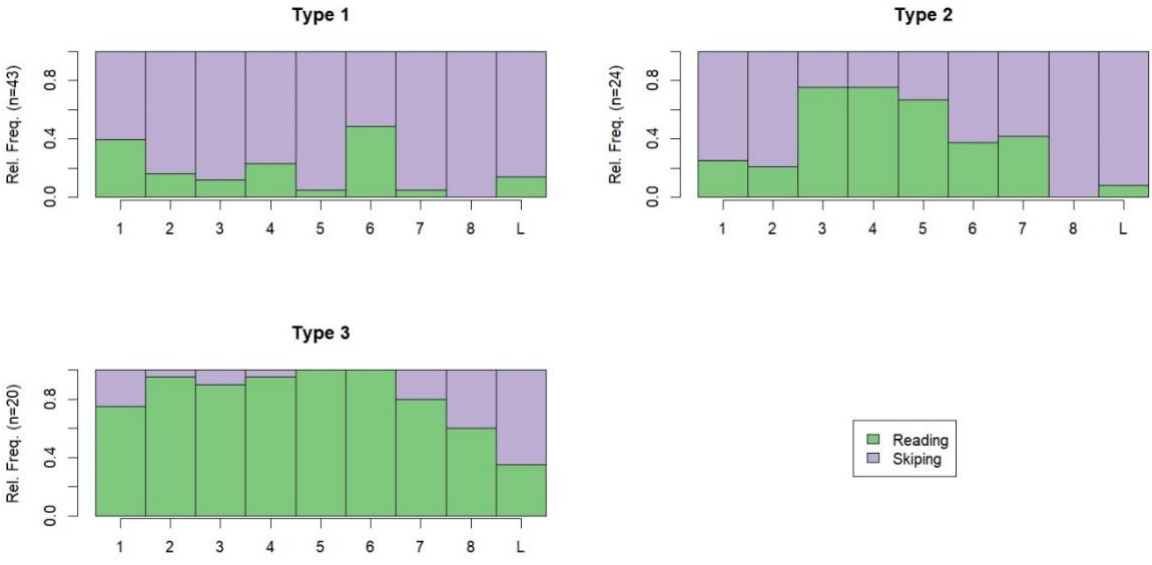

Fig. 13. 9th grade Science workbook interaction clusters

Three types of clusters were identified based on the students' interaction with a workbook on Science in the 9th grade. The first type is characterized by low chapters related frequency across all chapters. The second cluster is characterized by an average level of relative frequency of interaction (about 0.5) with 3-7 chapters, while interaction with the first and last third of the workbook was much lower. The third cluster is characterized by a high (more than 0.8) relative frequency of interaction with chapters in the first two thirds of the workbook, with a drop trend observed after the 6th chapter.

