

UTRECHT UNIVERSITY  
Department of Information and Computing Science

---

**Applied Data Science master thesis**

**Analysing the performance of international students at  
the UU**

**First examiner:**

Gerard Barkema

**Candidate:**

Sebastiaan Barneveld 2665271

**Second examiner:**

Lynda Hardman

**In cooperation with:**

Jurre Stiekema

July 8, 2023

## **Abstract**

This study was done on behest of Utrecht University in order to find out where they stand on the performance of international students and if there are specific groups of students who are struggling the most. The Dutch Ministry of Education, Culture and Science published a report with the findings that international bachelor students have seen less academic success than Dutch students over the period of 2011-2021. This study shows that the data suggests that international bachelor students are performing worse at the UU too. But it also highlights that the missing data of international students shows a deviating pattern from the missing data of Dutch students. Furthermore, groups of international students who are obtaining less academic success than their peers are identified based on a combination of student characteristics and their study programme. This gives the UU specific groups of students who are more likely to require student aid in order to succeed. The methods outlined by this study can be used to find other groups, not just limited to international students, with the aim of improving the chances of academic success at the UU.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Factsheet October . . . . .	5
1.2	Greece & Geoscience . . . . .	8
1.3	Literature review . . . . .	8
<b>2</b>	<b>Data</b>	<b>11</b>
2.1	Dataset . . . . .	11
2.2	Data collection . . . . .	11
2.3	Data preparation . . . . .	13
2.4	Data description . . . . .	14
2.5	Data privacy and ethics . . . . .	14
<b>3</b>	<b>Method</b>	<b>15</b>
3.1	Factsheet comparison . . . . .	15
3.2	Student group identification . . . . .	16
<b>4</b>	<b>Results</b>	<b>17</b>
4.1	Factsheet comparison . . . . .	17
4.2	Student group identification . . . . .	21
<b>5</b>	<b>Conclusion</b>	<b>26</b>
5.1	Discussion & further research . . . . .	27
	<b>Bibliography</b>	<b>32</b>
<b>6</b>	<b>Appendix</b>	<b>33</b>
6.1	Factsheet . . . . .	33
6.2	Student groups and study combinations . . . . .	34

# 1. Introduction

Utrecht University (UU) is a highly acclaimed institution that attracts thousands of international students every year to participate in its English language academic Bachelor's and Master's programmes. The university offers an international environment with hundreds of exchange programmes and Europe's largest Summer School. The UU also offers joint Master's programmes developed as part of the Erasmus Mundus programme and an extensive range of English-language academic programmes, including 12 Bachelor's programmes, 109 Master's programmes and over 200 courses for exchange students. Top lecturers are regularly recruited from abroad to teach these programmes and maintain the University's prestigious reputation.

Recently the Dutch government published a factsheet suggesting that the national performance of international students has decreased to below the performance of Dutch students [1]. Whereas historically they have outperformed Dutch students. The department of international marketing offered a thesis project to investigate whether this trend holds for students at Utrecht University. The UU wants to keep its status as a prestigious University and be an attractive option for students looking to study abroad. Therefore it is in their interest to maintain a high graduation rate without compromising on the quality of the education offered.

The research questions are:

1. How does the national average performance of students compare to the performance of students at the UU?
2. What factors can identify groups of students whose graduation chances would benefit from the various aid programs the UU offers?

The following hypotheses for each research question were posited:

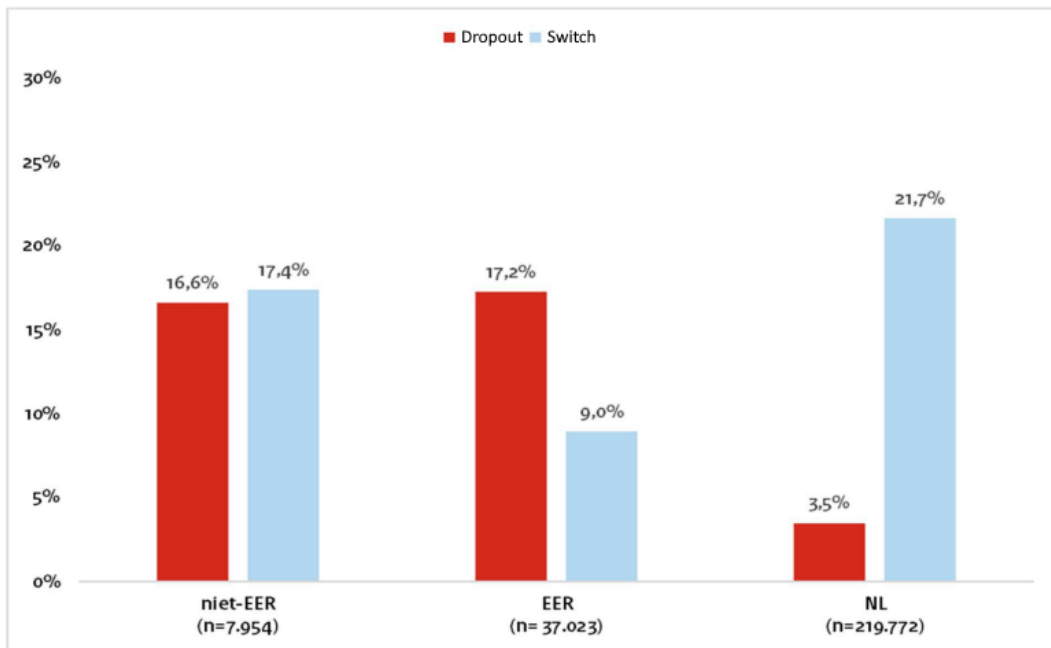
1. Native and international students at the UU perform better than the Dutch national average.
2. Certain backgrounds and study programme combinations will have a performance below expectations. Thus we expect that these groups will benefit from student aid.

## 1.1 Factsheet October

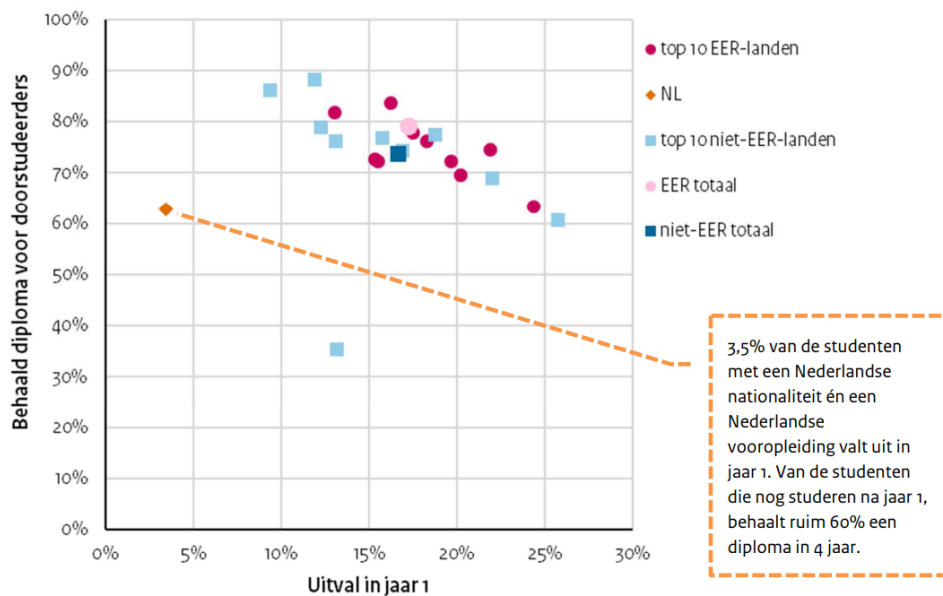
In October 2022 the Dutch ministry of Education Culture and Science published a report in which they discussed the study successes of international students pursuing a bachelor degree[1]. This report was about the average national performance of bachelor's students. The report suggested that international students at the university may not be performing as well as their domestic counterparts. Historically, international students have outperformed the Dutch students, which is why this is surprising.

International students are grouped into two different categories: Students from the European Economic Region (EER) and students from other countries (non-EER). This distinction is not only important in terms of geographic or cultural similarity, but because EER students pay the same tuition as Dutch students whereas for non-EER students it is significantly more expensive to study in the Netherlands.

Both EER and non-EER students drop out more often in all the years than Dutch students. Dutch students switch more often to a different Dutch degree. This can be explained by the fact that international students switch to degrees in other countries. However, switching to a foreign country is counted as dropping out by the ministry. In both EER and non-EER between  $\frac{1}{4}$  to  $\frac{1}{3}$  of the students start a degree they will abandon in the first year.



