Algorithm Audit

Dienst Uitvoering Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

Preventing Prejudice

Addendum

May 2024

Table of Contents

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are three-fold:

Normative advice commissions

Forming diverse, independent normative advice commissions that advise on ethical issues emerging in real world use cases, resulting over time in algoprudence

Technical tools

Implementing and testing technical tools for bias detection and mitigation, e.g, bias detection tool, synthetic data generation

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see for instance our AI Policy Observatory and position papers

1. Executive summary

Below are the background and design of this study (§1.1) and the findings (§1.2).

1.1 Background and design of the study

This report is an addendum to the report Preventing prejudice, which was sent to the Dutch parliament on March 1, 2024[1]. In that report, an independent external investigation commissioned by the Education Executive Agency of the Netherlands (DUO) is presented. This investigation was conducted in response to media reports about the overrepresentation of students with a migration background in the so-called College Grant Check (in Dutch: controle uitwonendenbeurs, abbreviated as CUB). At the time of publishing the report Preventing prejudice, enriched data from Statistics Netherlands (CBS) about the origin of the monitored students was not yet available. On May 6, 2024, the CBS provided DUO and Algorithm Audit with group-level data on the origin of students at various steps of the CUB process for the years 2014, 2017, 2019, 2021, and 2022. This addendum presents the analysis of the CBS data. This addendum should be understood in conjunction with the report Preventing prejudice.

Below is a summary of the design and results of the study.

1.1.1 Design of the bias analysis

This report presents the results of a bias analysis of the CUB process. The bias analysis is based on aggregated data provided by the CBS on the origin of more than 300.000 recipients of the college grant for students living away from home (in Dutch: 'uitwonendenbeurs', from here on referred to as 'college grant') in 2014, 2017, 2019, 2021 and 2022[2]. By measuring the distribution of students with a migration background in various steps of the CUB process, it can be investigated whether, and if so, where and to what extent bias is present in the various steps of the CUB process. In 2.1 Steps in the CUB process for which data has been obtained, each step relevant to the bias analysis is described in detail.

This report also presents research into whether the criteria used in the risk profile constitute proxy characteristics for students with a migration background. This concerns the criteria of education type, age and distance to

[1] *Intern onderzoek controle uitwonendenbeurs* (report DUO), attachment to *Kamerstukken II* 2023/24, 24724, nr. 220.

[2] https://www.cbs.nl/nl-nl/maatwerk/2024/21/ontvangers-uitwonendenbeurs-herkomst-2014-2017-2019-2021-en-2022

Box 1

What are proxy characteristics?

Proxy characteristics refer to seemingly neutral data that are strongly correlated with sensitive personal characteristics or a protected ground under non-discrimination law, such as ethnicity or religion. Due to a strong correlation, using proxy characteristics can (possibly unintentionally) also affect the sensitive group. In the context of the CUB process, it is relevant to what extent the characteristics of the risk profile (I. education type, II. age and III. distance to parent(s)) and/or the various possible combinations of these criteria are a proxy characteristic for migration background.

parent(s) and the combinations of these criteria. This proxy analysis can explain how the risk model is biased towards students with a migration background, even though this characteristic itself is not a criterion in the profile (see Box 1).

This report also describes to which extent the steps in the CUB process, after the risk profile has assigned a risk score, contribute to the bias towards students with a migration background. For example, the influence of the manual selection of students for control was investigated, as well as whether students with a migration background had more home visits that did not reveal any unduly use of the college grant. Finally, the degree of overrepresentation of students with a migration background is measured for the group of students who DUO considered to have been receiving a grant unduly, for the group that subsequently appealed to DUO, and for the group for which this appeal was successful.

Depending on the data available for the individual years, some or all of these questions will be answered for the years 2014, 2019, 2021 and 2022. In 3.2 Research populations the composition of all student populations in this study is described. In 3.4 Research questions bias analysis with regards to migration background concrete research questions are introduced.

1.1.2 Definition of the term migration background

The term migration background used in this report is not utilized by the CBS itself. Instead, the CBS distinguishes between students who were born outside the Netherlands (migrants) or who have one parent born outside the Netherlands (children of a migrant) and students with Dutch origin. Within the two migrant categories, the CBS further differentiates between European and non-European migrants.

For the sake of clarity and alignment with the research question, the term migration background is used in this report and therefore represents a slightly different categorization compared to the CBS. Figure 1 provides an overview of the (sub)categories used in this report. When referring to students with a migration background, it includes students who were born outside the Netherlands or have one parent born outside the Netherlands. Where useful for analysis, the category migration background is divided into European and non-European migration backgrounds. Students who do not belong to any of these groups are referred to as students with Dutch origin.

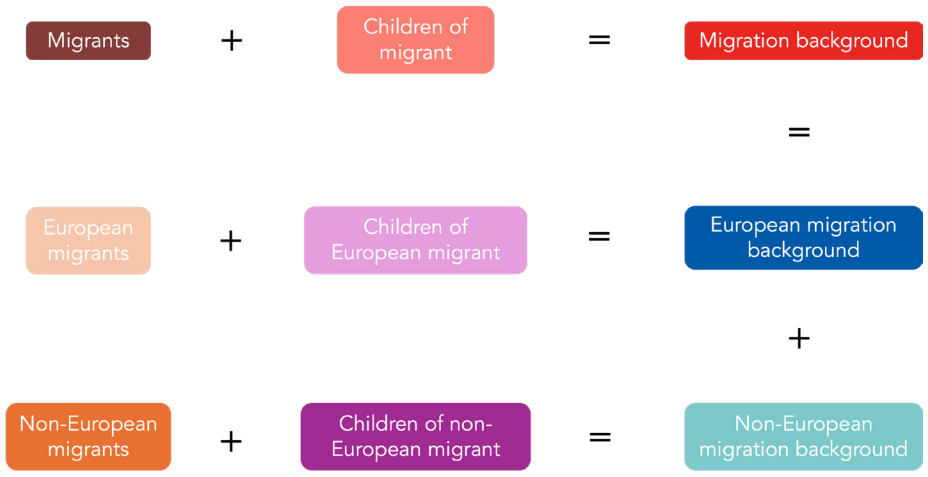

Figure 1 – Used subdivision of student categories

The groups with a migration background and Dutch origin together form the entire population of recipients of the college grant.[3] A detailed explanation and justification of this categorization, including its limitations, is provided in 3.3 Data provided by the CBS.

1.1.3 CBS data

By comparing data from different years of the CUB process, it can be determined whether any bias was a one-time occurrence or part of a structural trend. Aggregation statistics on the migration background of more than 300.000 students living away from home were requested from the CBS. These students fall into three categories: those who were part of the CUB process in 2014[4] and 2019[5]; those who were selected for a random sample in 2014 and 2017; or those who appealed to a decision by DUO regarding unduly use of the college grant in 2014, 2019, 2021, and 2022. Aggregation statistics cannot be traced back to individuals. These statistics only reflect characteristics of groups larger than 10 people. Unless otherwise stated, the results presented in this summary correspond to the data from 2014, as this is the most representative dataset for university, hbo and mbo students. Algorithm Audit sees no objection to using data that is ten years old because this research examines the functioning of the CUB process in the past. The research populations analyzed are described in more detail in 3. Research methodology.

[3] Students who come to the Netherlands from abroad to study do not receive the living-out college grant and therefore are not part of the population being studied here.

[4] 2014 was the last year before the introduction of the loan-based system, the population consists of mbo, hbo, and wo students. This is relevant because starting from the academic year 2023-24, the basic study grant (in Dutch: basisbeurs) has been reintroduced for all students.

[5] 2019 was the last year before the standard CUB process was disrupted by the COVID-19 pandemic. Due to the pandemic, approximately 33% fewer home visits could be conducted in 2020-2022. Note that in 2019, the population receiving the college grant mainly consisted of mbo students.

Box 2

Link with bias analysis in report Preventing prejudice

The bias analysis described in the report Preventing prejudice measured whether students in certain education, age, or distance categories were more often selected for a home visit than could be expected based on the risk score assigned by the risk profile. The bias analysis presented in this addendum examines bias towards students with a migration background. Both regarding the risk profile and in later steps of the CUB process, it is measured whether students with a migration background are more frequently selected for control than students without a migration background. This bias analysis deepens the analysis in the first report as it investigates the one-dimensional and multi-dimensional proxy nature of the three profiling categories concerning migration background.

Box 3

Bias

In this report, bias means the following: significant deviations in the demographic ratios compared to the source population (the population of all recipients of a college grant) that arise due to disproportionate selection of certain students during the CUB process. If a significant overrepresentation of students with a migration background is measured in various steps of the CUB process, this gives rise to the conclusion that the CUB process (and/or the various steps therein) is biased towards this demographic. Bias here does not imply conscious bias, i.e. premeditated and purposeful selection of a certain demographic by DUO or individual employees. In this report, bias refers to measured disproportions in the data as an (unintended) effect of CUB process steps (also known as bias). There is no bias if the demographic ratios remain virtually the same throughout the course of the CUB process.

1.2 Findings

This section presents the main findings of the bias analysis.

Finding 1 – The CUB process as a whole has been biased towards students with a non-European migration background.

In the CUB process, there was a strong bias towards students with a non-European migration background. Students with a non-European migration background were classified as high risk by the risk profile 2 times more often than students with Dutch origin. The group was manually selected for a home visit 6.2x more often. Ultimately, students with a non-European migration background had a 3.0x greater probability of receiving an unjustified home visit than students with Dutch origin.

Taking into account the substantially different populations in different years (largely due to the introduction of the loan-based system[6]), there is a clear structural trend in bias towards students with a non-European migration background.

Figure 2 provides an overview of the distribution of students from different origins per step of the CUB process. Figure 2 is further explained in the findings below and in 4. Results.

[6] In the period 2015-2023, only students enrolling in vocational education (mbo) were entitled to a college grant. Further details can be found in 3.2 Research populations.

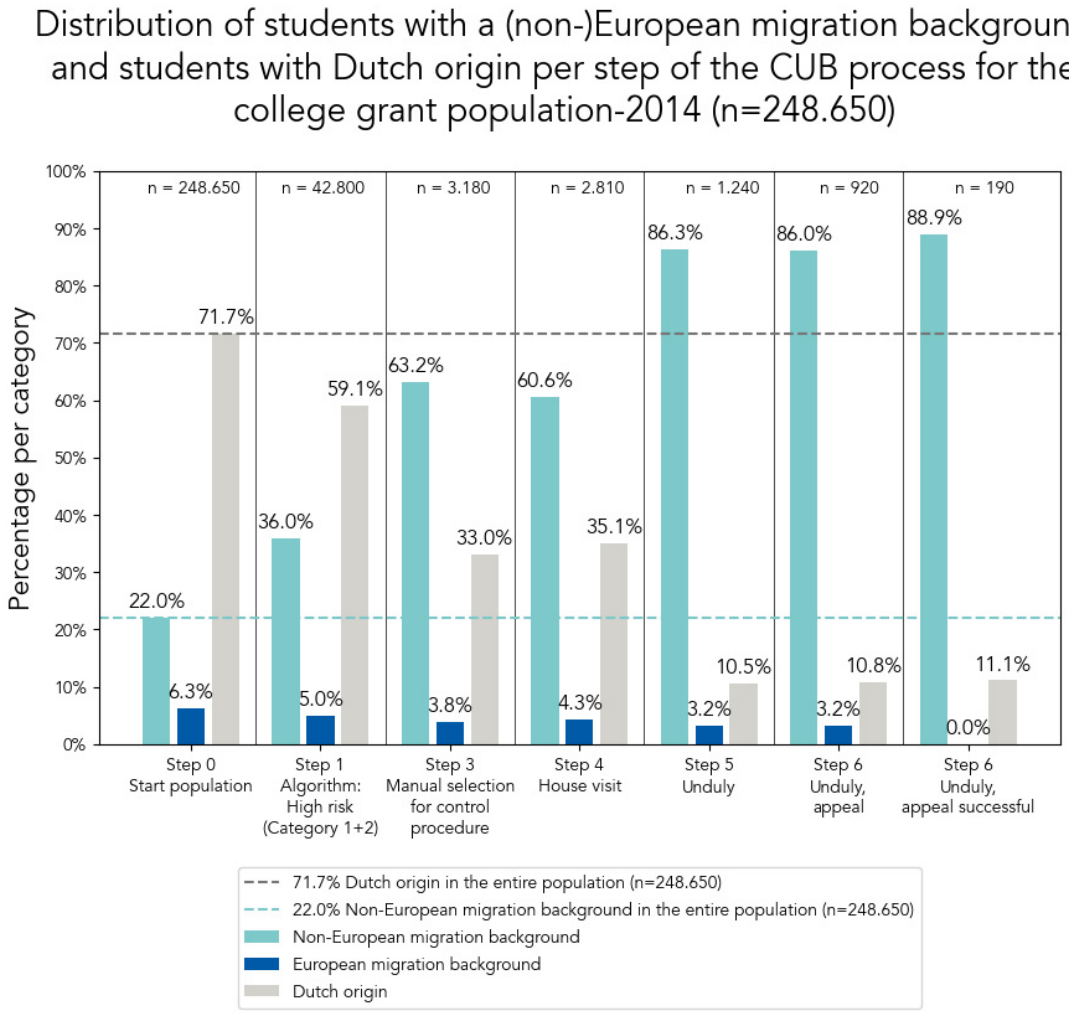

Figure 2 – Distribution of students with (non-)European migration background and students with Dutch origin per step of the CUB process in the college grant population-2014 (n=248.650)[7]

[7] Because the aggregation statistics compiled by CBS are rounded to the nearest ten, it is possible that percentages do not add up exactly to 100%. The same applies to the sum of the numbers.

