Investigating students’ journey through a computer science program using exam data: three new approaches

Madeleine Lorås
NTNU
Hallvard Trætteberg
NTNU
Kshitij Sharma
NTNU

Abstract

A computing student will over the first three years of their studies complete approximately 20 exams and even more attempts due to failures and retakes. Details about all exam attempts are stored in a national database called Common Student System (Felles studentsystem, FS). Although access to FS, in general, is restricted, anonymized data about exam attempts can be provided for research and is a potential goldmine of data that, if used right might be a useful tool for educators and teachers. In this study, we explore this data in an attempt to conceptualize three new approaches to assessing student performance. Firstly, we relate students' final grade point average (GPA) to their performance in all courses in the first two years. Additionally, we propose a new indicator of student performance called "struggle factor," which is calculated using the number of exam attempts. Lastly, we investigate how students perform in different course subjects and types. Both the proposed use of FS data and the new approaches to performance indicators are relevant for educators wanting to understand the educational design of a study program and the students’ journey.

Key words: Performance, Exam data, Educational design, Computing education

Introduction

In this study, we explore the FS data on our own students in an attempt to gain insight into student performance, both success, and failure. The overall research objective was to investigate what ways general institutional data can be used to assess student performance and aspects of the educational design. The research questions were as follows:

- How can FS data be extracted and managed ethically and practically?
- What insight into student performance can FS data give that could help in educational design?

In this context, student performance means both academic success, failure, and progression of the individual student, as well as the context of such behavior. For example, grades, attempts, and failed exams are indicators of performance, while the context relates to what kind of courses are challenging (e.g., computer science, mathematics or unrelated courses, project work or individual).

An essential underlying motivation for this work was the hunt for analytic tools on the study program level that can give educators insights into the educational design of study programs, both regarding practical implementation and quality assurance. Educators in this context means teachers and lecturers in various courses, as well as study program leaders, education managers and administrators working with teaching quality. In Norwegian higher education, as well as the rest of the world, universities and colleges are required to adhere to corporate governance, which entails finding good educational indicators to report, as well as useful planning and management tools. This contribution is an attempt to use already existing data in a new way, from educators who have experience “from the inside” of study program development and management.

Methodology

This research is designed as a retrospective quantitative study [3, 6]. The FS database allows us to go back over ten years, which makes it a useful dataset for a longitudinal study. Since the data is from the past, and any new data will be looking to the past as well, it must have a retrospective view. Lastly, the data is only quantitative as the exam submissions are not part of the database. Such studies are considered useful for establishing relationships and enables the dynamics of change and flow to be caught [3].

Population and program

The study investigates two undergraduate study programs; program A and B. Both these programs accept approximately 150 students each year and are considered relatively hard to get into, although the admission grade point average (GPA) for program B is significantly higher than A [8]. The gender balance for program A is consistent around 20%, while Program B has about 30% (+- 5% from year to year).

As you can see from Table 1, the educational design of the two programs are very similar. The main differences are related to the mathematics requirements and the number of optional courses. Program A has fewer mandatory mathematics courses than program B, which in turn opens up for more freedom to choose courses. For the first two years, however, the course plan is fixed.

Table 1: Educational design of Program A and B

Course	**General description**	**Semester in program**	
		Program A	Program B
Philosophy	Philosophy	1	1
CS1	Intro to programming, Python	1	1
Web 1	Introduction to web technology	1	
Math 1	Basic mathematics level 1	1	1
Discrete math	Discrete mathematics	2	1
CS2	Object oriented programming, Java	2	2
Arduino lab	Programming and technology, Arduino	2	2
Networks	Networks	2	4
Math 2	Basic mathematics level 2		2
Circuits	Electric circuits		2
Computers	Computers and digital design	3	3
Algorithms	Algorithms and data structures	3	3
SD1	Software development 1	3	
Security	Security in ICT systems	3	
Digital society	Digital society	3	
IoT lab	Programming and technology, IoT		3
SD2	Software development 2	4	4
HCI	Human computer interaction	4	4
DB	Databases	4	4
Statistics	Statistics		4
Web 2	Web development 2	5	5