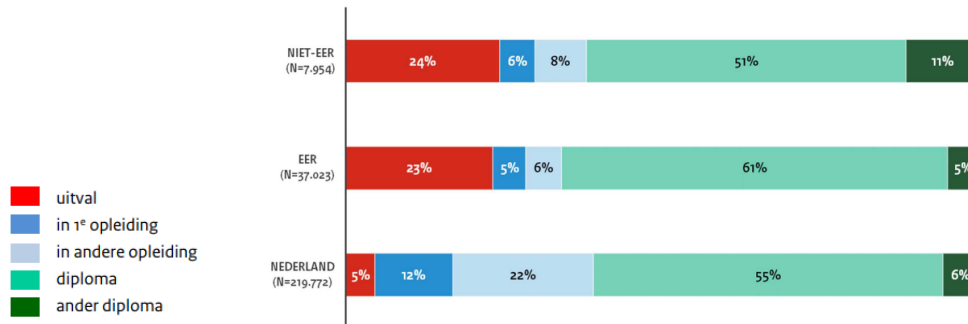
**Figure 1.1:** Percentage of first year students that drop out and switch within the first year averaged over the cohorts 2011-2017. The translation for 'niet-EER' is non-EER. Source: [1]



**Figure 1.2:** The correlation of first year students that drop out within the first year and the graduation rates for those who continue studying. Translation for the x-axis: Dropout in year 1. For the y-axis: Obtained degree for students that continued. Source: [1]

The orange box in figure 1.2 is information added by the Dutch Ministry of Education, the translation is:

3.5% of the students with a Dutch nationality and a Dutch previous education drops out in the first year. Of the students still studying after year 1, more than 60% obtains a degree within 4 years.



**Figure 1.3:** Outcomes of students from the 2011-2017 cohorts four years after starting a bachelor's programme. Source: [1]. The translation of the legend from top to bottom is: Dropout, Still studying, Studying different programme, Graduated, Graduated different programme.

It is important to note that this data is based on the cohorts of 2011-2017. A cohort is the starting year of a students programme. This was done to include everyone who completed their degree within 4 years (3 years + 1 year delay) as this report was published in October 2022. This means that the students who started during the impactful COVID-19 period are not present in this data, but the subset of students who graduated during the lockdowns are. The report does not focus on the effects of COVID-19 on the students.

Figure 1.2 shows that international students who do not drop out obtain a degree more often than Dutch students within four years, this might be explained by the fact that they drop out more often. 34% of Dutch student are still studying after 4 years, which is substantially higher than the 11% of EER-students and 14% of the non-EER students (see figure 1.3). Figure 1.1 shows the quit and switch ratios of the students. The graph shows that Dutch bachelors students switch their degrees 21.7% within the first year. Students likely switch to a different university upon finding that their pro-

gramme of choice is a poor fit. The UU only keeps track if students have left the university and not whether they enrolled in another university, therefore this data is unavailable. The dropout rates of the UU might be lower than the data suggests due to this distinction between dropouts and switchers not being in the data.

To analyse the performance of UU students the international department of marketing created this thesis project.

## 1.2 Greece & Geoscience

An example of a specific student body to analyse is the Greek master students at the geoscience faculty. The potential Greek students, who considered applying to the UU, thought the performance of their peers had declined. Upon hearing this rumour, the UU came with the question whether Greek master students at the geoscience faculty had started to perform worse recently. And if it was possible to find other clusters of students who required more attention from the university in order to increase their chances of graduating.

## 1.3 Literature review

Understanding the factors that influence the academic performance of international students is of vital importance in ensuring their successful adaptation and educational attainment in a foreign educational environment. This literature review will discuss what factors play a role in the chances of academic success for international students. Furthermore, it will explore the potential of exploratory data analysis as a powerful tool to identify patterns, trends, and distinctive student groups within the international student population. International students experience many stressors when it comes to studying abroad. Moving to a new environment is often traumatic for the student and one of the major problems for international students [2]. The host university can provide aid for the international student by properly informing and advising them on travel, housing and taking care of their



physiological needs [3]. These studies have been done in English-speaking countries, but it is likely that their findings extend to non-English speaking countries such as the Netherlands since the host-language proficiency of international students does not directly impact their interactional and general adjustment[4]. Lillyman and Bennett [5] states that stress is one of the major negative contributors towards academic achievement and that the difficulties of writing in a second language can be overcome given mentoring. If this mentoring is present, they find that English language proficiency does not affect the academic achievement of international students. And that the focus should be on providing a non-threatening environment where students have access to different types of support that match with the student. Multiple factors such as the amount of years spent studying, the type of education, age, gender, country of origin can determine what type of aid best suits the student. In an effort to discover which facets of these factors are important and how they work together, exploratory data analysis will be used. Tukey *et al.* [6] discussed the importance of graphical techniques for visualising data as this helps with gaining insight into patterns of the data. Furthermore he emphasised the importance of analysing quality issues in the data. Understanding what the impact of missing values are and why they are missing are key issues to address before the analysis is conducted. Van Atteveltdt, Trilling, and Calderón [7] build upon this and explains how modern computational methods can be used to aid in exploring data. They highlight the importance in showing the relations between variables and how to best create graphs with multiple factors. Using the techniques as described in their book incorporates the important step of quantifying the significance of the finding with p-values. They also advocate the use of Analysis of Variance (ANOVA) for finding groups within the data. In summary, this literature review has examined the factors influencing international student performance, including the impact of stressors and the importance of support mechanisms. Furthermore, it has highlighted the potential of exploratory data analysis techniques, as discussed by Van Atteveltdt, Trilling, and Calderón [7], in uncovering meaningful patterns and relationships within the international student population. By leveraging

these methodologies, researchers can gain valuable insights to enhance educational experiences and improve academic outcomes for international students.

## 2. Data

This chapter aims to provide a comprehensive overview of the dataset used in this research, including the source, characteristics, and any preparation steps performed.

### 2.1 Dataset

The dataset that was used is a database of all student records. This database is maintained by the UU data team and was started during the 2011 schoolyear. The University deems the data before 2017 too unreliable to use, therefore the data from 2017 onward will be the data used in this thesis. Each schoolyear is split into four periods where students can obtain grades. The data encompasses all periods from the start of the 2017 schoolyear until the third period of the 2022 schoolyear. Henceforth all years discussed in this thesis will refer to a schoolyear which ranges from September until approximately the end of July at UU.

### 2.2 Data collection

The database is an aggregation of information gathered by a number of different sources. This thesis limits the scope to the following fields:

- Studentnumber, this field is a pseudonymized ID that can link to a individual student. The UU data team ensured that no personal identifying information such as name and address can be linked to the pseudonyms.
- Study programme, this field holds the name for the programme that the student is following
- Programme type, lists whether the study programme is a bachelor's

or a masters or a pre-masters programme

- Gender, lists the gender of the student
- Country of previous education, the country of the previous education before the student started at the UU
- Average grade, this is the average grade per student per study per period.
- Reason quit, is a list of reasons when a student has quit. The options are: Quit in the first half of the first year, quit in the second half of the first year, quit in year 2+, still active, graduated.
- Nationality, this lists the type of passport the student has. Which is often different from where the student followed their previous education.

The veracity and availability of the datafields depends on multiple factors. One of them is the data source:

- DUO (Dienst Uitvoering Onderwijs) is an educational institution in the Netherlands, responsible for the administration and implementation of various educational policies and programs. The following data fields are collected by them: *Student application type*, *Student gender*, *study programme*, *nationality*. This data is considered to be fully accurate and has no missingness.
- ISA (International Student admissions office) is a department that is part of the UU. This department enters the *country of previous education* from a drop-down menu based on what the students entered in their application form. The drop-down menu is to ensure that all entries are valid and there are no variations. However unfortunately this field is less reliable than hoped. For example 'Rusland' and 'Rusland (oud)' (translation: 'Russia' and 'Russia (old)') are both in active use throughout all years, without there being any distinction between them. Furthermore, a large percentage of this data is missing (see 2.1).
- Osiris, the University managed database is responsible for the fields *average grade*, *reason quit*. The *average grades* field is calculated from the

grades entered by the professors responsible from the courses. The present data is considered to be fully reliable, but there is missing data in 22.5% of the entries. Part of this can be explained by students who stop studying and therefore stop obtaining grades but they have not unregistered yet. The *reason quit* field is an automatically generated field when students leave the UU with their time of leaving.

Region	Type	Avg grade (%)	Country of prev. educ. (%)	Reason quit (%)
EER	Bachelor	25	38	9
	Master	28	28	3
NL	Bachelor	15	99	2
	Master	33	99	4
non-EER	Bachelor	30	46	12
	Master	25	22	2

**Table 2.1:** The percentage of missing data out of all student entries (not all unique students), grouped by student region and programme type.

Students have the right to ask for their data to be deleted, which can result in student IDs without any information linked to them. According to the data team at the UU, this has only happened a handful of times and does not significantly impact the results of this study.