Finding 2 – The risk profile used in the CUB process was biased towards students with a non-European migration background. The reason for this was the assignment of a higher risk to mbo students and to students who were registered close to their parental address.

In the utilized risk profile (which utilizes characteristics such as education type, age, and distance between the registered address and the parental address), bias occurred towards migration background. Students with a non-European migration background were 2.0 times more often classified as high risk by the risk profile than students with Dutch origin. This can be attributed to the proxy characteristics, particularly education type and distance to parent(s). Characteristics such as mbo 1-2 and a short distance to parent(s) correlate strongly with the group of students with a non-European migration background. Profiling based on these characteristics resulted in bias towards this demographic. An unsupervised bias detection method confirms these results.

The proxy analysis is further explained in 4.2 Results of research question 2 (risk profile) and 4.3 Results of research question 2a (proxy analysis). The results of the unsupervised bias detection method are explained in Appendix A – Unsupervised bias detection.

Finding 3 – Manual selection reinforced the bias of the CUB process.

After classification into risk categories by the risk profile, students were manually selected for inspection and home visits. In that manual selection process, written research, work instructions, certain exclusion criteria, and other procedural aspects played a role. The bias introduced by the risk profile is reinforced in this step. The study did not reveal any indications that this bias is due to personal prejudices of individual officials. Bias can also arise from the nature of work instructions, such as the exclusion of student housing from home visits and care facilities. Selection criteria within the group selected for conducting a home visit included factors such as the ratio between area and number of residents at the registered address, registration with family, and the distance from the registration to the parents' address. The question of how the bias in manual selection arose is beyond the scope of this study.

Students with a non-European migration background classified as high risk by the risk profile were 5.5 times more likely to be selected for inspection than students with Dutch origin in the same risk category. Students with a non-European migration background who were classified as low risk by the risk profile were 1.8 times more likely to be selected for inspection than students without a migration background in the same risk category. It should be noted that the assigned risk scores themselves are already biased.

The analysis of the manual selection of students is further explained in 4.4 Results of research question 3 (manual selection) and 4.5 Results of research question 3a (extent to which algorithm prediction is followed).

Finding 4 – Due to the bias in the CUB process towards students with a non-European migration background, a considerable amount of unduly use has been identified in this group. This is largely attributed to excessive scrutiny of this demographic.

Due to the bias of the entire CUB process, students with a non-European migration background were relatively more often selected for a home visit than students with Dutch origin. In the case of this group, this selection was also more often unjustified: students with a non-European migration background were 3.0 times more likely to receive a home visit that later did not reveal unduly use than students with Dutch origin.

Even among the group of students where unduly use is eventually determined, the proportion of students with a non-European migration background is large. Throughout the steps of the CUB process, a magnifying effect emerges concerning the group with a non-European migration background. Students with Dutch origin were relatively less frequently inspected, and unduly use is less frequently detected in this group.

Whether students with a migration background also make unduly use of the college grant more frequently cannot be determined based on the available data. This is because this ratio cannot be isolated from the bias of the CUB process, and because the random sample is too small to measure this independently of the CUB process. Additionally, it could not be investigated whether there is any bias in the home visit process itself beyond the mentioned magnifying effect.

The overrepresentation of this group of students in the unduly use population is further explained in 4.6 Results of research question 4 (unduly use).

Finding 5 – The group of students who appeal to a determination of unduly use consists largely of students with a non-European migration background. No bias has been identified in the appeal process itself.

From the various reference years (2014, 2019, 2021, and 2022), a consistent picture emerges that the appeal population consists of 79-85% students with a non-European migration background. This broadly confirms the image presented in investigative journalism regarding the strong overrepresentation of students with a migration background who appeal to a decision by DUO.[8]

On the other hand, no bias is identified in the appeal step of the CUB process. Appeals are equally likely to be successful for all students regardless of their origin.

The overrepresentation of this group of students in the appeal populations and equal treatment during the appeal process is further explained in 4.7 Results of research question 5 (appeal populations).

[8] B. Belleman, B. Heilbron & A. Kootstra. *De discriminerende fraudecontroles van Duo*. Investico Onderzoeksjournalisten, 2023

2. Overview of the CUB research

Below is an overview of the CUB process for which data has been obtained (§2.1), an overview of steps taken after publication of previous studies into the CUB process (§2.2) and an overview of these studies (§2.3).

2.1 Steps in the CUB process for which data has been obtained

In this addendum, data has been obtained on five steps in the CUB process. In the Report Preventing prejudice, the CUB process is divided into seven steps. Data has been obtained from five of these steps; it was not relevant to request data from the other two steps[9].

Steps from the CUB process that are relevant to the research:

- **Source population:** All recipients of a college grant for a given reference date. These are all students who are registered as students at mbo, hbo or wo and who are registered at an address that is not the address of their parent(s). It is possible that it will later become clear that the student was not entitled to the college grant. This group is referred to as the *college grant population*.

- **Risk profile:** Assigning a risk score between 0-180 to all students in the college grant population based on a risk profile. For the precise functioning of the risk profile, please refer to the report Preventing prejudice. For this addendum, it is sufficient to note that the risk profile resulted in a risk score based on three criteria:
  - age
  - education
  - distance to parent(s)

- **Manual check:** Reviewing a *risk score population per region* by a DUO employee. A DUO employee examines the list of students in order from high to low risk scores. The employee manually selects whether a home visit will take place or not, taking into account work instructions. More details on this step can be found in the report Preventing prejudice.

- **Unduly use:** Processing of home visit results. Determining duly and unduly use. The population that has made unduly use of the college grant is referred to as the unduly use population.

- **Appeal procedures:** Students for whom DUO has determined that they made unduly use of the college grant (the *unduly use population)* can appeal against this decision. This population is referred to as the *appeal population*.

The report Preventing prejudice describes all steps of the CUB process in detail.

[9] Division by region (step 2 from the report Preventing prejudice) is made solely from a practical standpoint. Students are not checked by DUO at this step. The home visit carried out by an external party (step 4 from the report Preventing prejudice) is included for the sake of clarity in the step overview. However, the outcome of this step is processed in step 5 (feedback, handling, and follow-up actions). Data regarding step 5 has been requested from CBS.

2.2 Recent chronology of CUB research

For a description of the chronology of the CUB process until March 1, 2024, please refer to the report Preventing prejudice. Below is a concise summary of relevant events since then.

> March 1, 2024 - The report Preventing prejudice has been shared with the parliament as part of DUO's internal investigation. The external research commissioned by the Ministry of Education, Culture and Science (OCW) and conducted by PricewaterhouseCoopers (PwC) has also been shared with the parliament.

> March 1, 2024 - The government has responded to the investigation into the College Grant Check. In its response, the government mentions, among other things:

"DUO has engaged independent researchers from non-profit organization Algorithm Audit. DUO will have the CBS conduct further research in the coming period into the overrepresentation of students with a migration background in the audit process, using data that was not available to PwC within the duration of the study."[10]

> March 1, 2024 – Director-General of DUO and Minister Dijkgraaf of OCW have apologized for indirect discrimination in the CUB process.[11]

> March 21, 2024 - The parliament has adopted a motion by member of the parliament Soepboer in which the parliament requests the government to

"Also let the CBS explicitly look at the correlations between the parameters used and a person's nationality in this research. In the follow-up process, the (statistical) models used and the awareness of the choices for the use of these parameters and this model will also be considered."[12]

> May 6, 2024 – The CBS shares the data requested by DUO. Those data were made public on May 21, 2024.[13]

[10] *Kamerstukken II* 2023/24 p.6, 24724, nr. 22

[11] *Kamerstukken II* 2023/24, 24724, nr. 220

[12] *Kamerstukken II*, 2023/24, 24724, nr. 237 ter vervanging van de op 21 maart 2024 ingediende motie *Kamerstukken II*, 2023/24, 24724, nr. 234.

[13] See https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/overzicht-aanvullend-statistisch-onderzoek-per-jaar

2.3 Other research on the CUB process

In addition to Algorithm Audit's investigation, other investigations into possible bias in the CUB process have also been conducted or announced. On behalf of the Ministry of Education, Culture and Science, PwC investigated, among other things, how the CUB process came about and whether the CUB process could lead to discrimination.[14] Based on a data study in which migration statistics of students living away from home were analyzed at postal code level, PwC concluded that it cannot be established with complete certainty that like cases were treated equally in the CUB process in the period 2012-2023.

The Dutch Data Protection Authority is also investigating the CUB process. The result of that study is expected in June 2024.

The main difference between this addendum and the other studies is the availability of the CBS data on the migration background of the student populations studied. This makes it possible to in this addendum determine quantitatively whether there was bias towards students with a certain migration background in the CUB process.

[14] The entire study consisted of more questions. Onderzoek misbruik uitwonendenbeurs PwC, 21 januari 2024 V1F6 https://www.rijksoverheid.nl/documenten/rapporten/2024/03/01/eindrapport-pwc-rapportage-onderzoek-misbruik-uitwonendenbeurs

3. Research methodology

The aim of this follow-up study is explained (§3.1). The research populations (§3.2) and data requested from the CBS (§3.3) are described. Additionally, the methodology for analyzing bias in the CUB process with regards to migration background is explained (§3.4).

3.1 Aim of the analysis

The aim of this research is to trace possible (multidimensional) bias towards students with a migration background in different steps of the CUB process. By carrying out a measurement per step, targeted follow-up research can be conducted into where possible bias occurs in the CUB process. This analysis is called a *bias analysis*. An overview of relevant steps in the CUB process can be found in 2.1 Steps in the CUB process for which data has been obtained.

3.2 Research populations

For this study eight populations were selected for further investigation. Two populations consist of students to whom the CUB process applied in 2014 and 2019. Two populations refer to the students selected in the 2014 and 2017 random samples. The remaining four populations consist of the students who initiated appeal procedures in 2014, 2019, 2021 and 2022. The composition of the populations and the reason for their use in this study are explained below.

CUB populations

The first CUB population concerns students who received a college grant before the introduction of the loan-based system. The reference date 01-02-2014 applies to this population. This population, consisting of university, hbo and mbo students, is referred to as the *college grant population-2014* (n=248.649). The second CUB population concerns students who received a college grant after the introduction of the loan-based system. For this population, consisting of mbo students and phasing-out hbo and wo students who were entitled to a college grant before 2015, the reference date is 01-02-2019. This population is referred to as the *college grant population-2019* (n=50.233). If possible, the analysis for 2019 will focus entirely on mbo students (n=36.630), as these students belong to the primary CUB target group.

Box 4

Statistical significance

In this analysis, aggregate statistics were used on the origin of the entire population of students living away from home, compiled by the CBS. Therefore, no estimates are made in this analysis. The figures reflect the actual situation at population level. The population in this case concerns students who applied for the college grant on the first of February in 2014 or 2019 and who were subject to the CUB process. In this report, the term statistical significance is used only in connection with the random samples.