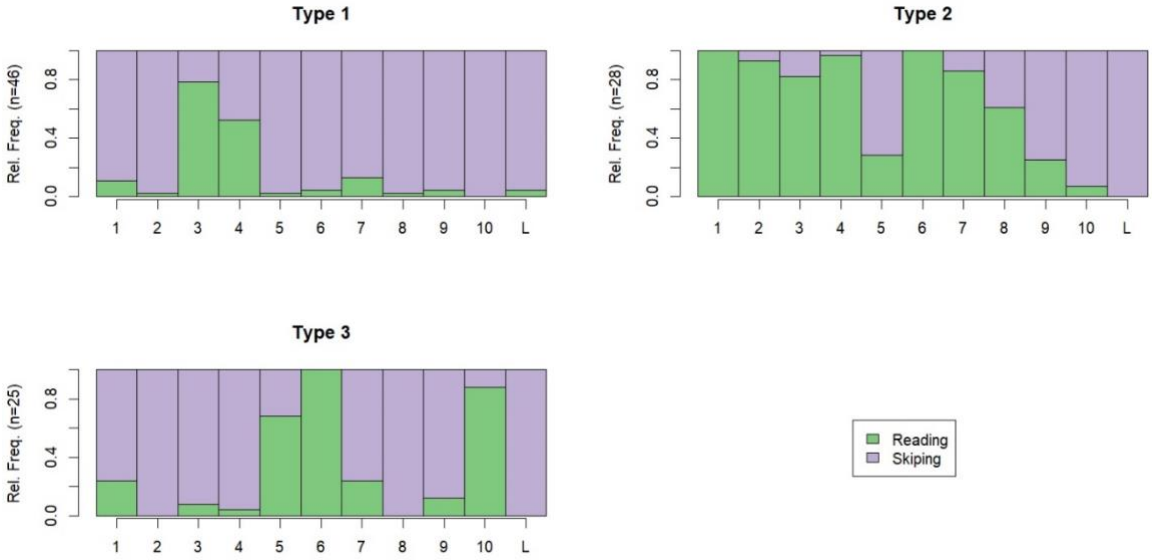

Fig. 14. 9th grade Math workbook interaction clusters

Three types of clusters were identified based on the students' interaction with a workbook on Math in the 9th grade. The first cluster is characterized by a low relative frequency of interaction with almost all chapters except chapters the 3rd and the 4th ones. The second cluster is characterized by a high level of relative frequency of interaction with chapters in the first half of the workbook (with the exception of a low level of interaction with the 5th chapter) and a gradual drop in frequency after the 7th chapter. The third cluster consisted of students with a high relative frequency of interaction with several chapters of the workbook, but not characterized by a systematic interaction.

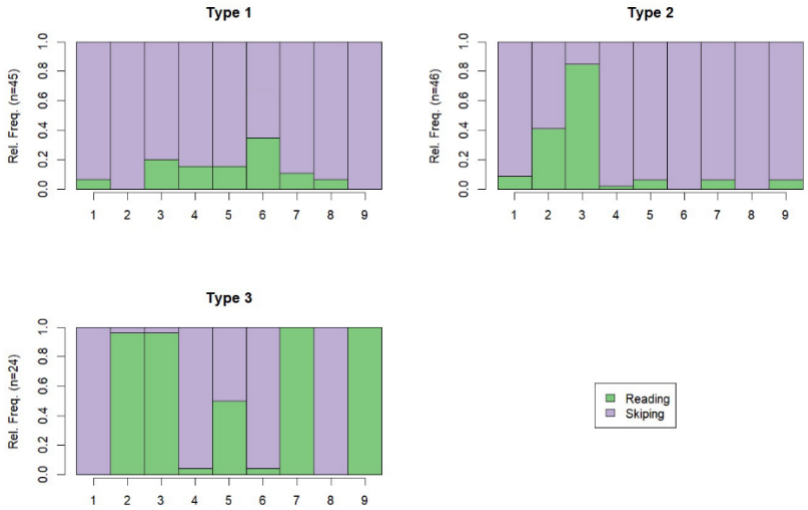

Fig. 15. 9th grade Estonian language workbook interaction clusters

Three types of clusters were identified based on the students' interaction with a workbook on Estonian language in the 9th grade. The first cluster is characterized by a low relative frequency of interaction with almost all chapters of the workbook. The second cluster consisted of students who interacted mostly with the second and third chapters. The third cluster consisted of students with a high relative frequency of interaction with several chapters of the workbook, but not characterized by a systematic interaction.

Next step was to create a typology of clusters to generalize strategies of students' interactions with the workbooks in all grades and subjects. Therefore five strategies were defined:

- "One-time use" describes a pattern of students who have interacted with Opiq only in a frame of one chapter.
- "Unsystematic use" includes students who have interacted with Opiq several times, although they interacted with unrelated chapters with gaps between them.
- "Fast drop" cluster describes an interaction habit, when related frequency of chapters' interaction significantly goes down approximately after the first third of a workbook.
- "Late drop" cluster is based on the similar pattern, the main difference is that drop happens approximately after the second third of the study year.
- "Systematic use" cluster describes an interaction with every chapter in the workbook.