## 2.3 Data preparation

An SQL query was used to extract the desired information from the database. This query selected all the students that had finished their application to the university, and had been admitted. This leaves all of the students that should be studying, which were then filtered to all of the students that enrolled in an education programme, and not just a single course. The output of this query was stored in a CSV file.

As the interest lies on the results per students, only the last entry of each student per programme is used for calculating the comparison figures with the factsheet.

## 2.4 Data description

The dataset has 870714 rows, with 91302 unique students. Of these students there are 44725 students who only did a bachelors, 29256 students who only did a masters, and 15246 students who did both at Utrecht university. Some students only did a Premaster or a Post-initial master's programme. These 2075 students will be left out of the analysis, as requested by the University. The international student population is split into 3960 bachelor students and 8260 master students. Each row in the dataset represents a student's grade for a specific period, year, and study programme. Therefore, a student who pursues two different programmes simultaneously and receives grades for both programmes in all four periods will have eight observations for a single year. There are 4706 students in the dataset who have attempted two studies simultaneously and there was one student that managed to have 61 entries in the six years that have reliable data. Additionally, each observation contains background information on the student, such as: nationality, gender, prior education, et cetera for a total of 49 columns. The fields that were used are mentioned above.

## 2.5 Data privacy and ethics

The protection of data privacy and adherence to ethical guidelines are paramount when working with sensitive datasets. This study uses a dataset that consists of pseudonymized student IDs without any personally identifiable information such as names, addresses, or contact details. The use of pseudonymized IDs ensures the anonymity of individual students and reduces the risk of re-identification. The necessary permissions and approvals to access and analyze the dataset were obtained, ensuring compliance with legal and ethical standards. Additionally, all analyses and results are presented at an aggregate level to further protect the privacy of individual students. No attempt has been made to link the pseudonymized IDs to any external or additional identifying information. The findings of this study are presented in a manner that maintains the privacy of the individuals represented in the dataset.

## 3. Method

This study was a combination of an exploitative and exploratory data analysis and relied heavily on the methods outlined by Van Atteveldt, Trilling, and Calderón [7] and Tukey *et al.* [6]. Python and in particular the pandas package [8] was used for the data analysis, with the matplotlib [9] and seaborn [10] packages for creating the visualisations. The scipy [11] package was used for statistical tests. The statistical tests in this analysis are used as outlined by Fisher [12]:

**P-value** The P-value is used to reject the assumption that there is no significant relationship or difference between variables. It represents the probability of observing the data (or more extreme results) with that assumption. If the p-value is low (typically below 0.05) then there is evidence to reject the assumption of the variables having no relationship, suggesting that the observed effect or relationship is statistically significant.

**ANOVA** ANOVA is a statistical method used to compare the means of two or more groups to determine if there are any statistically significant differences among them. It calculates an F-statistic, which represents the ratio of the between-group variability to the within-group variability.

**Z-score** Z-scores are a measure of how many standard deviations a particular data point is from the mean of a distribution. This helps identify outliers and extreme values in your data by quantifying how far a value deviates from the average.

### 3.1 Factsheet comparison

To compare the performance of UU students with the national averages cohorts 2017 and 2018 are used since these are the years after 2016 that have

had 4 years since. This is not a direct comparison with the factsheet as that uses the cohorts from 2011 till 2017, but in this instance it is not a trend comparison and therefore it is suitable. There are two additional caveats: Firstly, in the year 2017 no country of previous education data was collected. Secondly, the data only continues until the third period of 2022. Meaning that for cohort 2018 there are only  $3\frac{3}{4}$ 's years instead of the usual 4 years. Despite these differences, the comparison with the factsheet will be made using the 2017 and 2018 cohorts, with the knowledge that the reported graduation rate will be slightly below the actual graduation rate.

### **3.2 Student group identification**

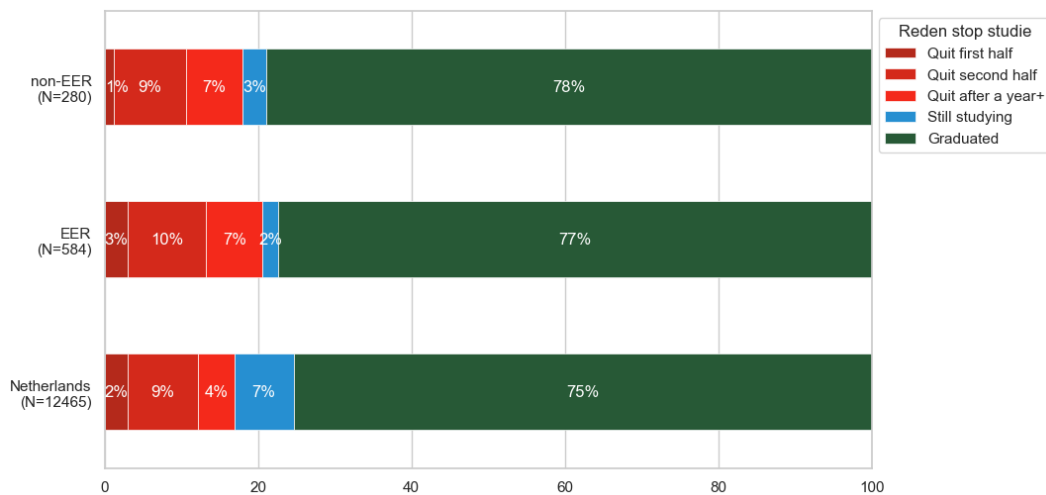
For identifying groups of students that perform less well, an exploratory data analysis approach was used. At first clustering was attempted, however due to the homogeneity in the data and the large amounts of missing data this quickly proved to be out of scope for the time constraints that were given. Instead certain population groups within the dataset were examined for deviating behaviour. Multiple combinations of student information columns and study outcome columns were used. Many of the fields present in the dataset were not complete or not required for the comparison with the factsheet but contained information nonetheless. For the combination of student origin and student outcome the UU defined an interesting group as: No less than 5 students with an average performance worse than the performance of the Dutch students. With the underlying assumption that the Dutch performance should be attainable for all international students, thus if that is not achieved, then this is a signal for the UU that specific help for this group might be needed.



## 4. Results

This chapter presents the results that were obtained by researching the hypotheses of this thesis. The chapter is split in two parts, one for each research question, and describes the key findings.

### 4.1 Factsheet comparison

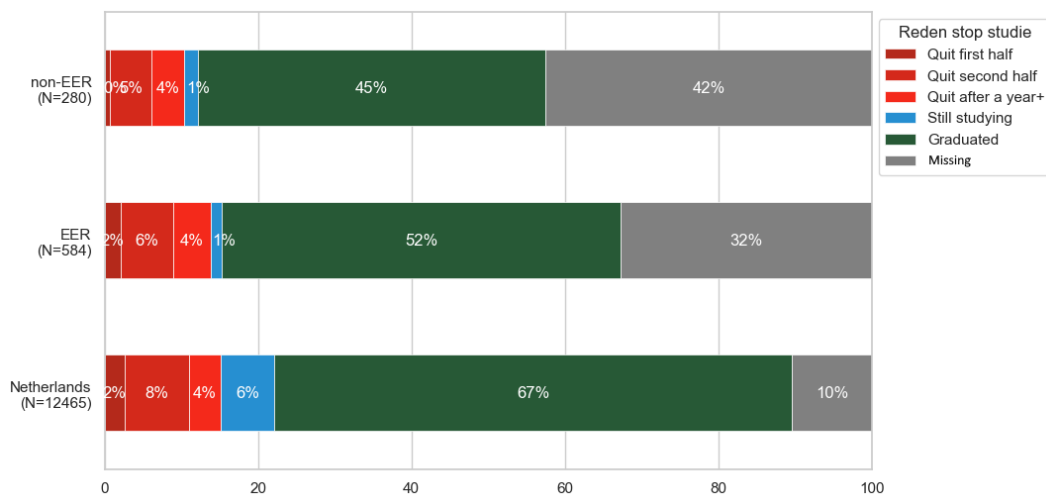


**Figure 4.1:** The outcomes of students who start a pursuit of a bachelor's degree at the UU after 4 years.

UU student groups significantly outperform the national averages according to figure 4.1. The percentages of students who quit in the UU figures are equal to the quit + different degree + graduated different degree percentages of the factsheet due to the fact that only switchers in the factsheet are counted as dropouts in this study.