College grant population-2014 and college grant population-2019 are followed for the relevant steps of the CUB process. An overview of all relevant steps can be found in 2.1 Steps in the CUB process for which data has been obtained.

Random sample populations

The random sample populations concern students who were drawn in 2014 and 2017 for a random sample. These populations are referred to as the *random sample population-2014* and *random sample population-2017*. Both the CUB populations and the random sample populations are examined because these populations provide different information. The difference between the two types of populations consists of the fact that the risk profile and manual selection have been applied to the CUB populations. The risk profile and manual selection were not applied to the random sample populations because students were randomly selected for control. The data from the random sample provide information about the unduly use percentage per group, regardless of the application of the risk profile and manual selection for control from the CUB process. Results of the analysis of a relationship between the used criteria from the risk profile and unduly use of the college grant based on the random samples can be found in the report Preventing prejudice. Alternative samples from other years are not available.

Appeal populations

In addition to the CUB populations and the random sample populations, four additional populations are examined to accurately analyze the appeal procedures. College grant population-2014 and college grant population-2019, determined on the basis of the respective reference dates 01-02-2014 and 01-02-2019, contain less than 50% of the appeal procedures that take place following a selection moment and home visit in calendar year 2014 or 2019. The reason for this is that many students apply for a college grant after February 1st, and then in the same calendar year, they are selected for a home visit, receive an irregularity decision, and appeal against the decision. To obtain a comprehensive understanding of the entire appeal population, it was decided to also analyze students who initiated an appeal procedure in the entire calendar year 2014 or 2019. These groups are referred to as *appeal population-2014* and *appeal population-2019*. In addition, the *appeal population-2021* and *appeal population-2022* have been requested. CUB population-2021 and CUB population-2022 were not requested as the CUB process was disrupted in these years by corona measures.[5]

Algorithm Audit sees no objection to the use of ten-year-old data because this research examines the functioning of the CUB process in the past.

3.3 Data provided by the CBS

For the college grant population-2014, college grant population-2019, random sample population-2014, random sample population-2017, appeal population-2014, appeal population-2019, appeal population-2021 and appeal population-2022, aggregated data on the country of origin and country of birth of students was requested from the CBS. The CBS has compiled the aggregation statistics for the following two variables for each population:

> **Country of Birth**[15] – The country where a person was born. Values: *Born in the Netherlands, born outside the Netherlands.*

> **Country of origin**[16] (new classification) – Characteristic indicating in which country someone was born or where one of their parents was born. The country of origin of persons born abroad is determined by their own country of birth. For individuals born in the Netherlands, the country of origin is determined by the country of birth of the parents. When both parents were born abroad, the mother's birth country is decisive in determining the country of origin. The birth data of the mother are known more often than those of the father. If the mother was born in the Netherlands or the mother's country of birth is unknown, the father's country of birth is used. Values: *Netherlands, Europe (excl. Netherlands) and outside Europe.*[17]

Third-generation migrants are not included in this study. These are students who were born in the Netherlands, both of their parents were born in the Netherlands, but one or more of their grandparents was born abroad. The CBS does not keep track of this specific group. Therefore, potential bias of the CUB process towards this group has not been investigated.

Since 2022, the CBS no longer uses the term migration background. Instead, a new classification is used, formerly known as 'population with a Western or non-Western migration background'. In the new classification, it is more important where someone was born, and less important where someone's parents were born. The new classification is based on continents (the Netherlands, Europe excluding the Netherlands and outside Europe) and no longer based on Western and non-Western countries.[18]

[15] Also see https://www.cbs.nl/nl-nl/onze-diensten/methoden/begrippen/geboorteland

[16] Also see https://www.cbs.nl/nl-nl/onze-diensten/methoden/begrippen/herkomstland

[17] The countries classified as Europe and outside Europe can be found here: https://www.cbs.nl/nl-nl/maatwerk/2024/08/landenindeling-van-de-variabele-herkomstland-2022

[18] Also see https://www.cbs.nl/nl-nl/nieuws/2022/07/cbs-introduceert-nieuwe-indeling-bevolking-naar-herkomst

Box 5

Unknown

There are conceivable variations in which it cannot be determined into which category a student falls. For example, because it is not known where the parents were born or because the term parent is different in a foreign legal system than in the Dutch legal system. Including provisions for all possible nuances goes beyond the scope of this study. The CBS can answer specific questions about dealing with such situations.

Based on the combination of the variable country of birth and country of origin, the CBS has distinguished students into five migrant groups:

1. **Non-European migrants:** these students were born outside of Europe.
2. **European migrants:** these students were born within Europe but outside of the Netherlands.
3. **Children of non-European migrants:** these students were themselves born in the Netherlands, but at least one of their parents was born outside Europe.
4. **Children of European migrants:** these students were themselves born in the Netherlands, but at least one of their parents was born in Europe, but not outside the Netherlands.
5. **Dutch origin:** these students and both their parents were born in the Netherlands. Third-generation migrants belong to this group.

The combination 'not born in the Netherlands x country of origin Netherlands' cannot occur for logical reasons (someone with Netherlands as country of origin is by definition born in the Netherlands) and has been excluded from the analysis. A schematic overview of the above groups is given in Figure 3. A similar overview in which the above terms are linked to the term migration background can be found in Figure 1.

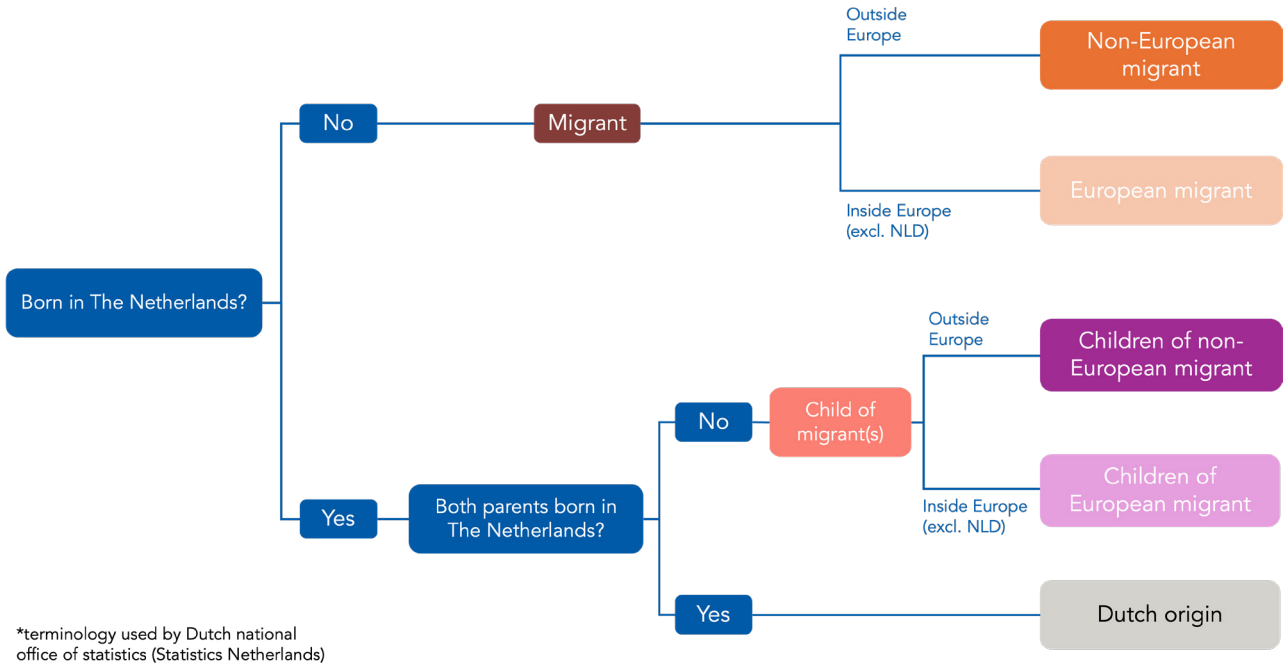

Figure 3 – Overview of the categories of students as distinguished by the CBS

Country of birth and country of origin, and the combination of these variables, are requested for the following (sub)populations:

- College grant population-2014, College grant population-2019, random sample population-2014, random sample population-2017, appeal population-2014, appeal population-2019, appeal population-2021 and appealpopulation-2022
- Education type (univariate, 4 values: mbo 1-2, mbo 3-4, hbo, wo)
- Age (univariate, 5 values: 15-18, 19-20, 21-22, 23-24, 25-50)
- Distance (univariate, 9 values: 0km, 1m-1km, 1-2km, 2-5km, 5-10km, 10-20km, 20-50km, 50-500km, unknown)
- Education type x age: (bivariate, 4x5=20 combinations)
- Education type x distance (bivariate: 4x9=36 combinations)
- Age x distance (bivariate 5x9=45 combinations)
- Education type x age x distance (trivariate: 4x5x9=180 combinations)
- Risk category (univariate, 6 values: Risk category 1, 2, 3, 4, 5, and 6)
- Selected for control (univariate, 2 values: selected for control or not)
- Outcome of home visit (univariate, 4 values: duly, unduly, home visit could not take place, not selected for control)
- Appeal procedure (univariate, 2 values: appeal or no appeal)
- Appeal outcome (univariate, 3 values: successful, partially successful, unsuccessful)
- Risk category x selected for control (bivariate: 6x2=12 combinations)
- Risk category x outcome of home visit (bivariate: 6x2=12 combinations)

Distributed over five different years, aggregate statistics about the migration background of 1.401 groups of students are requested. For the college grant population-2014 and -2019, aggregate statistics about the migration background of the above 341 groups are requested. For the random sample population-2014, aggregate statistics about the migration background for 146 groups are requested (all the above groups without trivariate, selected for control, and risk category x selected for control). For the random sample population-2019, aggregate statistics about the migration background for 116 groups are requested (the same as for the random sample population-2014 but without the values hbo and wo for education type). For the appeal population-2014, -2019, -2021, and -2022, aggregate statistics about the migration background for 129, 114, 99, and 115 groups respectively are requested. The exact data delivery can be found on the CBS website.[4]

3.4 Research questions bias analysis with regards to migration background

Using the data above, the ratio of students per combination of country of origin and country of birth can be measured for all steps of the CUB process and for the different (sub)populations. Based on the following research questions, potential bias is investigated for each step of the CUB process (see 2.1 Steps in the CUB process for which data has been obtained).

Research question 1 (Step 0 – initial population)
What was the distribution of students with a migration background in the college grant population-2014 and -2019?

Research question 2 (Step 1 – risk profile)
Were students with a migration background in 2014 and 2019 overrepresented in the higher risk categories that followed from the risk profile?

Research question 2a (Proxy analysis of the risk profile)
Is there a relationship between criteria used in the risk profile and students with a migration background?

Research question 3 (Step 3 – manual selection)
Were students with a migration background in 2014 and 2019 more often manually selected for a control procedure than students with Dutch origin?

Research question 3a (Extent to which algorithm prediction is followed)
To what extent does the classification of the risk profile (high or low risk) align with manual selection for a control procedure, in particular for students with a migration background compared to students with Dutch origin?

Research question 4 (Step 5 – unduly use)
What was the distribution of students with a migration background in the group of students considered to have been receiving a grant unduly?

Research question 5 (Step 6 – appeal procedure)
What was the distribution of students with a migration background who appealed DUO's decision in 2014, 2019, 2021 and 2022?

4. Results

The proportion of students with a migration background in the entire college grant population is presented (§4.1). This is followed by the bias analysis of the risk profile (§4.2), proxy analysis of the risk profile (§4.3), bias analysis of the manual check (§4.4), an analysis of the extent to which the algorithm's prediction was followed during manual selection (§4.5), determination of the proportion of students with a migration background in the population of improper use (§4.6) and in the appeal population (§4.7). The code used to produce the analysis below is available online.[19]

4.1 Results of research question 1 (initial population)

This paragraph presents the results of the bias analysis for step 0 of the CUB process. A precise description of the college grant population-2014 and -2019 is given in 3.2 Research populations. Histograms displaying the distribution of variables relevant to the CUB process are shown in the report Preventing prejudice.

Research question 1

What was the distribution of students with a migration background in the college grant population-2014 and -2019?