Frequency of strategies in using E-Books in different grades and subject areas is presented on the Fig. 16. The main trend in the 3rd grade is the almost complete absence of students who have studied at least half of the Estonian, Science or Mathematics workbook. However, if in the case of the Estonian language workbook the most frequent cluster was Fast drop (53%), then in other subjects, students most often interacted with only one chapter (68% in Mathematics and 61% in Science).

In the 6th grade, students also often study only one chapter in the Estonian language workbook, however, 8% have studied it completely. The interaction with the Science workbook can be considered the most detailed – 29% of students studied it completely.

Late-Drop cluster includes 31% of the students in the case of studying Math and 25% in the case of studying Science. It is worth noting the lack of students who have studied the entire workbook, as well as the high number of Unsystematic Users – 22% in Estonian Language, 26% in Mathematics and 30% in Science. However, One-time use remains the most common strategy in each of the three items.

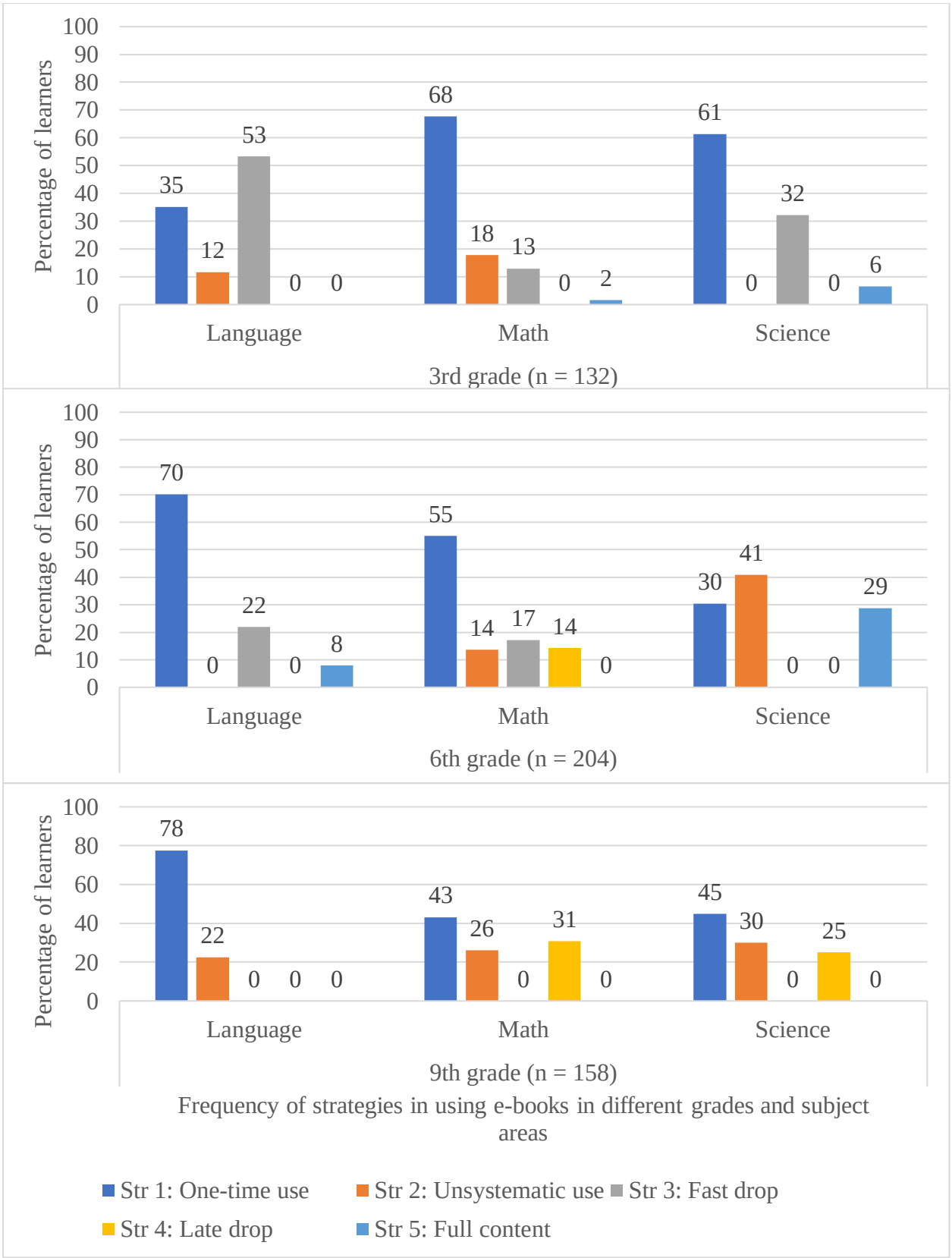

Fig. 16. Frequency of strategies in using E-Books in different grades and subject areas

Conclusions

Presented results are the first attempt to use two innovative technologies that are just beginning to gain popularity in pedagogy: process mining and sequence analysis. The current study of student interaction had several limitations:

- the information was collected without the participation of scientists, i.e., without developing research design;
- some interactions were not included in the final dataset (in particular, there is no information about swaps and scrolls, based on which reading process could be analysed);
- dataset is not homogeneous; some students interacted with Opiq for only a couple of minutes, and some for more than 10 hours;
- information about teachers' instructions devoted to Opiq (in other words, whether its use was mandatory) is missing.

However, process mining and sequence analysis algorithms seem promising for studying the Opiq log data, provided that the shortcomings of the previous data collection are eliminated.

References