Prog.lang	Programming languages	5	5
Low level	Low level programming	5	5
Cognitive arc	Cognitive architectures	5	5
Infosys	Information systems	5	5
AI1	Intro to AI	5	5
Info retrieval	Information retrieval	5	5
Software arc	Software architecture	6	6
OS	Operating systems	6	6
AI2	AI methods	6	6
Compilers	Compiler construction	6	6
Data mining	Data warehouses and data mining	6	6
Bachelors	Bachelors project	6	
Security 2	Software security	6	
Analytics	Intelligent text analytics	6	
Management	Technology management		6
Physics	Basic physics		6

Key:

Unique for one of the programs	The same for both programs	Same course, but different semesters

Data collection

In order to collect the necessary data in an ethical yet practical way, certain precautions had to be made. Although exam results are considered public information, the detailed data we were aiming for is not readily available. Therefore, we decided that there was a need to use the FS database; however, there were two main issues with that. Firstly, academic staff does not have direct access to FS for valid privacy reasons. Hence, a member of the administration who does have such access would need to extract the data on our behalf, which leads to a second issue: we would be collecting the data without informed consent from the students. With these issues in mind, we prepared a plan for anonymizing the data, which was approved by Norwegian Centre for Research Data (NSD).

The data set was extracted from FS by a member of the administration. The data set at this point uses the student's student number as an identifier, and in order to anonymize the data, this was replaced by a random identifier. This anonymized dataset did not contain any directly identifiable personal information and was therefore satisfactory to use in research. The random identifier does not relate to the original student number; hence, the researchers had no way of identifying a student.

Data set

The data set from FS contains exam results from students in two different study programs, classes starting in 2011-2015. That is five classes of students who have completed the first three years. In total, that is 1 809 students, who have completed 38 024 exams in 44 unique courses. For many study programs, students will have more and more freedom to choose courses as they progress. Therefore, this study limited the time span to the first three years and only looked at mandatory courses in those three years.

The data extracted from FS consisted of every exam attempt so that a student will appear several times in the data set. However, we are in this case, also interested in following a student as opposed to a course. Therefore, the data needed to be restructured so that every row in the dataset represented one student and their individual progression. For the purpose of analysis, it was useful to view both versions of the data, which is visualized in Figure 1:. Data set 1 shows the original data, exam-based, with each individual exam result per row. While data set 2 contains student-based data, with a student per row and corresponding courses. In addition, data set 2 includes individual variables such as GPAs, program, class, etc.

Figure 1: Visualization of the two different data sets

Data set 1		
Student 1	Course A	**Grade**
Student 1	Course B	…
Student 1	Course C	…
Student 2	Course B	…
Student 3	Course A	…
Student 3	Course C	…
…		
Student N	Course X	Grade

Data set 2					
Student 1	**Course A**	**Course B**	**Course C**	**…**	**Individual variables**
Student 2	…	…	…		…
Student 3	…	…	…		…
Student 4	…	…	…		…
Student 5	…	…	…		…
Student 6	…	…	…		…
…					
Student M	Course X	Course Y	Course Z	…	Individual variables

Variables

With the two datasets described above, there are many different variables that could be calculated in order to learn more about student performance and educational design. Performance variables can generally be viewed at four levels: institution, study program, course, and individual. Traditionally, institution and program-level variables are concerned with throughput and dropout rates. Courses are commonly assessed by pass/fail rates and grade distributions, while individuals are often measured by GPA and failure rates. In this study, we are only able to address program, course and individual levels and will focus on individual performance. Table 2 gives an overview of possible variables.