Answer to research question 1

Of the 248.650 students who were eligible for the college grant on 01-02-2014 (college grant population-2014), 22% had a non-European migration background. Of these:

> 9.5% were non-European migrants and
> 12.5% were children of a non-European migrant.

6.3% of the college grant population-2014 belonged to the group with a European migration background. Of these:

> 3.1% were European migrants and
> 3.2% were children of a European migrant.

In other words, 28.3% of the students in the college grant population-2014 had a migration background, and 71.7% of the students were of Dutch origin. See Figures 4-6.

Of the 36.630 recipients of a college grant on 01-02-2019, who studied at vocational schools (mbo students within the college grant population-2019[20]), 47.7% belonged to the group with a non-European migration background. Of these:

> 27.8% were non-European migrants and
> 19.9% were children of a non-European migrant.

[19] https://github.com/NGO-Algorithm-Audit

[20] A clarification on why only mbo students in 2019 are being examined can be found in 3.2 Research populations.

5.2% of the college grant population-2019 belonged to the group with a European migration background. Of these:

> 2.2% were European migrants and

> 2.9% were children of a European migrant.

In other words, 52.8% of the students in the college grant population-2019 had a migration background, and 47.2% of the students were of Dutch origin. See Figures 7-9.

Elaboration on research question 1

First, the distribution of all migrant groups is presented for the college grant population-2014. Following this, the distribution is presented for mbo students within the college grant population-2019.

Diagrams and numbers 2014

Figure 4 shows the distribution of students with a migration background and students with Dutch origin in the college grant population-2014.

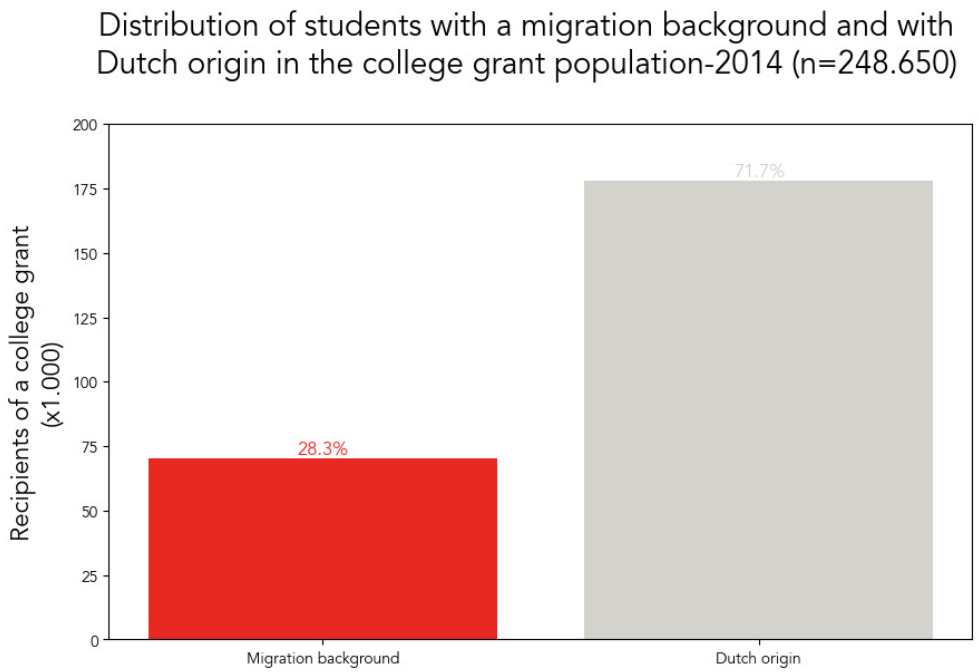

Figure 4 – Distribution of students with a migration background and with Dutch origin in the college grant population-2014 (n=248.650)

Figure 5 shows the distribution of students with a (non-)European migration background and students with Dutch origin in the college grant population-2014.

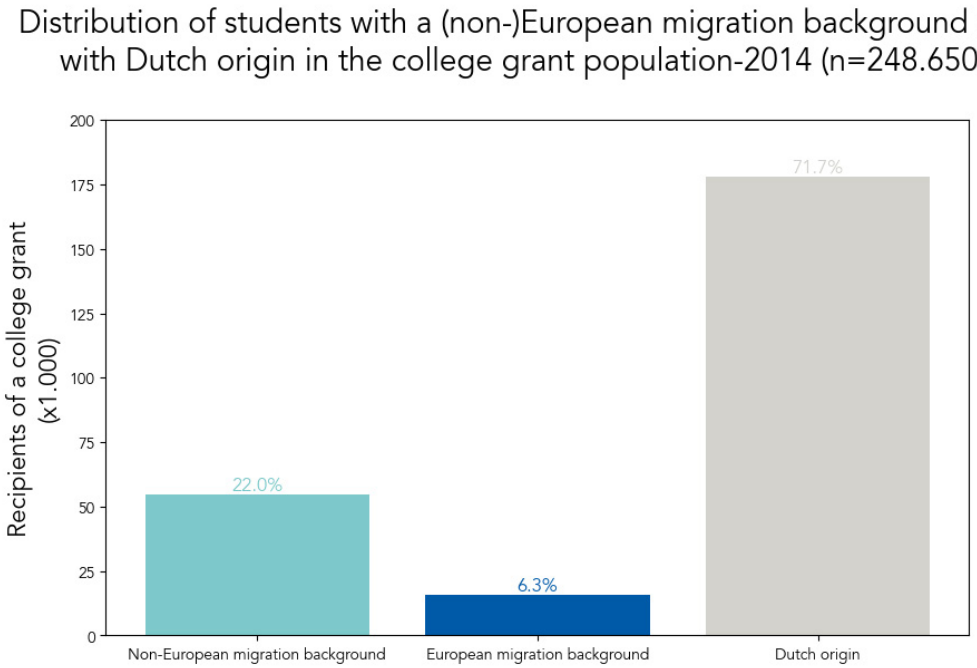

Figure 5 – Distribution of students with a (non-)European migration background and with Dutch origin in the college grant population-2014 (n=248.650)

Figure 6 shows the distribution of (non-)European migrants, children of (non-)European migrant, and students with Dutch origin in the college grant population-2014.

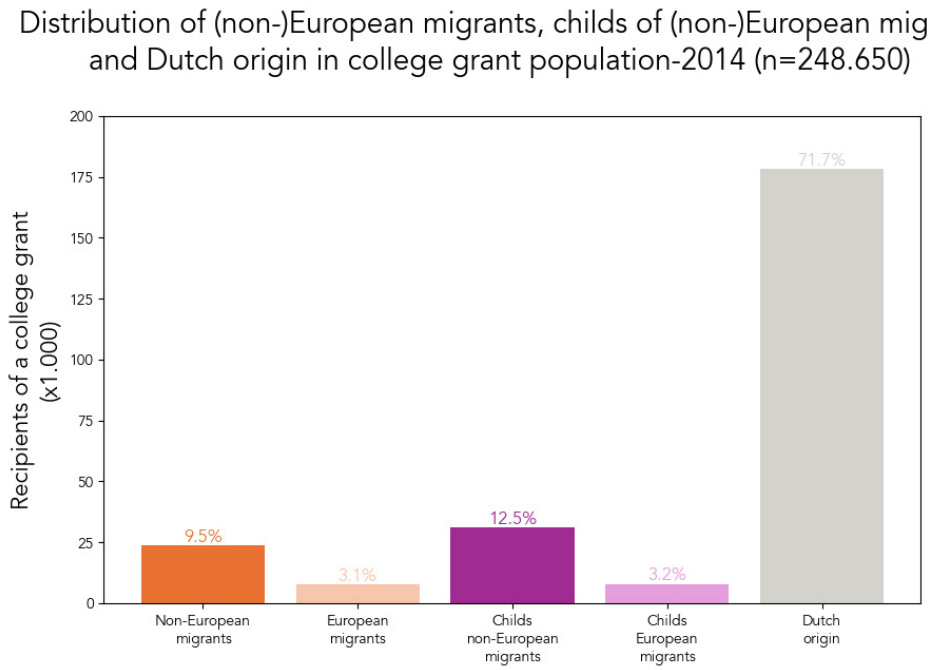

Figure 6 – Distribution of (non-)European migrants, children of (non-)European migrant, and stud ents with Dutch origin in the college grant population-2014 (n=248.650)

Diagrams and numbers 2019

Figure 7 shows the distribution of mbo students with a migration background and mbo students with Dutch origin in the college grant population-2019.

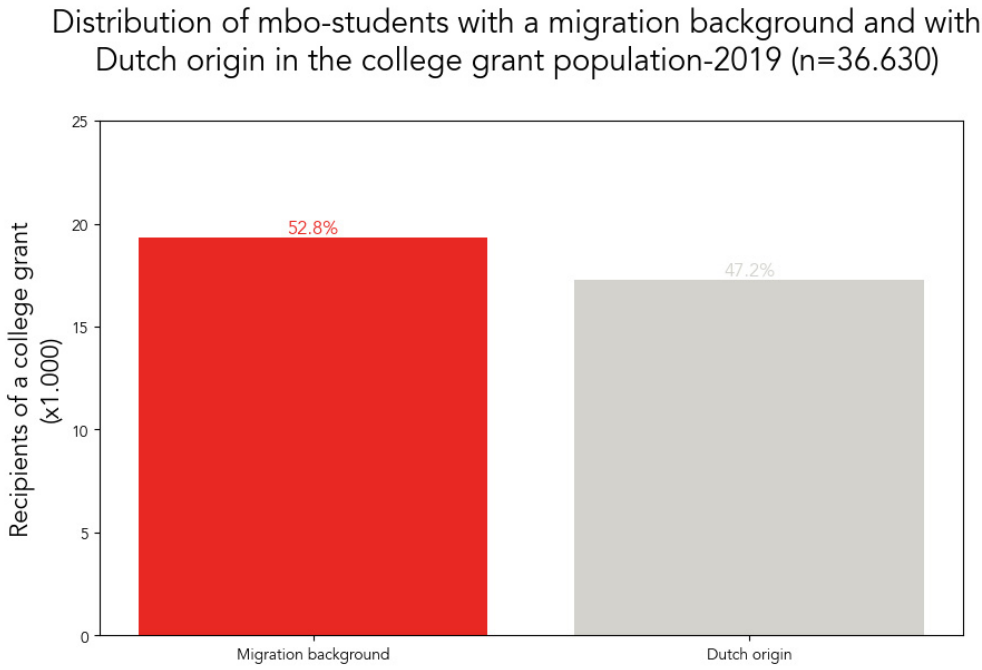

Figure 7 – Distribution of mbo students with a migration background and mbo students with Dutch origin in the college grant population-2019 (n=36.630)

Figure 8 shows the distribution of mbo students with a (non-)European migration background and mbo students with Dutch origin in the college grant population-2019.

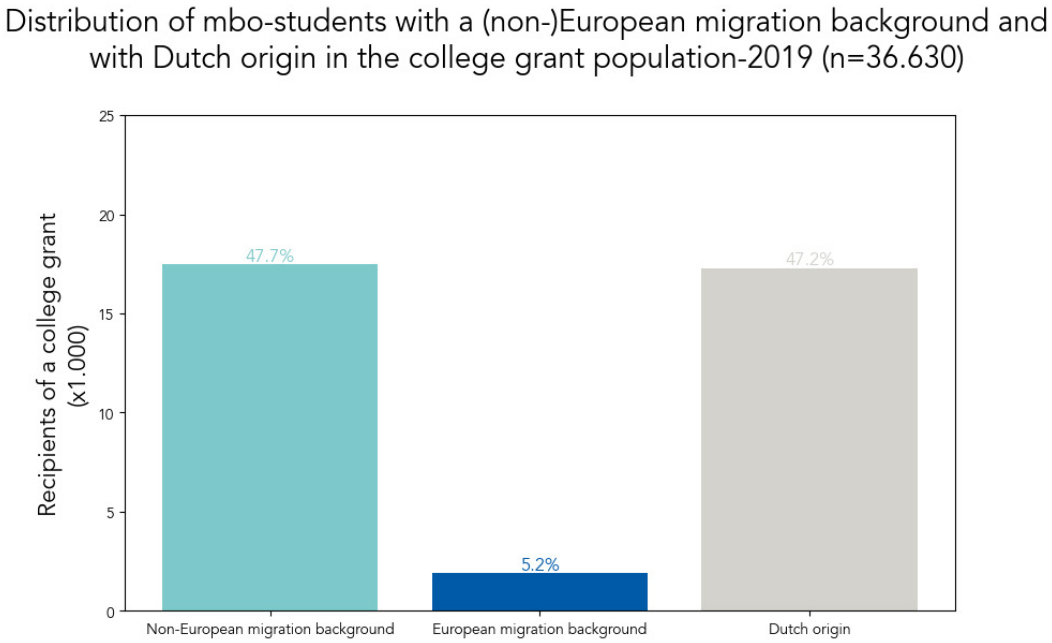

Figure 8 – Distribution of mbo students with a (non-)European migration background and mbo students with Dutch origin in the college grant population-2019 (n=36.630)

Figure 9 shows the distribution of (non-)European migrants, children of (non-)European migrant, and students with Dutch origin, only for mbo students.