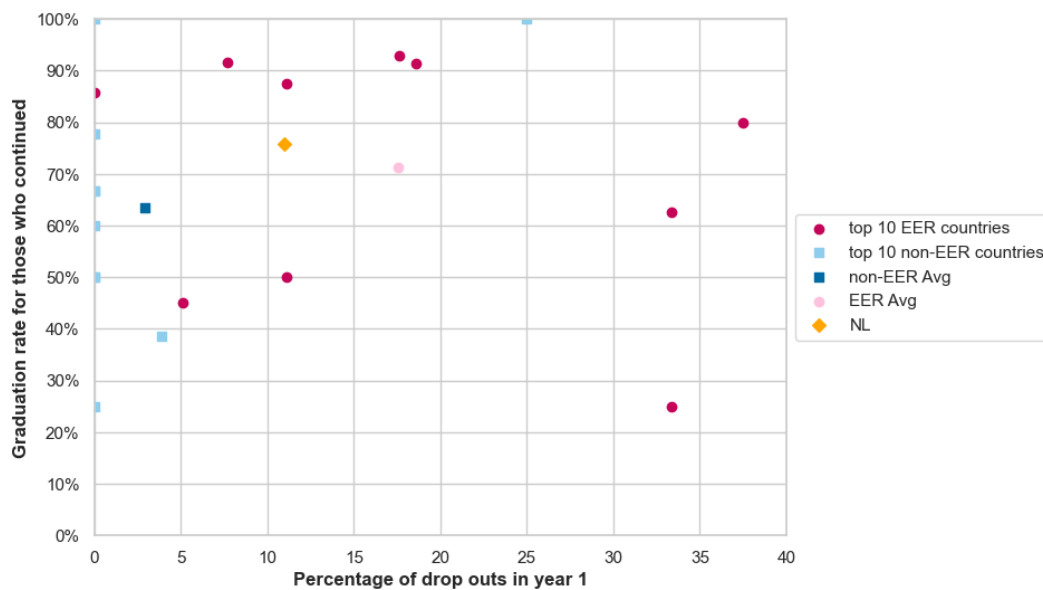
However, upon incorporating the missing data in the *reason quit* column (see fig. 4.2) it becomes unclear how well each student group performed. In an attempt to use this data despite the missingness, one could easily impute if the missing data amongst each field relies on the Missing Completely At Random assumption (MCAR). However the p-value for the relationship be-

tween the missing *reason quit* data and the *student region* data is  $1.47 \times 10^{-82}$ , indicating that the MCAR assumption does not hold[11]. This means that there is a link between the students region and the likelihood of them having a missing outcome of their studies. This missingness is an interesting find as the UU does not know at the time of writing why any data in the reason quit field would be missing. Beyond the missing data it can be seen that the graduation rate of the Dutch students is higher at the UU with 67% than at the national average of 55% despite the unknown outcome for 10% of the UU students.



**Figure 4.2:** The outcomes of students who start a pursuit of a bachelor’s degree at the UU after 4 years, with missingness shown.

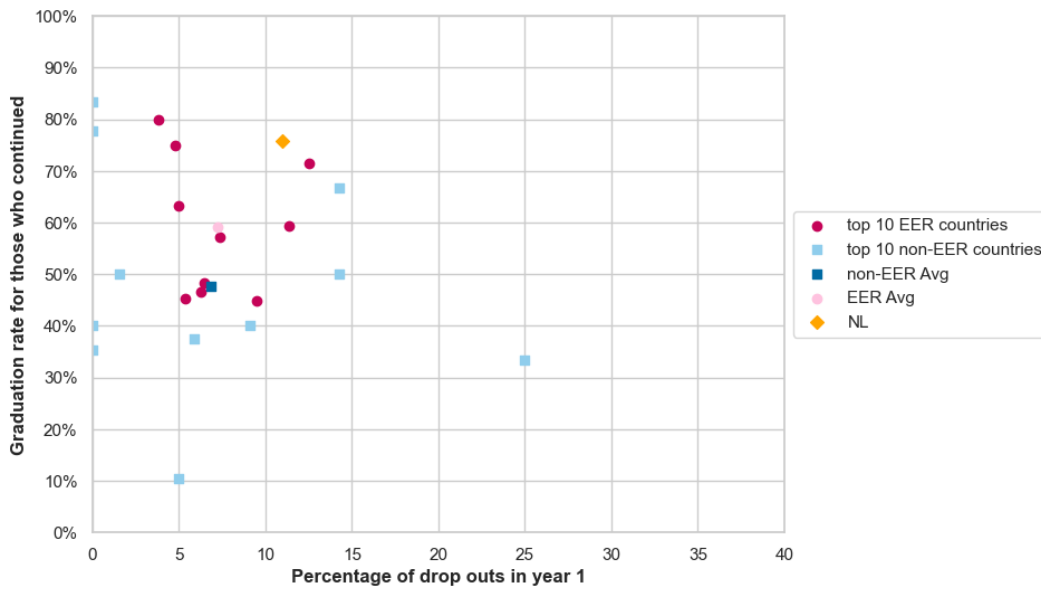
Figures 4.3 and 4.4 show two different versions of the comparison with figure 1.2 (the dropout/graduation graph of the factsheet). Figure 4.3 shows the data plotted using the *country of previous education field* as requested by the UU. However due to the large percentage of missing data, the unknown country represents 67% and 73% for EER and non-EER students respectively. This can also explain the much more spread out results than were found in the national figures. The UU has a smaller sample-size, but the missing data lowers the sample size per country even more. Despite that, there still is a visible trend of the non-EER countries dropping out less often in year 1 than Dutch or EER countries. The graduation rates of students who continue after the first year is higher at the UU for Dutch students. The lower average performances of EER and non-EER students at the UU may



**Figure 4.3:** Percentage of students that drop out in the first year and the graduation rate within four years for those who continue studying. Split per country

partly be explained by the missing *reason quit* data. For the non-EER students however it seems that another factor is important. No students out of the 10 largest non-EER countries quit in the first half of the first year and most of those who quit did so after the first year. When including all non-EER students this deviating behaviour is still visible, suggesting that there is a reason non-EER students quit later during their studies than EER and Dutch students (see table 4.1).

Figure 4.4 shows the distributions based on nationality instead of the country of previous education. Each point in the image now represents a significantly larger student body. Despite the larger sample sizes the figure is still more spread out than the figure from the factsheet. Interestingly the EER nationalities are all huddled together as a cluster and the non-EER nationalities surround the EER cluster. The performance of individual nationalities is heavily impacted by the missing *reason quit* data. The datapoint on the (5%,10%) location represents Indian students, whose *reason quit* data is missing for 70% of the students. Despite the missingness there is still an interesting find in this data. The Dutch ministry of education found that there are significant differences in performance based on country, but in the UU



**Figure 4.4:** Percentage of students that drop out in the first year and the graduation rate within four years for those who continue studying. Split per nationality

Reason quit	non-EER (%)	EER (%)	NL (%)
Quit first half year 1	6.9	14.8	17.4
Quit second half year 1	51.7	49.4	54.9
Quit in year 2+	41.4	35.8	27.7

**Table 4.1:** The distribution of when students drop out.

dataset the performance per nationality seems to be mostly the same. The differences in graduation rate can mainly be explained by the differences in missingness. These two factors have a Pearson correlation coefficient [13] of  $-0.88$  with a p-value of  $5.5 \times 10^{-7}$ . Meaning that the higher the missingness the lower the graduation rate. The correlation increases to  $-0.93$  when filtering out some of the smallest countries. This correlation is across a combination of EER and non-EER countries. The implication of this finding is that the graduation rates for students who continue after year 1 of international UU bachelor students is roughly the same regardless of their nationality. This is only true under the assumption that the missingness is unrelated to student performance.

## 4.2 Student group identification

The following section contains the most interesting results that were obtained using exploratory data analysis in an attempt to characterise student groups with sub-par academic achievements.

### 4.2.1 Study programme and student background combinations

There are many ways to identify underperforming student groups. The UU's preferred manner is finding groups of at least 5 students who perform worse at a certain programme than the Dutch students. Unfortunately it appears that international bachelor students have a higher percentage missing *reason quit* data, lowering their observed graduation rates. Therefore students were also compared to the average performance of their group. When using the *country of previous education* data as a group identifier it is possible to find groups that perform worse than Dutch students or worse than the average performance of their group. However, most of these groups would belong to an unknown country, defeating the purpose of identifying groups of students. For this reason the *region* and *nationality* data will be used as group identifiers.

The 'UCU Liberal Arts and Sciences' programme had 100% of their outcome data missing which equates to a 0% observed graduation rate. This is the only programme that had 100% missing data, with 506 students. For this reason it came up with any possible combination of student background when looking for underperforming outliers and significantly affected the average graduation rate of certain nationalities that had large amounts of students enrolled in that programme. For these reasons that programme was filtered out of the following results and the graduation rates were recalculated after excluding 'UCU Liberal Arts and Sciences'. This is possible since the aim here is not to establish the absolute graduation rates, but to find differences in relative graduation rates.