Blummer, B. A., & Kenton, J. M. (2020). A systematic review of E-books in academic libraries: Access, advantages, and usage. *New Review of Academic Librarianship*, 26, 109–179.

Bogarín, Alejandro & Romero, Cristóbal & Cerezo, Rebeca & Sánchez-Santillán, Miguel. (2014). Clustering for improving Educational process mining. *ACM International Conference Proceeding Series*. 11-15. 10.1145/2567574.2567604.

Bruno Emond, Scott Buffett (2015). Analyzing Student Inquiry Data Using Process Discovery and Sequence Classification. *EDM*: 412-415

Brzinsky-Fay, C. (2022). NEET in Germany: Labour Market Entry Patterns and Gender Differences. *The Dynamics of Marginalized Youth* (pp. 56-86). London: Routledge. https://doi.org/10.4324/9781003096658-3

Cerezo, R., Bogarín, A., Esteban, M. (2020). Process mining for self-regulated learning assessment in e-learning. *J Comput High Educ* 32, 74–88https://doi.org/10.1007/s12528-019-09225-y

Herianto, Wilujeng, I. & Lestari, D.P. (2022). Effect of interactive multimedia e-books on lower-secondary school students' curiosity in a Science course. *Educ Inf Technol* 27. https://doi.org/10.1007/s10639-022-11005-8

MacNish, Jean & Bate, Frank & Stewart, Nigel. (2017). Using e-textbooks to support problem-based learning in science: Learning from the journey. 111-116. https://doi.org/10.1145/3175536.3175550.

Noor, A.B., Embong, A.M., & Abdullah, M.R. (2012). E-Books in Malaysian Primary Schools: The Terengganu.

Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Teck Kiang Tan, Lakshminarayanan Samavedham. (2022). The learning process matter: A sequence analysis perspective of examining procrastination using learning management system, *Computers and Education Open*, Volume 3, https://doi.org/10.1016/j.caeo.2022.100112.

Tim F. Liao, Danilo Bolano, Christian Brzinsky-Fay, Benjamin Cornwell, Anette Eva Fasang, Satu Helske, Raffaella Piccarreta, Marcel Raab, Gilbert Ritschard, Emanuela Struffolino, Matthias Studer. (2022). Sequence analysis: Its past, present, and future, *Social Science Research, Volume 107*. https://doi.org/10.1016/j.ssresearch.2022.102772.

W. Aalst and A. Weijters (2004). Process mining: A research agenda. Computers in Industry

Yaya, J. A. (2015). *Prospects and Challenges of E-Books in School Media Services in Nigeria : The Way Forward About EBooks*, 1(3), 92–98.