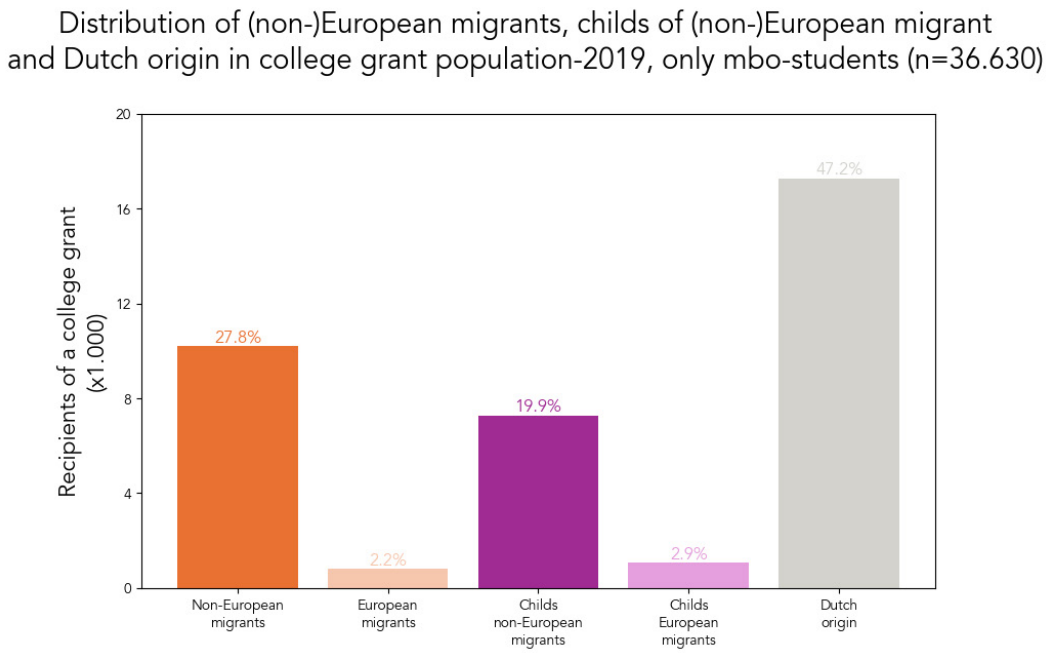

Figure 9 – Distribution of (non-)European migrants, children of (non-)European migrant, and students with Dutch origin, only for mbo students (n=36.630)

4.2 Results of research question 2 (risk profile)

Research question 2 addresses the possible overrepresentation of students with a migration background in step 1 of the CUB process (application of the risk profile). A description of how the risk profile functions is provided in 2.1 Steps in the CUB process for which data has been obtained.

Research question 2

Were students with a migration background in 2014 and 2019 overrepresented in the higher risk categories that followed from the risk profile?

Answer to research question 2

Yes. Students with a migration background were overrepresented in the higher risk categories following from the risk profile in both 2014 and 2019. Specifically, students with a non-European migration background, and then particularly children of a non-European migrant, were overrepresented in the higher risk categories in both years.

Elaboration on research question 2

Overrepresentation occurs when the proportion of students with a migration background in a risk category, as assigned by the risk profile, is greater than in the initial population (Box 3). The answer to Research question 1 shows that in the college grant population-2014, 22.0% of the students had a non-European migration background. In 2019, this was 47.7%.

The distribution of students with a (non-)European migration background and students with Dutch origin in the risk categories assigned by the risk profile is given in Figure 10 and Figure 12 for the college grant populations-2014 and -2019. The same distribution, but broken down into (non-)European migrants, children of a (non-)European migrant, and students with Dutch origin, is given in Figure 11 and Figure 13.

In 2014, there was an overrepresentation of 21.5 absolute percentage points of students with a non-European migration background in risk category 1 (very high) (43.5% compared to the average of 22.0%, see Figure 10). For the high-risk category, there was an overrepresentation of 11.2 absolute percentage points (31.2% compared to 22.0%)[21]. In 2014, an underrepresentation of students with a non-European migration background was observed in the three lowest risk categories. Consequently, the opposite effect is seen for students with Dutch origin. Based on the 2014 data, students with a non-European migration background were 2.0 times more likely to be assigned to a high-risk category (category 1 or 2) than students with Dutch origin.[22]

Figure 11 shows that in 2014, within the group of students with a migration background, particularly the group of children of a non-European migrant were strongly overrepresented in the highest risk categories. For example, 32.5% of students in risk category 1 (very high) belonged to the group of children of a non-European migrant. This decreases stepwise to 8.2% in risk category 5 (very low). This effect is not present to the same extent for the group of non-European migrants. For this group, the strong overrepresentation is noticeable in risk category 6 (unknown) of 26.4%.

The high number of non-European migrants with an unknown risk category can be explained as follows: a student living away from home is assigned to the unknown risk category if the distance category is unknown. The distance category is unknown if the address of both parents is unknown. This can occur if 1) it is known who the parents are, but their address is unknown, or 2) the parents are unknown. For non-European migrants, including refugees and immigrants, it is generally more common that the parents of a student are unknown. Therefore, there are relatively many non-European migrants with an unknown distance category, and thus an unknown risk category.

In 2019, the trend is similar for lower risk categories: students with a non-European migration background are underrepresented, and students with Dutch origin are overrepresented in low-risk categories (Figure 12)[23]. Based on the 2019 data, students with a non-European migration background were 0.9 times more likely to be assigned to a high-risk category (category 1 or 2) than students with Dutch origin. However, non-European migrants were even more overrepresented in risk category 6 (unknown) in 2019 than in 2014 (see explanation above). For children of a non-European migrant, a similar overrepresentation is visible in risk categories 1-2 as in the 2014 data (Figure 13).

---

[21] This deviation does not need to be tested for statistical significance since data for the entire population was available for analysis. Deviations are therefore not estimated but factually determined. See also Box 4.

[22] This probability is determined by comparing the probability that a student with a non-European migration background is classified as high risk to the probability that a student with Dutch origin is classified as high risk. The probability is determined based on the college grant population-2014.

[23] Note: the assigned risk categories per education type are not available in the CBS data. Figures 13-14 are therefore based on all students (mbo, hbo and wo) in the college grant population-2019, rather than only mbo students.

In summary, the risk profile in 2014 and 2019 clearly assigned a higher risk score to students with a non-European migration background, except for the unknown category. This means that these students were higher on the lists for manual selection for control in step 3 of the CUB process due to the use of the risk profile in proportion to students with Dutch origin. Whether students with a non-European migration background were actually more frequently selected for control is discussed in 4.4 Results of research question 3 (manual selection). First, a proxy analysis of the criteria used in the risk profile follows in 4.3 Results of research question 2a (proxy analysis).

Diagrams and numbers 2014

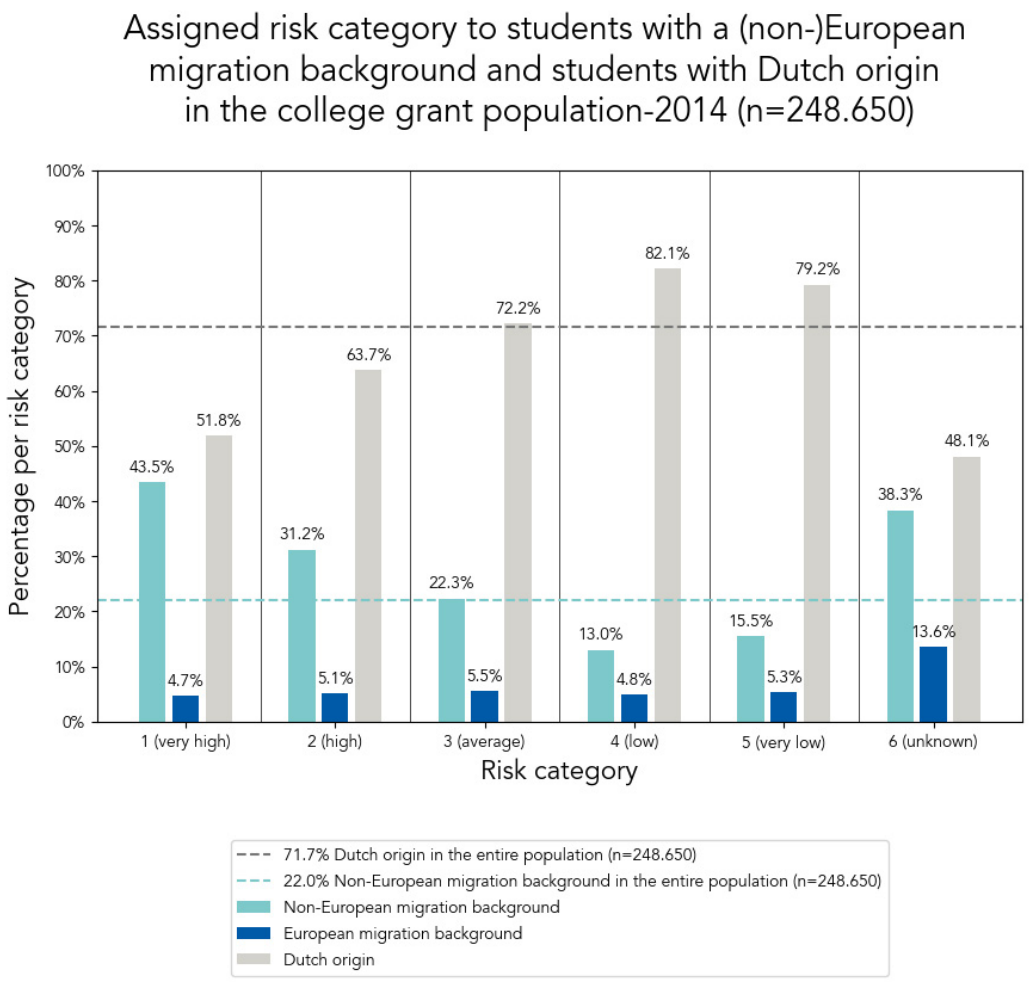

Figure 10 – Assigned risk category to students with a (non-)European migration background and students with Dutch origin in the college grant population-2014 (n=248.650)