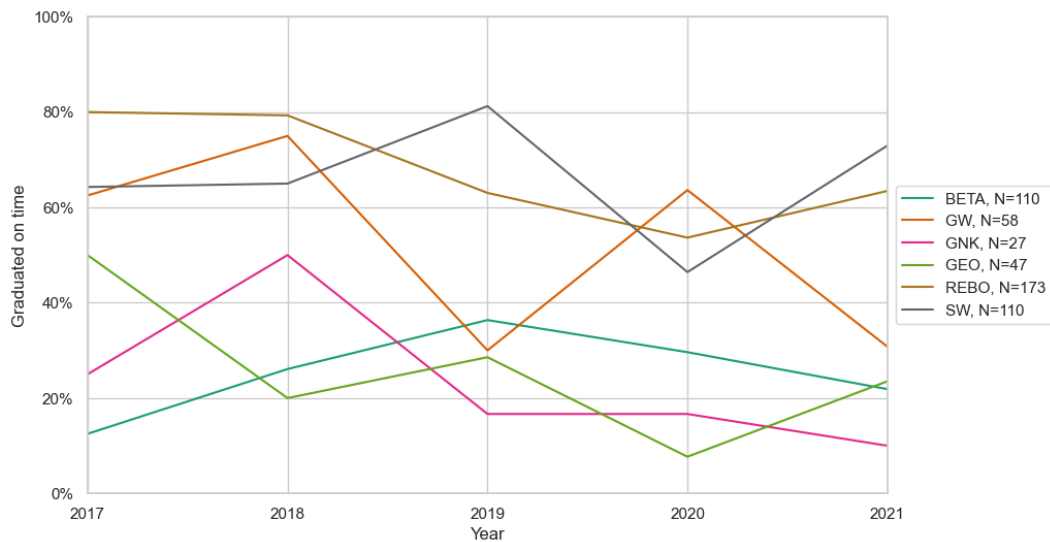
## Results

Study Programme	Region	N	Graduation rate (%)		z-score
			observed	expected	
History	EER	26	61.5	71.2	-2.944
History	non-EER	13	69.2	71.3	-0.456
Economics and Business...	EER	85	64.7	71.2	-3.576
Economics and Business...	non-EER	50	64.0	71.3	-3.106
Media en cultuur	EER	29	65.5	71.2	-1.828
Game and Media Technol...	EER	19	68.4	84.6	-4.208
Game and Media Technol...	non-EER	12	66.7	84.9	-3.784
Earth, Life and Climate	non-EER	14	50.0	84.9	<b>-7.815</b>
Global Criminology	EER	27	59.3	84.6	<b>-7.862</b>
Gender Studies (Resear...	EER	12	75.0	84.6	-1.981
Gender Studies (Resear...	non-EER	12	50.0	84.9	<b>-7.235</b>
Marine Sciences	EER	14	78.6	84.6	-1.341
Marine Sciences	non-EER	13	76.9	84.9	-1.727
Contemporary Theatre, ...	EER	11	72.7	84.6	-2.348
Gender Studies	EER	24	79.2	84.6	-1.582
Earth Structure and Dy...	EER	13	76.9	84.6	-1.648
Business Informatics	EER	13	76.9	84.6	-1.648
Business Informatics	non-EER	17	76.5	84.9	-2.087
New Media and Digital ...	non-EER	13	61.5	84.9	<b>-5.044</b>
Strategy, Competition ...	EER	13	0	84.6	<b>-18.229</b>
Bio Inspired Innovation	EER	16	75.0	84.6	-2.288
M&S for the Behavioura...	EER	15	73.333	84.567	-2.601
Linguistics	EER	13	76.9	84.6	-1.648

**Table 4.2:** The performance of student regions in programmes where they underperform with the 5 most deviating groups boldened. The programmes below the line are masters programmes. Only the programmes with at least 10 students were shown here, see appendix 6.2 for the full list.

In table 4.2 there are numerous examples of student subgroups that have a poor relation to their programme of choice. The appendix has more combinations of student groups and programme/faculty combinations (see appendix 6.2).

### 4.2.2 Greece

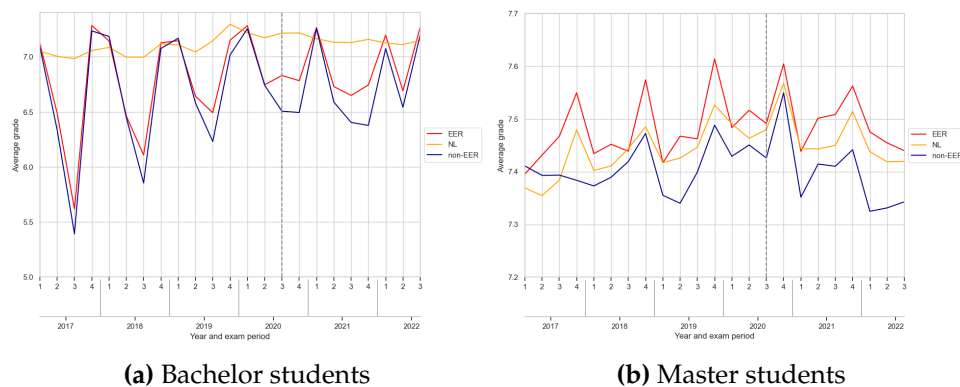


**Figure 4.5:** The percentage of Greek students who graduated without delay per schoolyear, split per faculty.

For the Greek master students the interest lay in the trend of the student performance due to the perceived change of academic success. The *reason quit* field did not show any significant changes due to the missing data. However, the *study duration* field did. This field states how much study delay a student experienced if they graduate. It relies on the *reason quit* data and therefore has the same missingness. Figure 4.5 shows the percentage of student that completed their degree in the standard amount of time. The inverse percentage being the students who fail to complete their degree or those who are delayed. Due to the smaller sample size there is a lot of variance visible in the graph, but the reported downwards trendline for the geoscience faculty (GEO in the graph) seems to be present. Especially the period of 2017-2020 showed a clear decline in performance with only 1 out of the 13 students graduating on time in 2020. The x-axis in this figure represents the year of graduation and not the cohort year. This allows for 2021 to be plotted as well since the study duration is included in calculating the percentage of students who graduate in the standard amount of time.

### 4.2.3 Other findings

For more recent data, graduation rates are not available yet. But other performance metrics in the dataset such as average grades and first year dropouts can still offer insights. When looking at the average grades of bachelor students in figure 4.6a a wave-like pattern emerges. Each year starts off with high grades, then in the second and third period the grades lower and in the final period students achieve high grades again. This pattern holds for all regions until the start of the COVID-19 lockdown in the Netherlands, halfway through the third period of 2020. The lockdown affected the grades obtained during the third period and those after that. The difference between international and Dutch students is very pronounced. During the first and the last period of each year the international students perform similarly to the Dutch students. But the slight dip the Dutch students experience during the year is magnified by the international students. During these two period both EER and non-EER students struggle with their studies. Before the pandemic there is a visible trend of a less pronounced dip each year for the international students. Since the pandemic however, the trend of the decrease of the amplitude is no longer present.



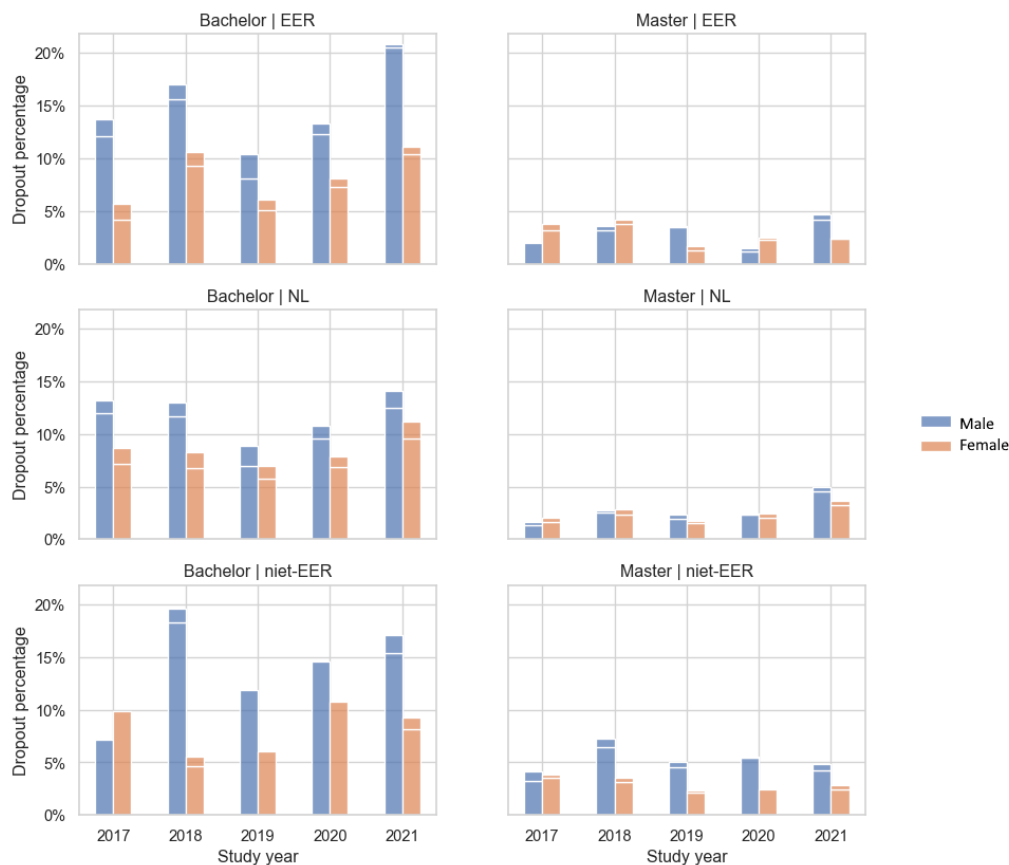
**Figure 4.6:** The averages grades obtained by students per period per year, split by region. The vertical dashed line indicates the first grades that were obtained during the COVID-19 related lockdown in the Netherlands