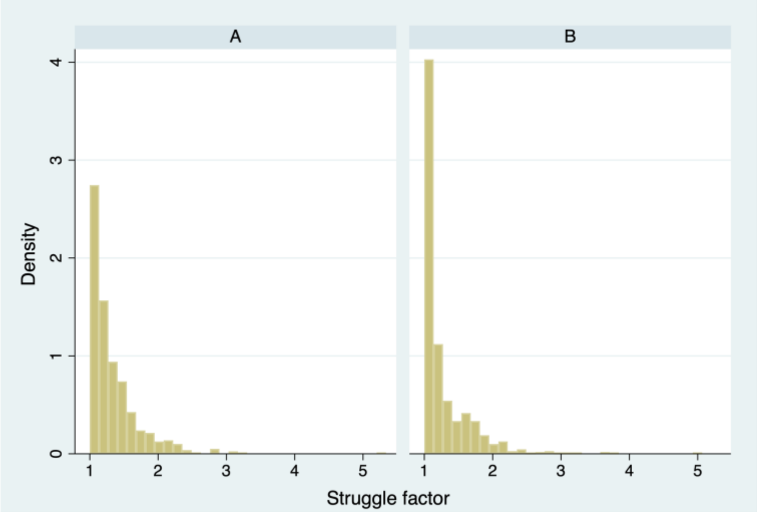
Table 2: Overview of possible performance variables

	Overall	**Year**	**Semester**	**Program**
Program	Within a program: the average grade overall	Within a program: the average grade for one year of exams	Within a program, the average grade of each semester	Drop-out rate
Course	For each course, the overall average grade for all years Pass/fail rates	For each course, the average grade for each year taught	For each course, the average grade for each year taught	For each course, the average grade for one year within a program
Individual	For each student, what is their overall average grade	For each student, what is their average grade per year	For each student, what is their average grade per semester	

GPA

For the analysis in this study, we have chosen to examine two GPA variants: the overall GPA for the student and the 6th semester GPA. The overall GPA was calculated by averaging the individual student's grades for all courses in the first three years. The 6th semester GPA, on the other hand, only consisted of grades in semester six. The reason for wanting to examine both these GPA variants was that the 6th-semester courses are on a higher level and therefore should represent the final competence level. Whereas the overall GPA includes all courses, and in that way, might include low points periods in the students' academic progression. In other words, looking at only semester six GPA will exclude the possibility that students needed a few semesters to "get the hang of things."

Figure 2: Histograms of struggle factor for programs A and B.



The struggle factor

In addition to the GPA indicators, we have also explored a new approach. We have observed as educators that some students need several attempts to pass a course, or to get an acceptable grade. We were, therefore, curious if this could be a possible approach to performance that does not directly relate to GPA. To calculate this “struggle factor” we used the following formula:

$$Struggle\ factor = \frac{Total\ number\ of\ exam\ attempts}{Total\ number\ of\ courses\ completed}$$

Calculating the struggle factor like this for our data set resulted in a number between 1 and 5.2, for each student in the dataset. A struggle factor of 1 means that the student had the same number of exam attempts as courses completed, which indicates no struggle. A struggle factor of 5.2, on the other hand, indicates that the student has attempted 5.2 more exams than courses completed. In Figure 2 you can see that most students have a struggle factor of between 1 and 2. However, some students have a higher struggle factor, which is why the right tale of the histogram is so long.

It is important to note here that the assessment regime at NTNU allows students to retake exams for different reasons, which is why the number of attempts can be higher than the number of courses. Firstly, students who take an exam and fail will, of course, have the option to retake that exam. The same goes for students who are unable to take an exam because of medical reasons. A second possibility is retaking an exam to improve a grade, which is also possible. The last possibility is what is often referred to as a “tactical retake”, where students can intentionally choose not to receive a grade in an exam in order to qualify for the retake. This is often something students can do when they believe they will not get the grade they aimed for, or for some other reason, wanted access to the retake exams. Retake exams are organized in the summer, and students only have access to these if they failed or did not receive a grade. If a student wants to improve a grade, they would have to take the exam the next time the course is offered.

Course types and sets

A different perspective on performance is looking at how students perform in different course types. On the highest level, we can group the course into computer science (CS), mathematics and other courses (physics, philosophy, management, etc.). From Table 1, we can see that there are 30 CS courses, four mathematics courses and three other courses across both programs. Furthermore, some courses can be viewed as sets that build on each other. For example, CS1 and CS2 or the different mathematics courses.

Analysis and results

When it comes to exploring ways to evaluate student performance using FS data, we will in the following section present three approaches. Firstly, we visualize the first two years of courses with respect to the overall and 6th semester GPA. Secondly, we explore the struggle factor and what insights it can give into student performance. Lastly, the courses were grouped by type, and the differences evaluated. Both the method of analysis and the results will be presented for each approach.