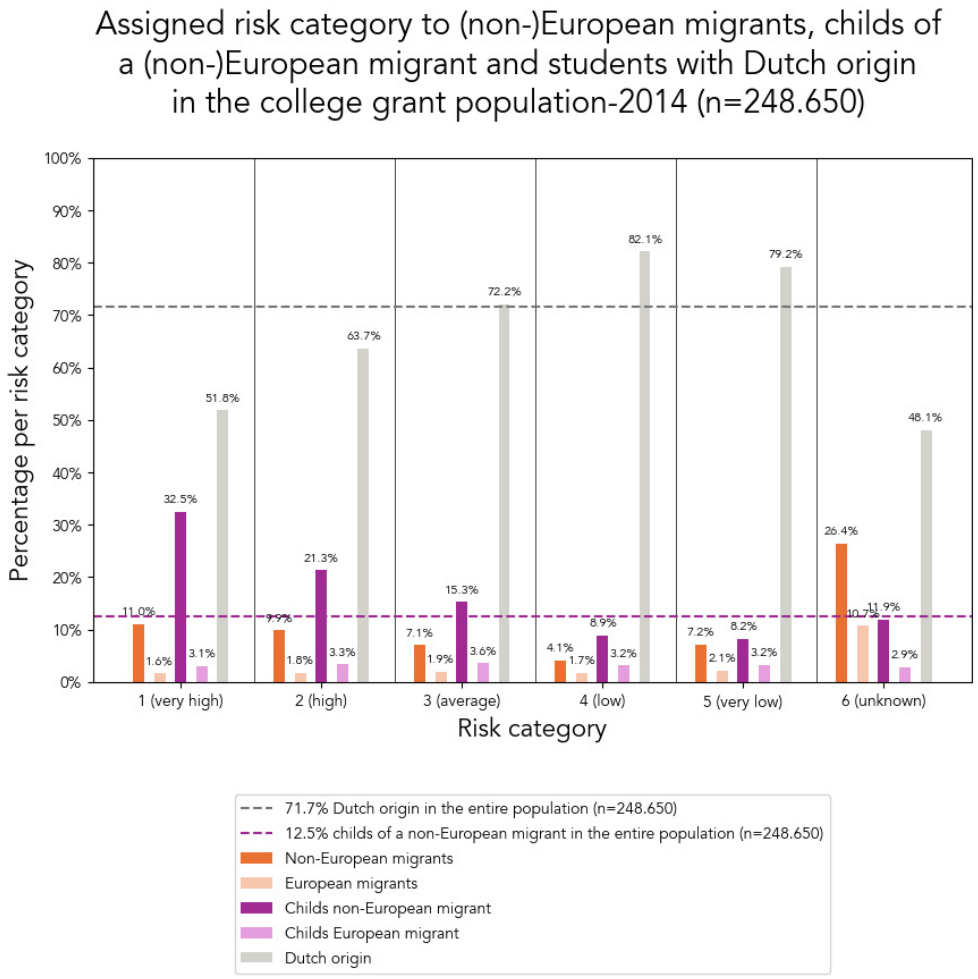

Figure 11 – Assigned risk category to non-European migrants, childs of a (non)-European migrant and students with Dutch origin in the college grant population-2014 (n=248.650)

Diagrams and numbers 2019

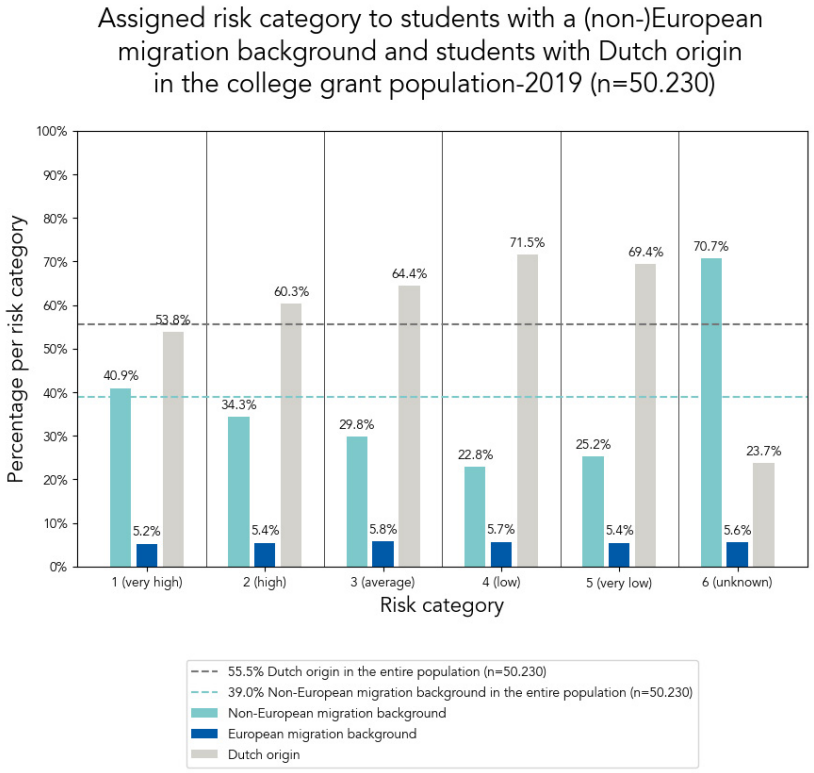

Figure 12 – Assigned risk category to students with a (non-)European migration background and students with Dutch origin in the college grant population-2019 (n=50.230)

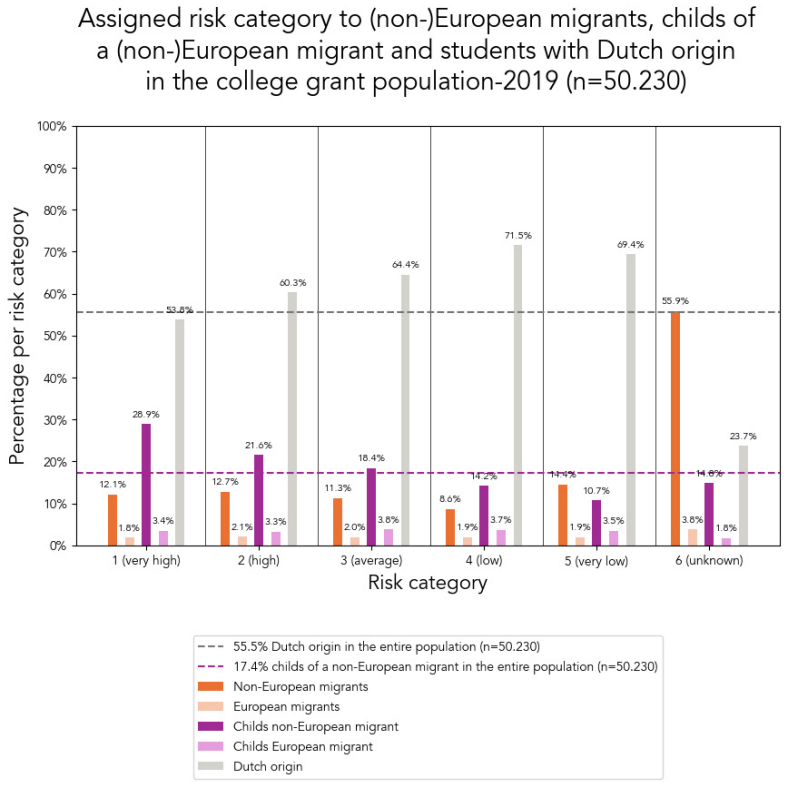

Figure 13 – Assigned risk category to non-European migrants, childs of a (non)-European migrant and students with Dutch origin in the college grant population-2019 (n=50.230)

4.3 Results of research question 2a (proxy analysis)

Research question 2a examines the relationship between criteria used in the risk profile and students with a migration background.

It is investigated whether the characteristics used in the risk profile (education type, age, and distance to parent(s), and a combination of these factors) are proxy characteristics for students with a migration background.

Research question 2a

Is there a relationship between criteria used in the risk profile and students with a migration background?

Answer to research question 2a

For all characteristics and combinations thereof, there is a relationship with students with a non-European migration background, although the strength and form of this proxy characteristic vary. The characteristics of education type and distance to parent(s) contribute to the bias of the risk profile against students with a non-European migration background through their proxy characteristics. The characteristic of age has the opposite effect. The proxy characteristics of each criterion are explained as follows:

- **Education type:** strong proxy characteristic for students with a non-European migration background. mbo students are significantly more likely to have a non-European migration background than students with Dutch origin. Since mbo students were assigned a higher risk by the risk profile, students with a non-European migration background were also assigned a higher risk.
- **Age:** inverse proxy characteristic for students with a non-European migration background. Older students, mainly in the age category 25-50, are more likely to have a non-European migration background than students with Dutch origin. Since younger students were assigned a higher risk by the risk profile, students with a non-European migration background were assigned a lower risk.
- **Distance to parent(s):** strong proxy characteristic for students with a non-European migration background. Students living away from home with an address close to the parental address are more likely to have a non-European migration background than students with Dutch origin. Since students registered close to the parental address were assigned a higher risk by the risk profile, students with a non-European migration background were also assigned a higher risk.

A two-dimensional analysis, by looking at the proxy characteristics of young mbo students (education type and age combined) and wo students who live close to their parents (education and distance combined), confirms the above picture. The inverse proxy characteristic of age does not offset the strong proxy characteristic of education type or distance.

Elaboration on research question 2a

A characteristic has a proxy character if there is a relationship between the characteristic and the group of students with a non-European migration background.

Figures 14 and 20 show a clear relationship between the characteristic education type and students with a non-European migration background. For example, in 2014 and 2019, 63.3% and 72.1% of mbo 1-2 students had a non-European migration background, respectively. This decreases stepwise to 13.2% for wo students in 2014 and to 38.5% for mbo 3-4 students in 2019. The education type characteristic is thus a proxy characteristic for the non-European migration background characteristic.

For the characteristic age, there is only a relationship with students with a non-European migration background in the age category 25-50 years. In the age categories 15-18, 19-20, and 21-22, the share of students with a non-European migration background in 2014 is stable (Figure 15). In the age group 23-24 years, this slightly increases. In the age group 25-50 years, this share strongly increases to 43.2%. In 2019, this relationship is less strong; the increase is from 54.5% to 67.6% between the age categories 23-24 and 25-50. Based on the data from 2014 and 2019, students with a non-European migration background are older than students with Dutch origin (Figure 21). The risk profile assigned a higher risk score to younger students. Therefore, a low age is not a proxy characteristic disadvantageous to students with a non-European migration background. Regarding the age characteristic, students with a non-European migration background received a lower risk score than students with Dutch origin. How this effect relates to the other characteristics is further elaborated below.

There is a clear relationship between the characteristic distance to parent(s) and students with a non-European migration background. For example, in 2014 and 2019, 42.1% to 44.1% of students who live 1 meter to 1 kilometer from their parent(s) are students with a non-European migration background, and this percentage decreases stepwise to 11.1% for the distance category 50-500 kilometers in 2014 (Figure 16) and to 27.9% in 2019 (Figure 22). The distance to parent(s) characteristic is thus a proxy characteristic for the characteristic students with a non-European migration background.

In Figures 17-19 (2014) and Figure 23-25 (2019), the percentages of students with a non-European migration background are given for the various combinations of two of the three characteristics from the risk profile. These figures confirm the picture that emerges from the proxy characteristics of the individual characteristics. Below are some examples from 2014. The same picture applies for 2019.

In 2014, only 11.3% of 19-20-year-old wo students are students with a non-European migration background, while 65.8% of 23-24-year-old mbo 1-2 students belong to this migrant group (Figure 17). Also, the combination of the characteristics education type and distance to parent(s) shows that the combined characteristics largely show the expected proxy characteristic of the one-dimensional variant: of the wo students who live 50-500 kilometers from their parents, only 9.5% have a non-European migration background. For mbo 1-2 students who live 1 meter to 1 kilometer from their parent(s), 57.5% belong to this group (Figure 18). It is noteworthy, however, that the proportion of students with a non-European migration background among mbo 1-2 students changes little as the distance to parent(s) increases. There is little difference between these students who live close to their parents (57.5%) and students who live far from their parents (53.5%) and all distance categories in between. This proportion does shift as students live further away from their parental home in all other education types. When age is compared to distance to parent(s), a more fragmented picture emerges (Figure 19). Generally, the share of students with a non-European migration background is low among young students who live far from their parents (between 8.0-10.6%) and high among students who live close to their

parents (between 44.1-55.2%). For the distance category, the large share of students with a non-European migration background in the unknown category stands out. An explanation is provided for why this proportion is so high in the answer to Research question 2. The figures below show a similar picture for mbo students from the CUB population of 2019. For older mbo 1-2 students, the share of students with a non-European migration background is high (Figure 23). For mbo 1-2 students, the distance to parent(s) plays a limited role in indicating the share of students with a non-European migration background (Figure 24). Except for the unknown distance category, for which an explanation was provided in answering Research question 2. For the entire mbo student population it applies that older students who live further from the parental address are less likely to have a non-European migration background. It is precisely younger students who live close to their parents typically who are more likely to belong to the group of students with a non-European migration background (Figure 25).