For master students the average grades graph shows a completely different trend (see figure 4.6b). Here EER students perform slightly better than the Dutch students and the non-EER students perform slightly worse. All



three student groups have roughly the same pattern on a much smaller scale than the bachelor students. The y-axis on the master student grades figure is much narrower. Grades post-COVID are trending downwards slightly, but there is no clear 'problem period' that was visible in figure 4.6a. Interestingly there is no peak visible on the first period of each year.

The other performance statistic that was looked into is first year dropout rates. Figure 4.7 shows that bachelor students quit much more often than master students. But also that male, and especially male international students have a higher dropout rate. The difference between the groups was measured with ANOVA and is statistically significant with an F-value of 4.20 and a p-value of 0.045.



**Figure 4.7:** The dropout rates in the first year, per year, grouped by the type of programme, the student region and the gender. The small strip in the bar indicates what percentage of the dropouts quit at what time of the year. The small part of the bar, if present, represents the students that quit in the first half of the year.

## 5. Conclusion

The hypothesis for the first research question was: Native and international students at the UU perform better than the Dutch national average. The student body is divided into three categories and therefore their performance is too. This thesis found that the performance of Dutch bachelors students at the UU exceeds the average performance of Dutch bachelors students in the Netherlands. The graduation rate at the UU is upwards of 67% whereas the national average is 55%. Unfortunately the current dataset does not allow for precise measurements as 10% of the student outcomes is missing. Besides the missing data, the 2018 cohort did not have the full 4 years which negatively impacts the graduation rate too. For the EER and non-EER students an accurate comparison with the factsheet is not possible due to the amount of missing data and therefore the hypothesis can not be accepted or rejected until further research has been done.

The hypothesis for the second research question was: Certain backgrounds and study programme combinations will have a performance below expectations. Thus we expect that these groups will benefit from student aid. The combination of student groups and their study choice reveals that certain combinations do not perform as expected. These findings confirm the initial hypothesis. This mismatch could be due to cultural factors or the education in the home country of the student not aligning well with the required barriers of entry at the UU. As shown in the example with Greek geoscience students it is possible to find trends in subgroup performance. This might allow for the UU to find and intervene on certain groups sooner. Based on student characteristics alone it is possible to identify groups of students who will be more likely to struggle than other groups. This study found that the combination of gender, student region and programme type identifies a struggling subgroup with male international bachelor students underperforming.

## 5.1 Discussion & further research

In the conclusion it is stated that an accurate comparison of the international student performance with the factsheet is not possible due to the amount of missing data. However that does not mean that the UU cannot gain insight from this. Not only is the current method of entry for this datafield not working as hoped, but its errors are related to the origin of the student. This fact is extremely interesting. There appears to be a systematic error in acquiring student outcome data that affects international students more often. It is in the universities best interest to quickly remedy this, not only to increase the quality of the data they have access to, but also to identify if there are other processes that negatively affect international students, since somewhere during the processing their data is treated differently than the Dutch students. Furthermore, the investigation of the missing data in the *reason quit* column might give an answer to why there seems to be remarkably little difference in student performance based on nationality. Figure 4.4 shows that there is more variance in the dropout rates in the first year for non-EER students than for EER-student. As many EER students have more cultural similarity with the Netherlands this can indicate that culture and nationality affect the chances of a students success. However the graduation rates for students that continue are approximately the same across EER and non-EER, suggesting that this is not the case. Cultural and/or national differences are expected to play a role in the likelihood of a students success. The International Student Office at the UU has stated that they have empirically experienced different levels of academic success based on the nationality of the students. Therefore it appears more likely that the missingness is related to student performance, which immediately begs multiple questions: Why is it only related for international students? What are the actual dropout values? However this remains speculative until further research has been done.

If the UU is able to retrieve or otherwise fix the missing data issue, a new research in the comparison between the international students at the UU and on a national level would be interesting as this could lead to the next

question: Why do international students perform differently at the UU? Or, in the case that they perform similarly, the question could be asked: Why don't international students at the UU outperform the national average when the Dutch students do? However if the UU is not able to retrieve the missing data some way, then an investigation into why the *reason quit* data for international students has a different behaviour than for Dutch students could also yield results that may help students.

The average grades graphs (figures 4.6a and 4.6b) show that non-EER students obtain lower grades than the Dutch students for both bachelor and master students. This suggests that non-EER students perform worse at the UU than Dutch students. This can not be stated with certainty using the current data, but the lower grades indicate towards this conclusion. Further research can confirm or deny this assumption.

For the Greek students it was confirmed that their success in the geoscience faculty has declined over the years. For the UU it would be interesting to research what other student bodies experience the same trend. Finding out if this decline is solely for Greek students or if other students in that faculty are experiencing the same, may help the UU with properly allocating their resources. Currently a student-centric approach is used to identify outliers in performance, but a programme-centric approach is another avenue of research that is recommended to pursue.

Table 4.1 has shown that non-EER students who quit, do so later on in their studies than other student groups. An investigation why non-EER students quit later provides insight into if, for example, they require more help in the later parts of their study, or if they should be advised not to hold on too long if they do not match with the programme.

The *country of previous education* datafield is currently relying on the input of international students themselves, which is a likely explanation for the large percentage of missing data. Changing this process can improve the quality of future data as the UU is more interested in the country of a student's previous education than in their nationality since previous education should be a stronger indicator towards student success.

The origin of the student is not the only factor that can help in predicting the chances of a student's success. The UU could benefit from including more personal data in finding 'at-risk' groups. Student age, location of residence and multiple combinations of other factors were found to be indicators of performance in other universities [14]. The combination of a more robust dataset and more personal information could allow for clustering approaches towards finding student groups. Clustering techniques can be more precise and save the UU time, which is why this study recommends the UU to do further research in this area.

As seen in figures 4.6a and 4.6b, bachelor and master students have different performance trends per year. Perhaps the different types of programmes start the year with differing difficulties of workloads, or there master and bachelor students have varying motivation levels at the start of the year. Investigating the cause of this difference can bring more insight into why and when student aid would be beneficial.

One potential factor that could influence the academic achievements of international students is an advice of the UU regarding housing. On the 9th of June 2022 it was reported that the UU advised international students who had not yet found housing, not to come [15]. As moving is one of the major stressors for international students this report might have increased the stress for these students. It would be interesting to explore if this led to a quantifiable impact on international student performance. This would have to be researched using a more up to date dataset, but it could provide insight into the direct effects that university communication have on the perceived stress of (international) students.

In summary this study recommends conducting further research, including but not limited to, the following topics:

- Why is there a correlation between students origin and their missing data?
- Why do non-EER students quit later during their study, and is this behaviour desirable?
- What groups can be found using modern clustering approaches if more personal data is included?
- Why do bachelor and master students have different grade trends per year?
- What insights can be found using a programme-centric approach into the data?
- How can the UU directly influence the levels of stress for international students with their communication?