Study program heat maps

The dependent variables for these approaches are the overall GPA (calculated using grades from all three years), the 6th semester GPA and the struggle factor. The independent variables in this analysis are the grades received in the mandatory courses of the first two years. Since the third year includes several optional courses, we were not able to analyze the data in a coherent way. Additionally, pass/fail courses were omitted since there was no variance in grade. For this analysis, we used data set 2.

A Pearson's correlation matrix was calculated to model the two study programs [2]. In order to visualize this, the correlations are summarized with a heat map as shown in Figure 3 and Figure 4, for programs A and B, respectively. In this heatmap, darker colors indicate a stronger positive correlation, while lighter colors are negative correlations. Every color shade relates to a specific Pearson correlation coefficient (r), as labeled on the right. Note that programs A and B have different coefficients. All variables are listed on both axes', hence creating the matrix; however, we have only included the upper half in order to make the map more readable. The first three variables from the top are the dependent variables; overall GPA, 6th semester GPA and struggle factor. Notice that since the struggle factor ranges from 1-5.2, meaning a high score equals a high level of struggle, the correlations are naturally negative. The remaining variables are the various courses in chronological order from first semester courses on the left/top to fourth-semester courses on the right/bottom (see Table 1 for details about the courses). In the following sections, we list the results for each dimension of the analysis; GPA, struggle factor and course types and sets.

GPA results

- The 6th semester GPA and overall GPA are highly correlated, as indicated by the dark color of the top left correlation box.
- Most courses seem to have a high correlation to overall GPA, as indicated by all the dark correlation boxes. Exceptions will be further discussed in the course type section.
- The correlations between courses and the 6th semester GPA are not as strong as the overall GPA, as indicated by the fact that the whole row for the 6th semester GPA is lighter than the overall GPA.
- For program A the relation between 6th semester GPA and the different courses is weaker than for Program B, as indicated by the lighter color of the 6th semester row.