Diagrams and numbers 2014

Education type

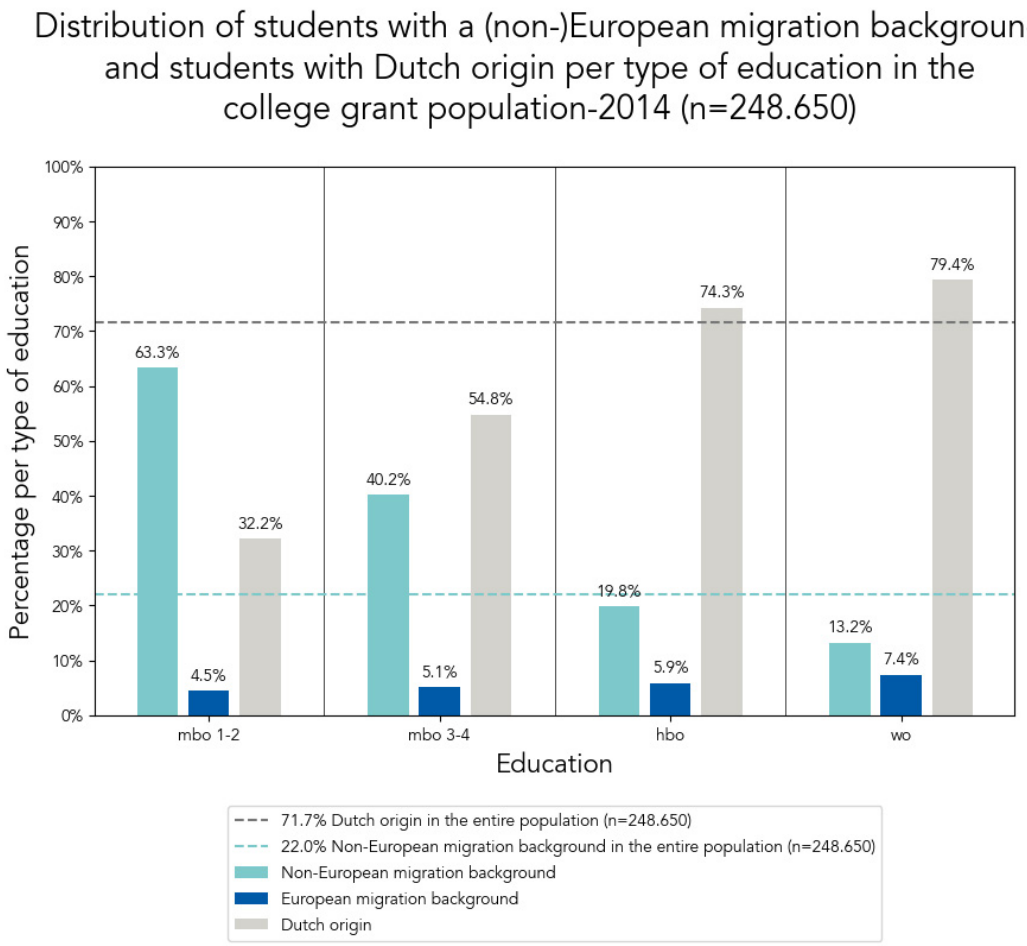

Figure 14 – Distribution of students with a non-European migration background and students with Dutch origin per education type in the college grant population-2014 (n = 248.650)

Age

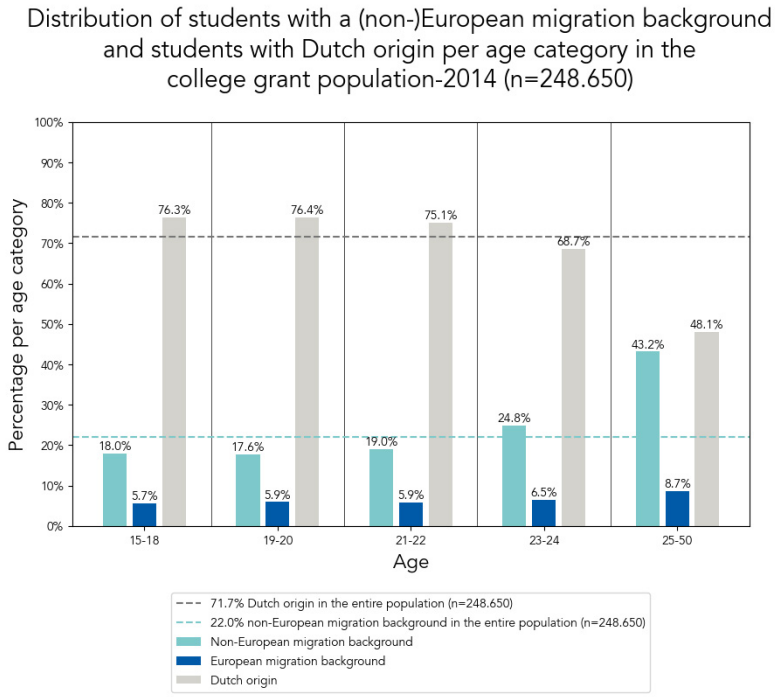

Figure 15 – Distribution of students with a non-European migration background and students with Dutch origin per age category in the college grant population-2014 (n = 248.650)

Distance to parent(s)

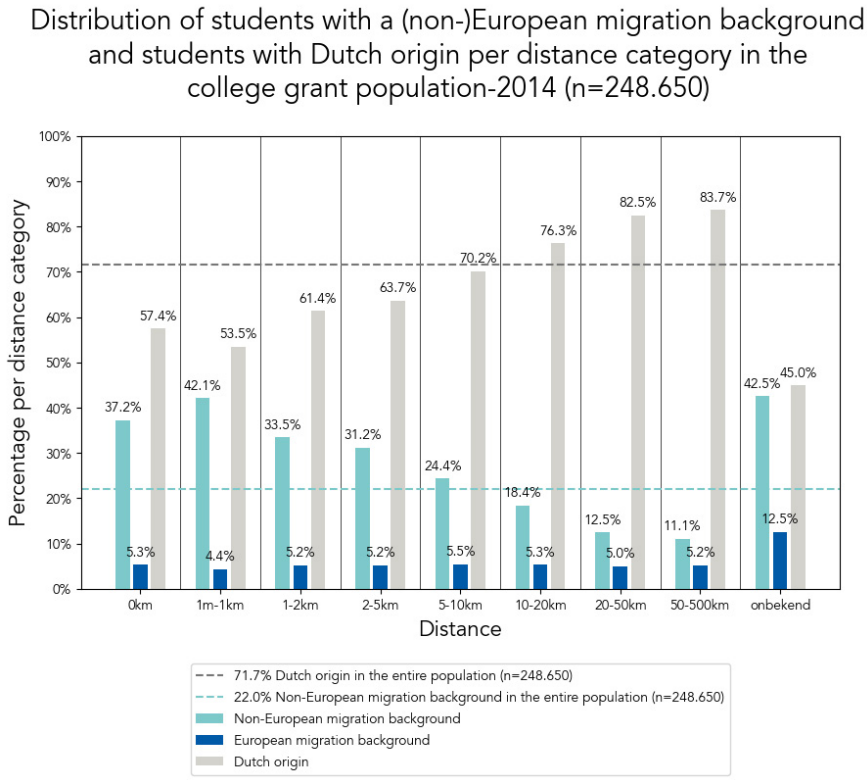

Figure 16 – Distribution of students with a non-European migration background and students with Dutch origin per distance category in the college grant population-2014 (n = 248.650)

Education type in combination with age

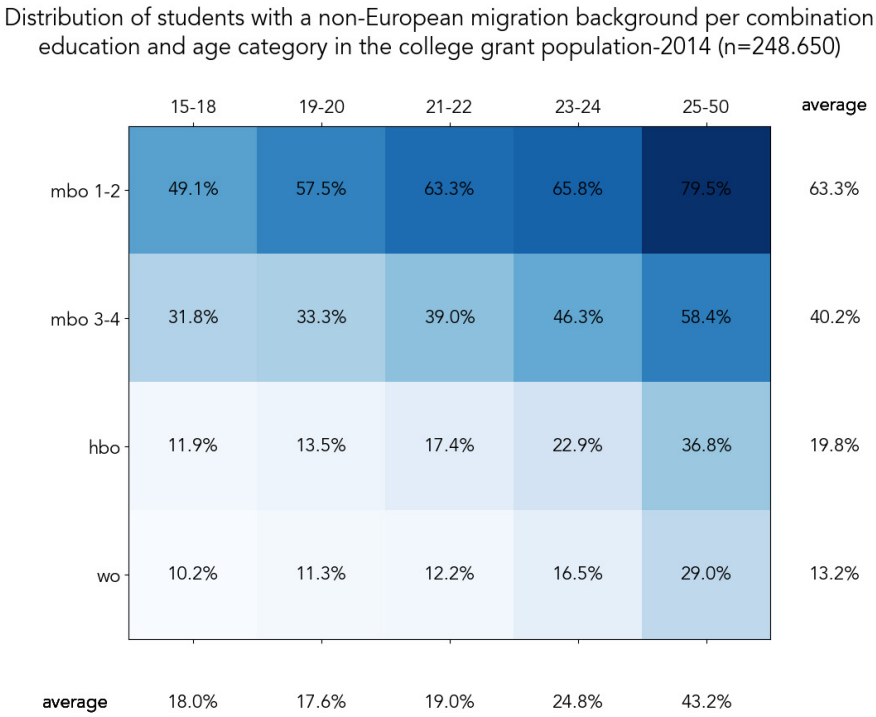

Figure 17 – Distribution of students with a non-European migration background per combination of education and age category in the college grant population-2014 (n=248.650)

Education type in combination with distance to parent(s)

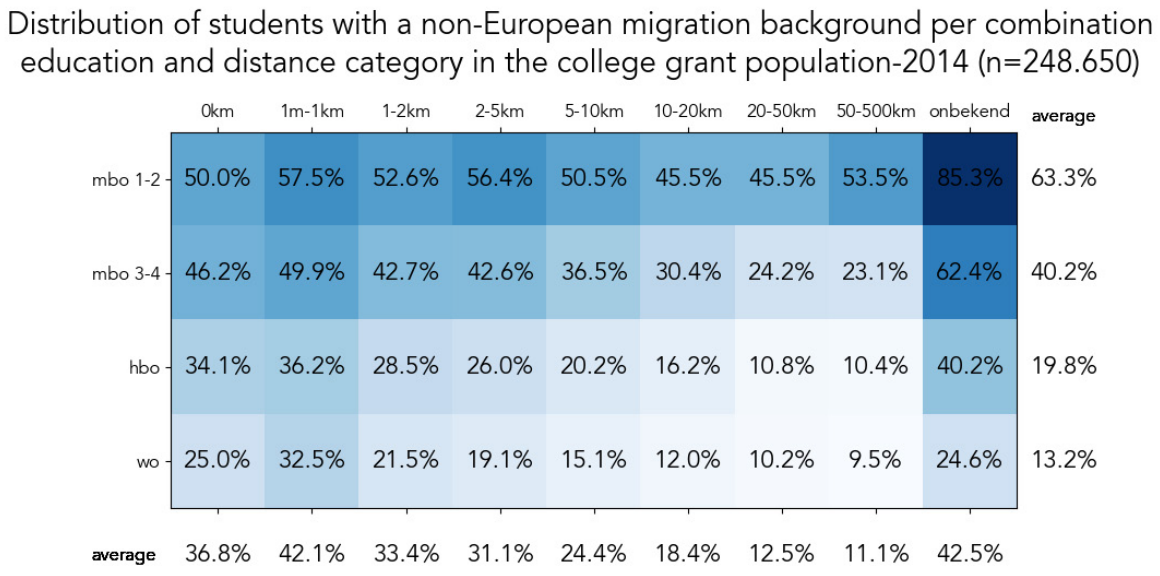

Figure 18 – Distribution of students with a non-European migration background per combination of education and distance category in the college grant population-2014 (n=248.650)

Age in combination with distance to parent(s)