# Bibliography

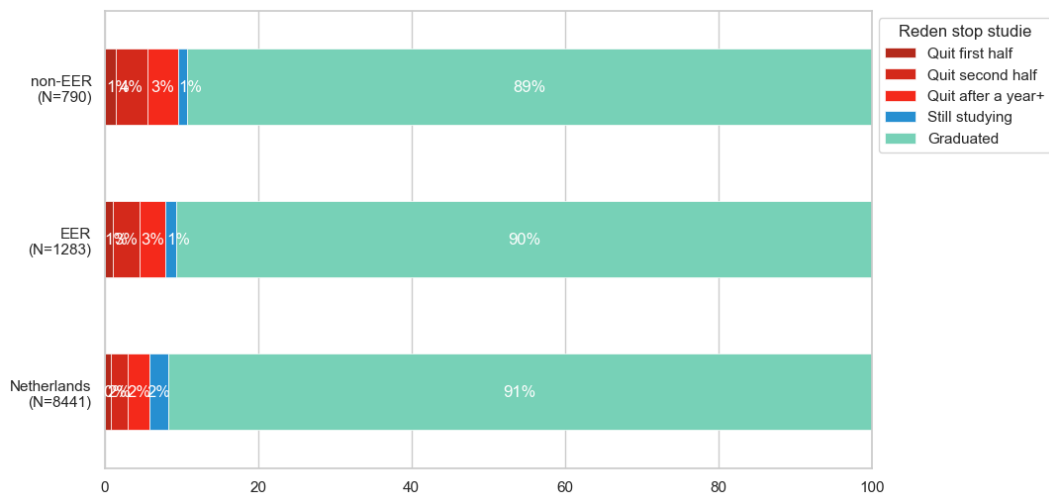
- [1] I. of education, "Factsheet oktober: Study success of international students in a university bachelor's programme," Oct. 2022, Royal Dutch Ministry of Education, Culture and Science. [Online]. Available: <https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/publicaties/2022/10/13/factsheet-studiesucces-van-internationale-studenten-in-de-wo-bachelor/Factsheet-Studiesucces+van+internationale+studenten+in+de+wo-bachelor.pdf>.
- [2] L. Brown and I. Holloway, "The adjustment journey of international postgraduate students at an english university: An ethnographic study," *Journal of research in International Education*, vol. 7, no. 2, pp. 232–249, 2008.
- [3] Y. Zhou and J. Todman, "Patterns of adaptation of chinese postgraduate students in the united kingdom," *Journal of Studies in International Education*, vol. 13, no. 4, pp. 467–486, 2009.
- [4] M. Wilczewski, R. Wang, J. Du, *et al.*, "Cultural novelty and international students' experience: A five-country study," *Higher Education*, pp. 1–22, 2022.
- [5] S. Lillyman and C. Bennett, "Providing a positive learning experience for international students studying at uk universities: A literature review," *Journal of Research in international Education*, vol. 13, no. 1, pp. 63–75, 2014.
- [6] J. W. Tukey *et al.*, *Exploratory data analysis*. Reading, MA, 1977, vol. 2.
- [7] W. Van Atteveldt, D. Trilling, and C. A. Calderón, *Computational Analysis of Communication*. Wiley Blackwell, 2022.
- [8] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: 10.5281/zenodo.3509134. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>.
- [9] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.
- [10] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. [Online]. Available: <https://doi.org/10.21105/joss.03021>.
- [11] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

- [12] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in statistics: Methodology and distribution*, Springer, 1970, pp. 66–70.
- [13] I. Cohen, Y. Huang, J. Chen, *et al.*, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.
- [14] J. Morrison, B. Merrick, S. Higgs, and J. Le Métails, "Researching the performance of international students in the uk," *Studies in Higher Education*, vol. 30, no. 3, pp. 327–337, 2005.
- [15] "Universiteit utrecht adviseert internationale student: 'geen kamer? kom niet'," 2022. [Online]. Available: <https://www.rtvutrecht.nl/nieuws/3421778/universiteit-utrecht-adviseert-internationale-student-geen-kamer-kom-niet>.

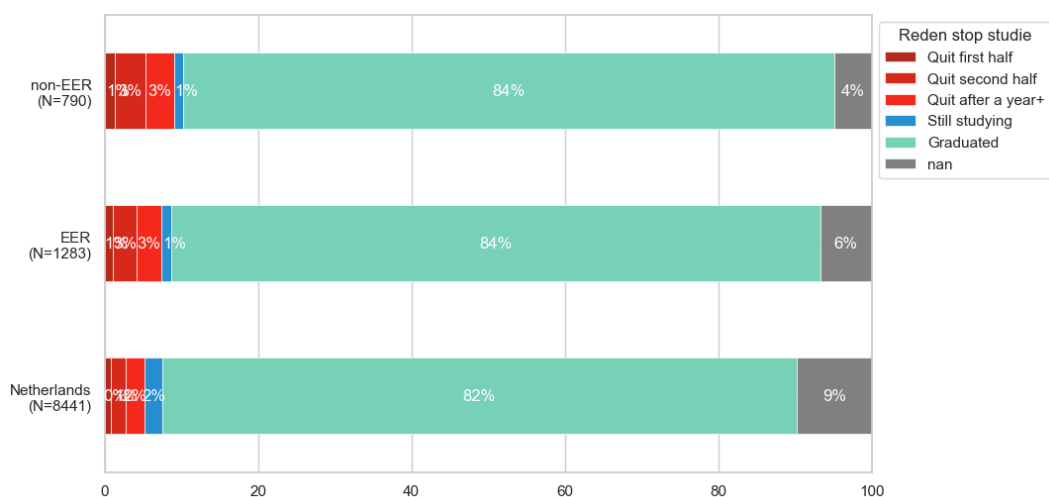


## 6. Appendix

### 6.1 Factsheet



**Figure 6.1:** The outcomes of students who start a pursuit of a masters degree at the UU after 4 years.



**Figure 6.2:** The outcomes of students who start a pursuit of a masters degree at the UU after 4 years, with missingness shown.

## 6.2 Student groups and study combinations

Study Programme	Region	N	Graduation rate (%)		z-score
			observed	expected	
Economics and Business...	Britse	7.0	71.4	82.1	-3.676
Economics and Business...	Bulgaarse	7.0	71.4	85.7	-4.913
Economics and Business...	Duitse	24.0	70.8	70.9	-0.064
Economics and Business...	Egyptische	6.0	50.0	57.1	-2.259
Economics and Business...	Finse	10.0	70.0	83.3	-5.462
Media en cultuur	Poolse	6.0	66.7	83.3	-5.28
UCR Liberal Arts and S...	Italiaanse	14.0	64.3	73.5	-4.47
Game and Media Technol...	Griekse	6.0	50.0	81.6	-3.749
Artificial Intelligence	Italiaanse	6.0	66.7	83.9	-2.041
Earth, Life and Climate	Britse	6.0	16.7	79.2	-7.415
Global Criminology	Italiaanse	7.0	71.4	83.9	-1.602
Marine Sciences	Britse	6.0	66.7	79.2	-1.483
Molecular and Cellular...	Italiaanse	6.0	66.7	83.9	-2.041
Public International Law	Duitse	7.0	85.7	89.4	-0.474
Public International Law	Spaanse	9.0	88.9	91.0	-0.305
Clinical Psychology	Turkse	12.0	91.7	92.1	-0.067
International Relation...	Duitse	6.0	83.3	89.4	-0.724
Social, Health and Org...	Duitse	11.0	81.8	89.4	-1.221
M&S for the Behavioura...	Duitse	8.0	75.0	89.4	-1.973

**Table 6.1:** The table of underperforming combinations of student nationality and programme choice.

Study Programme	Region	N	Graduation rate (%)		z-score
			observed	expected	
Faculteit Geestesweten...	Belgische	8.0	25.0	63	-6.746
	Litouwse	6.0	50.0	63	-1.999
	Spaanse	7.0	57.1	63	-0.98
Faculteit Recht, Econo...	Duitse	25.0	68.0	68	0.0
	Egyptische	6.0	50.0	68	-2.767
Faculteit Geowetenscha...	Duitse	15.0	73.3	74	-0.17
University College Roo...	Amerikaanse	9.0	77.8	82	-0.791
	Duitse	29.0	72.4	82	-3.245
	Italiaanse	14.0	64.3	82	-4.157
	Spaanse	7.0	71.4	82	-1.76