Figure 3: Heat map plot of the first two years for Program A. N=265

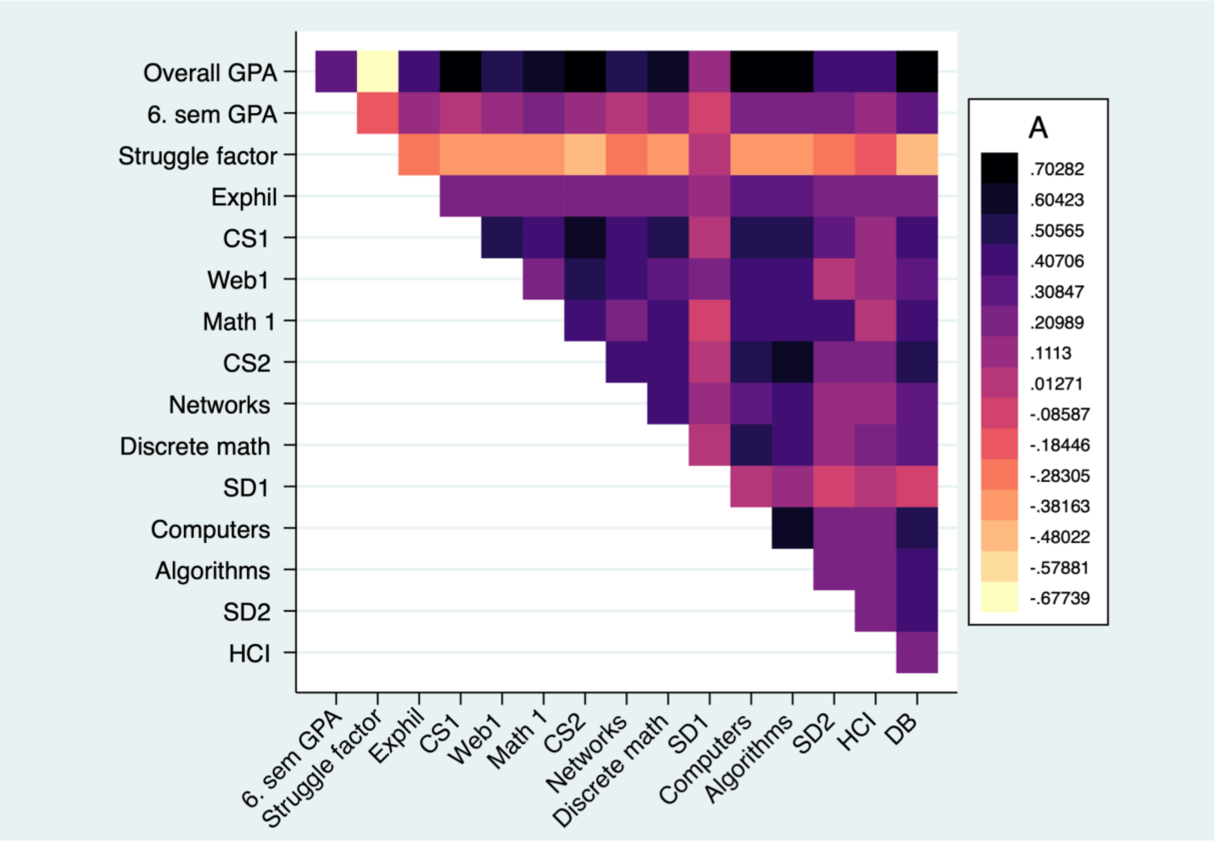
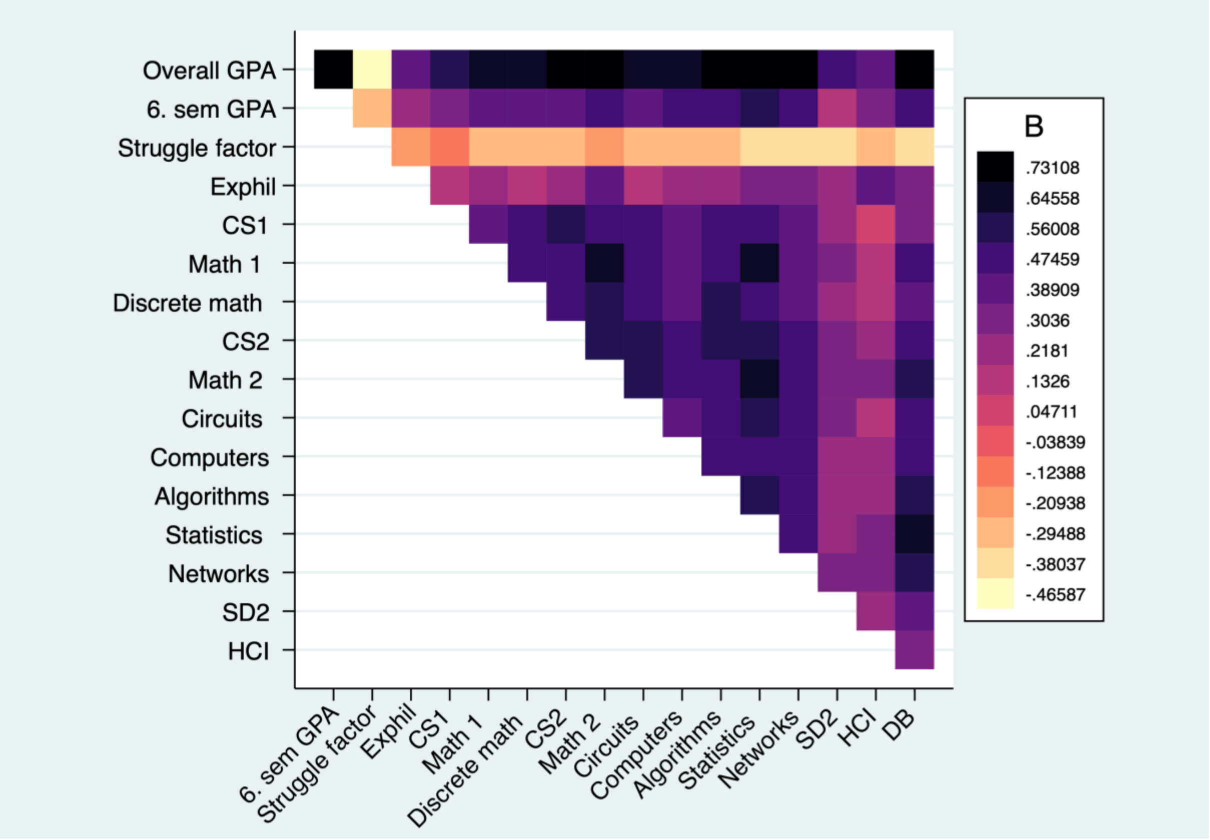


Figure 4: Heat map plot of the first two years for Program B. N=270



Struggle factor results

- There is a strong correlation between struggle factor and overall GPA, as indicated by the strong light color of the correlation box (second from the left, top row).
- The correlation to 6th semester GPA is not as strong as indicated by the less light correlation box below.
- The correlation to the different courses gradually becomes stronger as the students progress through the program as indicated by the gradually lighter correlation boxes (more distinct for program B).

Course types and sets results

Using the heat map to look at correlations between courses, we can learn about how various course types and sets relate or not. This is done by finding the intersecting correlation box for the two courses in question. In addition, one can investigate obvious outliers, that is areas which are lighter or darker than the surrounding areas. In the following, we have extracted the expected and not so expected correlations from this analysis.

Expected strong correlations:

- CS1 and CS2 have a high correlation.
- The different mathematics courses have a high correlation.
- Discrete mathematics and algorithms have a high correlation.

Notable weak correlations:

- Philosophy does not strongly relate to any other courses.
- SD1, SD2 and HCI do not strongly relate to other courses, GPAs or struggle factors. These can be seen as lighter columns/rows.

Discussion and related work

The overall research objective of this study was to investigate how general institutional data can be used to assess student performance and aspects of the educational design. We have explored variants of GPA, developed a potential new struggle factor and looked at how course types and sets relate, using exam data gathered from FS. In the following sections, we will discuss the application and implications of the results presented above, the methods we have used and the related research.

Using exam data to assess performance

To answer the research question of *what aspects of student performance can be evaluated using FS data*, we have examined overall GPA, 6th semester GPA, the struggle factor as well as course types and sets. Considering the two GPA variants, the overall GPA seems to be a better indicator than 6th semester GPA. This is based on the fact that overall GPA produces higher correlations to the courses and seemed to be more consistent. The use of GPA to indicate academic performance and success is common in educational research, although as the literature review done by York et al. found it does not always measure learning or growth in cognitive abilities [9]. In the current study, however, measuring performance or success with GPA is not directly looking at the summative learning growth. We have used GPA to indicate academic progression as a journey through a study program, where the actual grade in itself is not assessed, but the relation of a student's GPA to other grades. Previous work on performance in computing education research has also used GPA. In some cases to measure the effectiveness of certain educational approaches, teaching technologies or study behaviors [4, 5, 7]. In other cases, grades or GPA was used to differentiate students into high and low performers [4, 7], which was not the goal of our approach.

The struggle factor was a new approach to assessing student performance in a program with mixed results. To the authors' knowledge, similar approaches have not been reported on. On the one hand, the struggle factor seems to be a valid indicator of performance since the correlations to GPA and other courses is consistently high. On the other hand, it is difficult to interpret these results because the struggle factor is somewhat unprecise. It does not differentiate between retaking an exam because of a failed previous exam, tactical retakes and improving a grade. Nevertheless, the process of calculating and analyzing this variable provided some useful insights. In retrospect, the authors have discussed only including retakes based on a failed grade or calculating the time between the first attempt and the first passing grade. An additional aspect of the struggle factor is considering if it should be interpreted as a linear relationship. A possible perspective would be to square the number of attempts, hence emphasizing the students who have many retakes, which would, in fact, indicate struggle more. Lastly, this process sparked the idea of creating a struggle factor for courses as well as for students.