**Table 6.2:** The table of underperforming combinations of student nationality and faculty for bachelor students.

## 6.2 Student groups and study combinations

Study Programme	Region	N	Graduation rate (%)		z-score
			observed	expected	
History	niet-EER	13	69.231	71.348	-0.352
Philosophy, Politics a...	niet-EER	6	33.333	71.348	-4.288
Economics and Business...	EER	85	64.706	71.194	-2.755
Economics and Business...	niet-EER	50	64.0	71.348	-2.393
Media en cultuur	EER	29	65.517	71.194	-1.408
Geschiedenis	niet-EER	7	57.143	71.348	-1.731
Game and Media Technol...	EER	19	68.421	84.567	-3.241
Game and Media Technol...	niet-EER	12	66.667	84.937	-2.914
Earth, Life and Climate	niet-EER	14	50.0	84.937	-6.02
Global Criminology	EER	27	59.259	84.567	-6.056
Gender Studies (Resear...	EER	12	75.0	84.567	-1.526
Gender Studies (Resear...	niet-EER	12	50.0	84.937	-5.573
Marine Sciences	EER	14	78.571	84.567	-1.033
Marine Sciences	niet-EER	13	76.923	84.937	-1.331
Experimental Physics	EER	6	66.667	84.567	-2.019
Epidemiology	EER	8	62.5	84.567	-2.874
Diergeneeskunde	EER	7	0.0	84.567	-10.303
Contemporary Theatre, ...	EER	11	72.727	84.567	-1.808
Youth, Education and S...	EER	9	77.778	84.567	-0.938
Gender Studies	EER	24	79.167	84.567	-1.218
Theoretical Physics	niet-EER	7	42.857	84.937	-5.127
Earth Structure and Dy...	EER	13	76.923	84.567	-1.269
Earth Structure and Dy...	niet-EER	9	77.778	84.937	-0.989
Business Informatics	EER	13	76.923	84.567	-1.269
Business Informatics	niet-EER	17	76.471	84.937	-1.607
Environmental Biology	niet-EER	8	62.5	84.937	-2.922
Nanomaterials Science	EER	7	71.429	84.567	-1.601
Biofabrication	EER	8	62.5	84.567	-2.874
Computing Science	EER	10	80.0	84.567	-0.665
Social Policy and Publ...	EER	10	60.0	84.567	-3.578
Health and Environment	EER	10	0.0	84.567	-12.315
Health and Environment	niet-EER	7	0.0	84.937	-10.349
Earth Surface and Water	niet-EER	10	80.0	84.937	-0.719
New Media and Digital ...	niet-EER	13	61.538	84.937	-3.885
Strategy, Competition ...	EER	13	0.0	84.567	-14.041
Human Geography	EER	9	77.778	84.567	-0.938
Human Geography	niet-EER	8	75.0	84.937	-1.294
Bio Inspired Innovation	EER	16	75.0	84.567	-1.762
M&S for the Behavioura...	EER	15	73.333	84.567	-2.004
Sociology: Contemporar...	niet-EER	6	66.667	84.937	-2.061
Neuroscience and Cogni...	niet-EER	8	75.0	84.937	-1.294
Innovation Sciences	EER	8	75.0	84.567	-1.246
Media, Art and Perform...	niet-EER	8	62.5	84.937	-2.922
Migration, Ethnic Rela...	EER	10	80.0	84.567	-0.665
Linguistics	EER	13	76.923	84.567	-1.269
Cultural Anthropology:...	EER	10	70.0	84.567	-2.121

**Table 6.3:** The full table of underperforming combinations of student region and programme choice.

Study Programme	Region	N	Graduation rate (%)		z-score
			observed	expected	
Faculteit Recht, Econo...	Belgische	8.0	75.0	87.6	-3.861
	Britse	26.0	84.6	87.6	-1.657
	Franse	17.0	76.5	87.6	-4.958
	Ierse	13.0	84.6	87.6	-1.172
	Indiase	6.0	83.3	87.6	-1.141
	Indonesische	6.0	83.3	87.6	-1.141
	Italiaanse	61.0	86.9	87.6	-0.592
	Oekraïense	6.0	83.3	87.6	-1.141
	Oostenrijkse	10.0	80.0	87.6	-2.604
	Poolse	7.0	85.7	87.6	-0.545
	Roemeense	10.0	80.0	87.6	-2.604
	Russische	10.0	80.0	87.6	-2.604
	Slowaakse	6.0	83.3	87.6	-1.141
	Zweedse	7.0	71.4	87.6	-4.643
Faculteit Betawetensch...	Amerikaanse	11.0	63.6	83.0	-6.97
	Britse	14.0	71.4	83.0	-4.702
	Bulgaarse	7.0	57.1	83.0	-7.423
	Franse	11.0	72.7	83.0	-3.701
	Griekse	46.0	76.1	83.0	-5.07
	Ierse	10.0	80.0	83.0	-1.028
	Italiaanse	30.0	76.7	83.0	-3.738
	Roemeense	8.0	62.5	83.0	-6.281
Faculteit Geowetenscha...	Amerikaanse	23.0	82.6	88.1	-2.857
	Britse	49.0	81.6	88.1	-4.929
	Griekse	20.0	85.0	88.1	-1.502
	Ierse	10.0	70.0	88.1	-6.201
	Indiase	6.0	66.7	88.1	-5.679
	Portugese	6.0	83.3	88.1	-1.274
Faculteit Geneeskunde	Amerikaanse	9.0	77.8	85.1	-2.372
	Britse	17.0	64.7	85.1	-9.112
	Chinese	6.0	50.0	85.1	-9.314
	Griekse	15.0	80.0	85.1	-2.14
	Italiaanse	12.0	58.3	85.1	-10.057
Faculteit Geestesweten...	Amerikaanse	17.0	76.5	83.4	-3.082
	Britse	71.0	80.3	83.4	-2.83
	Chinese	19.0	63.2	83.4	-9.539
	Franse	11.0	81.8	83.4	-0.575
	Griekse	31.0	80.6	83.4	-1.689
Faculteit Sociale Wete...	Britse	23.0	78.3	91.4	-6.806
	Duitse	65.0	86.2	91.4	-4.542
	Griekse	34.0	82.4	91.4	-5.685
	Hongaarse	6.0	66.7	91.4	-6.554
	Italiaanse	15.0	86.7	91.4	-1.972

**Table 6.4:** The table of underperforming combinations of student nationality and faculty for master students. Combinations with a z-score > -0.5 were omitted

## 6.2 Student groups and study combinations

Study Programme	Region	N	Graduation rate (%)		z-score
			observed	expected	
History	nan	34	64.7	69.1	-1.218
Economics and Business...	Bondsrepubliek	13	53.8	78.0	-4.141
Economics and Business...	Duitsland	13	53.8	78.0	-4.141
Economics and Business...	nan	478	63.6	69.1	-5.707
Rechtsgeleerdheid	nan	1294	67.9	69.1	-2.049
Wiskunde	nan	227	50.2	69.1	-13.514
Kunstmatige Intelligen...	nan	414	55.6	69.1	-13.036
Spaanse taal en cultuur	nan	49	57.1	69.1	-3.986
Religiewetenschappen	nan	17	52.9	69.1	-3.17
Keltische talen en cul...	nan	14	64.3	69.1	-0.852
Kunstgeschiedenis	nan	76	64.5	69.1	-1.903
Liberal Arts and Scien...	nan	467	53.5	69.1	-15.999
Filosofie	nan	127	56.7	69.1	-6.632
Aardwetenschappen	nan	254	68.5	69.1	-0.454
Italiaanse taal en cul...	nan	13	53.8	69.1	-2.618
Natuur- en Sterrenkunde	nan	250	54.8	69.1	-10.73
Biologie	nan	630	57.8	69.1	-13.46
Muziekwetenschap	nan	50	56.0	69.1	-4.396
Literatuurwetenschap	nan	53	67.9	69.1	-0.415
English Language and C...	nan	202	56.4	69.1	-8.566
Taal- en cultuurstudies	nan	168	64.3	69.1	-2.953
Wiskunde & Toepassingen	nan	51	31.4	69.1	-12.777
Informatica	nan	358	54.7	69.1	-12.93
Scheikunde	nan	275	65.8	69.1	-2.597
Duitse taal en cultuur	nan	23	65.2	69.1	-0.888
Taalwetenschap	nan	50	66.0	69.1	-1.04
Franse taal en cultuur	nan	48	68.8	69.1	-0.099
Academische lerarenopl...	nan	123	1.6	69.1	-35.528
Game and Media Technol...	nan	58	62.1	69.1	-2.53
Global Criminology	Groot-Brittannië	12	66.7	79.9	-2.17
Diergeneeskunde	nan	369	0.0	69.1	-62.994
Health and Environment	nan	44	0.0	69.1	-21.753
Engelse taal en cultuu...	nan	33	66.7	69.1	-0.654
Gezondheidswetenschapp...	nan	19	0.0	69.1	-14.294
Strategy, Competition ...	nan	26	0.0	69.1	-16.722
Nederlandse taal en cu...	nan	15	66.7	69.1	-0.441
Franse taal en cultuur...	nan	12	58.3	69.1	-1.776
Duitse taal en cultuur...	nan	16	56.2	69.1	-2.449
Media, Art and Perform...	nan	17	52.9	69.1	-3.17
Multidisciplinary Econ...	nan	29	65.5	69.1	-0.92
Religious Studies	nan	18	55.6	69.1	-2.718

**Table 6.5:** The table of underperforming combinations of country of previous education and study programme. 'nan' indicates that the country of previous education is unknown.