When it comes to course types and sets the heat map analysis provided some very interesting results. We would expect all courses to correlate to GPA as a program is designed with the intention of building the students' knowledge and skills over time. Therefore, the noticeable discrepancies found for philosophy, SD1, SD2 and HCI are striking. There are several possible explanations for these inconsistencies. In the case of philosophy, it is a course that does not "fit in" to a CS program, and it is therefore understandable if students treat that course differently and thus perform differently. In future work, it would be interesting to examine the performance of students in these *other* courses intended to broaden the knowledge of students. The results for SD1, SD2 and HCI, on the other hand, are harder to explain. One possibility is a divergence in content and assessment regime; in other words, a lack of constructive alignment [1]. Another possible explanation could be the structure of the course and assessment. We know that these courses are largely based on project and/or group work, in contrast to the other courses, which are based on individual assignments. Nevertheless, these explanations are based on tacit knowledge about the courses and programs. The key takeaway here is that this process of analyzing a program can highlight courses that need further investigation.

Using exam data from FS

A second aspect of the current research was *how FS data could be extracted and managed in an ethical and practical way*. We have found that by anonymizing the exam data from FS using a random identifier, the data can be analyzed in an ethical way. However, managing and analyzing the data in a practical and useful way provided more challenging. The fact that data goes back over many years is very useful, but many things change over time, which provided some challenges. Firstly, programs change over time and tracking these changes is hard. The changes can be due to course alterations, for instance, names or scope. An example of this is that there has been an introductory mathematics course in the first semester of both these programs for at least ten years. However, the name of this course has changed, as has the number of credits. Identifying and accounting for changes like this is time-consuming, but very important for such analysis'.

Limitations

The main challenge for this study is that the data set is so vast that the number of conceivable variables and methods of analyzation is very large. It is very possible that other tools would provide interesting results and answers to our research questions. Another limitation is that exam data in this form is an aggregated and summative indicator of very complex phenomena. Researchers should be cautious when drawing conclusions just based on the exam data. In our case, the fact that both authors are involved in the study programs analyzed can be viewed as a strength because it informs the analysis.

Conclusion

This paper has summarized an attempt to use available exam data to find new ways to evaluate and assess study programs and student performance. We have looked at two variants of GPA and found that the overall GPA seems to be the better indicator of performance throughout a program. In addition, we have proposed looking at the relation between exam attempts and courses taken as a struggle factor. However, this variable might need some fine-tuning for future work. Lastly, we have found that examining course types and sets provides useful information about the design of a program.

In conclusion, the use of correlations visualized by heat maps was found to be very informative and we aim to further explore how this approach can be used in the future. An important motivation behind this work was to find useful tools for study program leaders and educational managers, tools that can inform decisions about study program design and enlighten possible inconsistencies between courses. What can we actually change and where are the possible rooms for actions are important future questions. At the student level, identifying and predicting challenging courses or transitions could help educators to implement impactful changes.

References

[1] Biggs, J. 1996. Enhancing teaching through constructive alignment. *Higher Education*. 32, 3 (Oct. 1996), 347–364. DOI:https://doi.org/10.1007/BF00138871.

[2] Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge.

[3] Cohen, L. et al. 2002. *Research Methods in Education*. Routledge.

[4] Edwards, S.H. et al. 2009. Comparing Effective and Ineffective Behaviors of Student Programmers. *Proceedings of the Fifth International Workshop on Computing Education Research Workshop* (New York, NY, USA, 2009), 3–14.

[5] Goold, A. and Rimmer, R. 2000. Factors Affecting Performance in First-year Computing. *SIGCSE Bull.* 32, 2 (Jun. 2000), 39–43. DOI:https://doi.org/10.1145/355354.355369.

[6] Johnson, B. and Christensen, L. 2012. *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. SAGE.

[7] Liao, S.N. et al. 2019. Behaviors of Higher and Lower Performing Students in CS1. *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (New York, NY, USA, 2019), 196–202.

[8] Lorås, M. et al. 2018. First year computer science education in Norway. *Proceedings from the annual NOKOBIT conference 2018* (2018).

[9] York, T.T. et al. 2015. Defining and Measuring Academic Success. *Practical Assessment, Research & Evaluation*. 20, 5 (Mar. 2015).