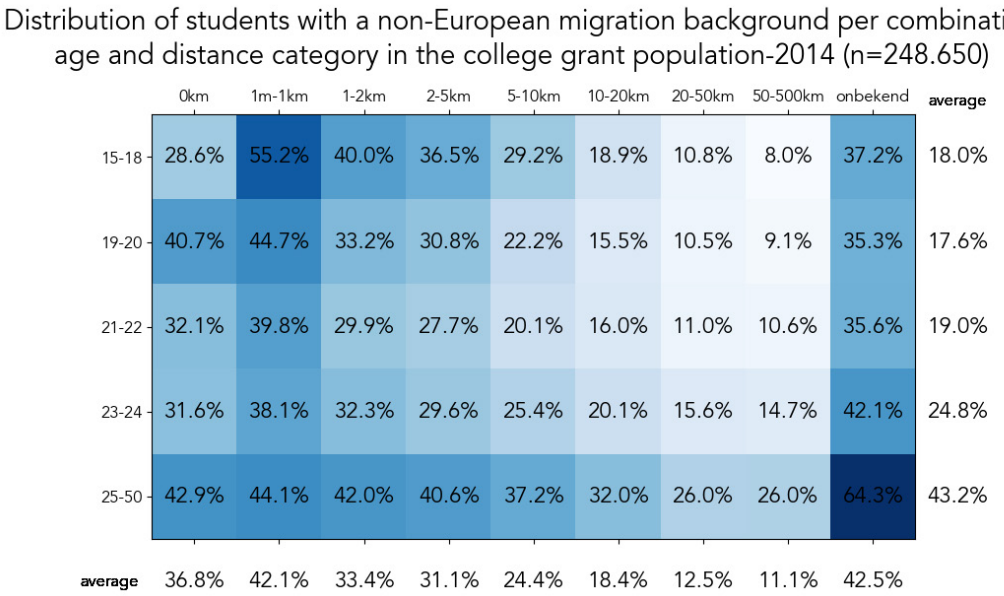

Figure 19 – Distribution of students with a non-European migration background per combination of age and distance category in the college grant population-2014 (n=248.650)

Diagrams and numbers 2019

Education type

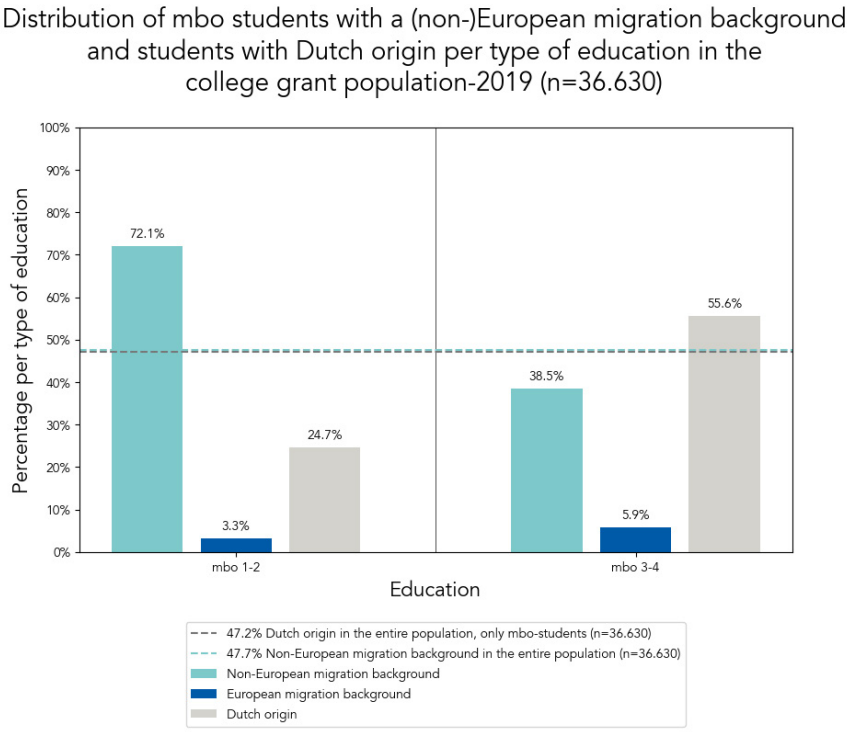

Figure 20 – Distribution of mbo-students with a non-European migration background and students with Dutch origin per education type in the college grant population-2019 (n = 36.630)

Age

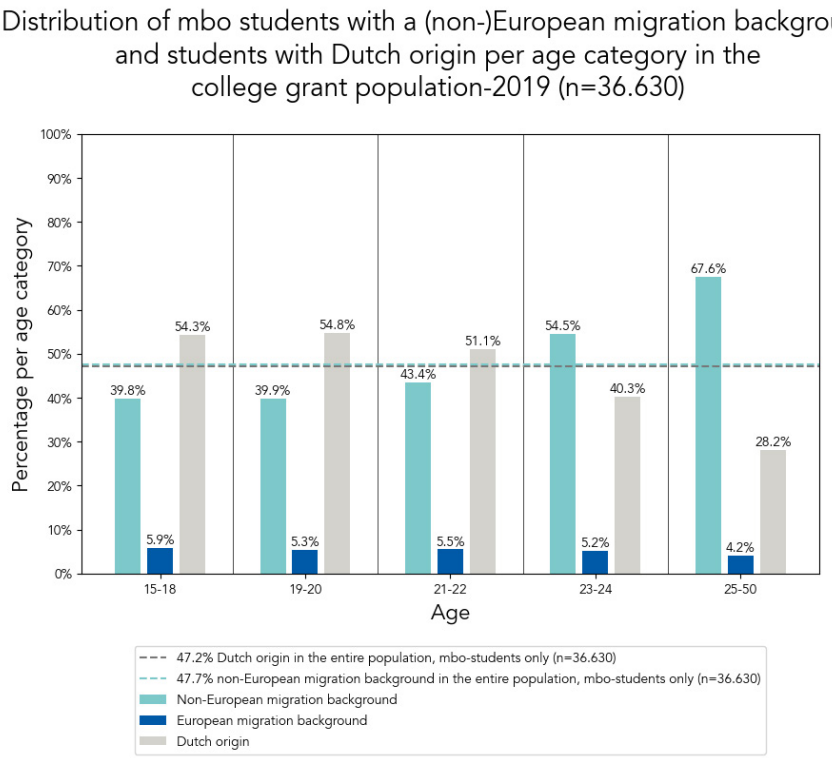

Figure 21 – Distribution of mbo-students with a non-European migration background and students with Dutch origin per age category in the college grant population-2019 (n = 36.630)

Distance to parent(s)

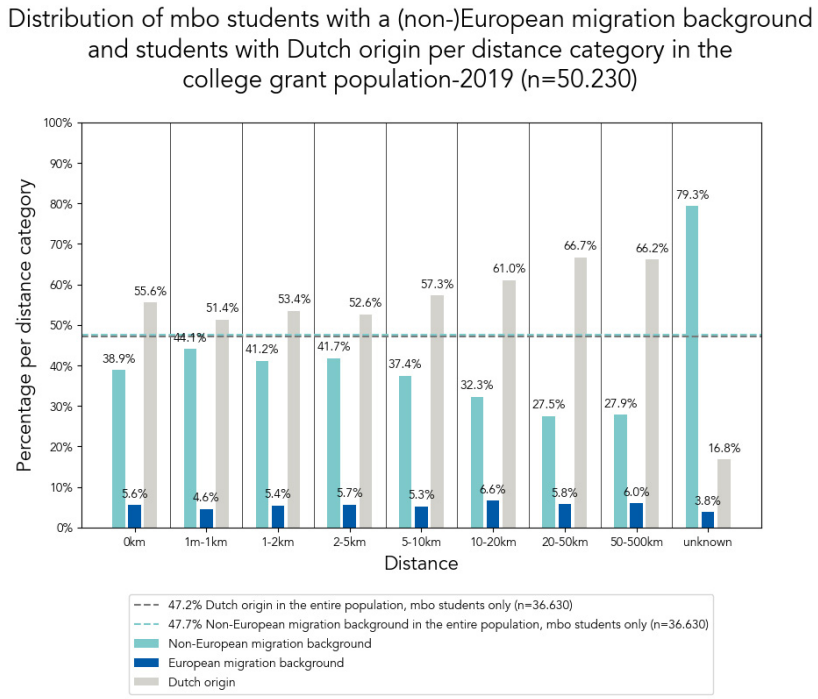

Figure 22 – Distribution of mbo-students with a non-European migration background and students with Dutch origin per distance category in the college grant population-2019 (n = 36.630)

Education type in combination with age

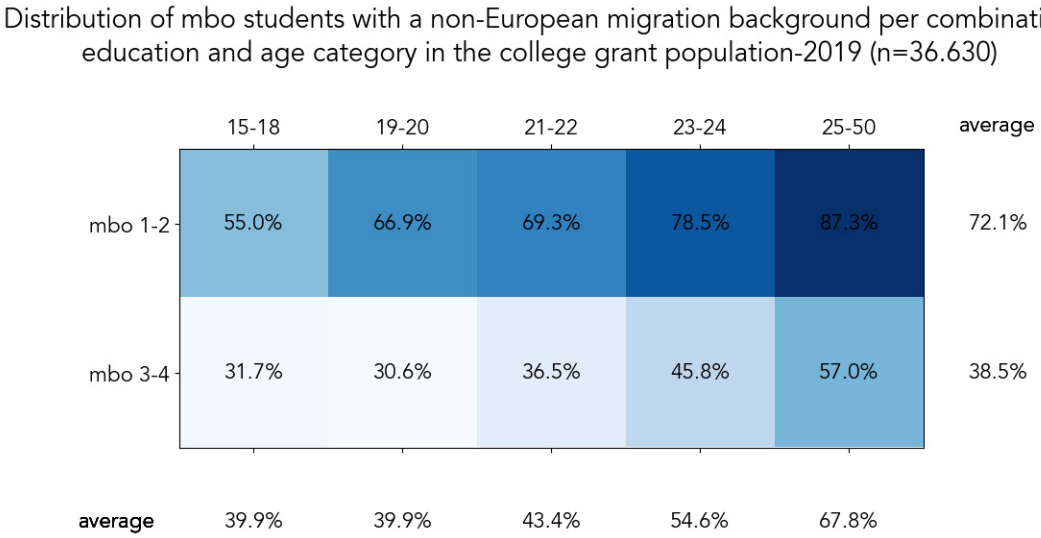

Figure 23 – Distribution of mbo students with a non-European migration background per combination of education and age category in the college grant population-2019 (n=36.630)

Education type in combination with distance to parent(s)

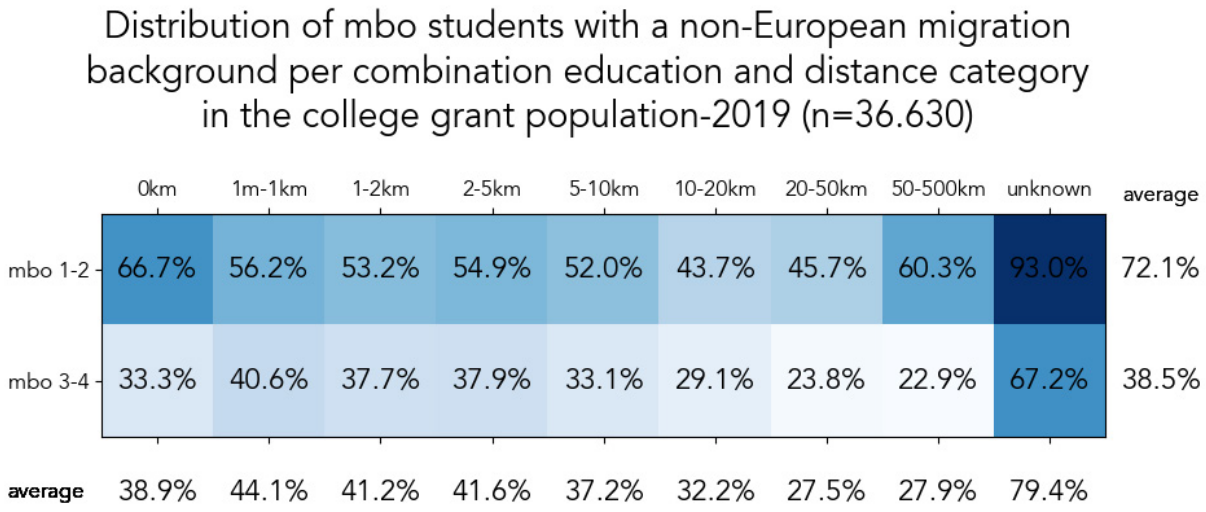

Figure 24 – Distribution of mbo students with a non-European migration background per combination of education and distance category in the college grant population-2019 (n=36.630)

Age in combination with distance to parent(s)

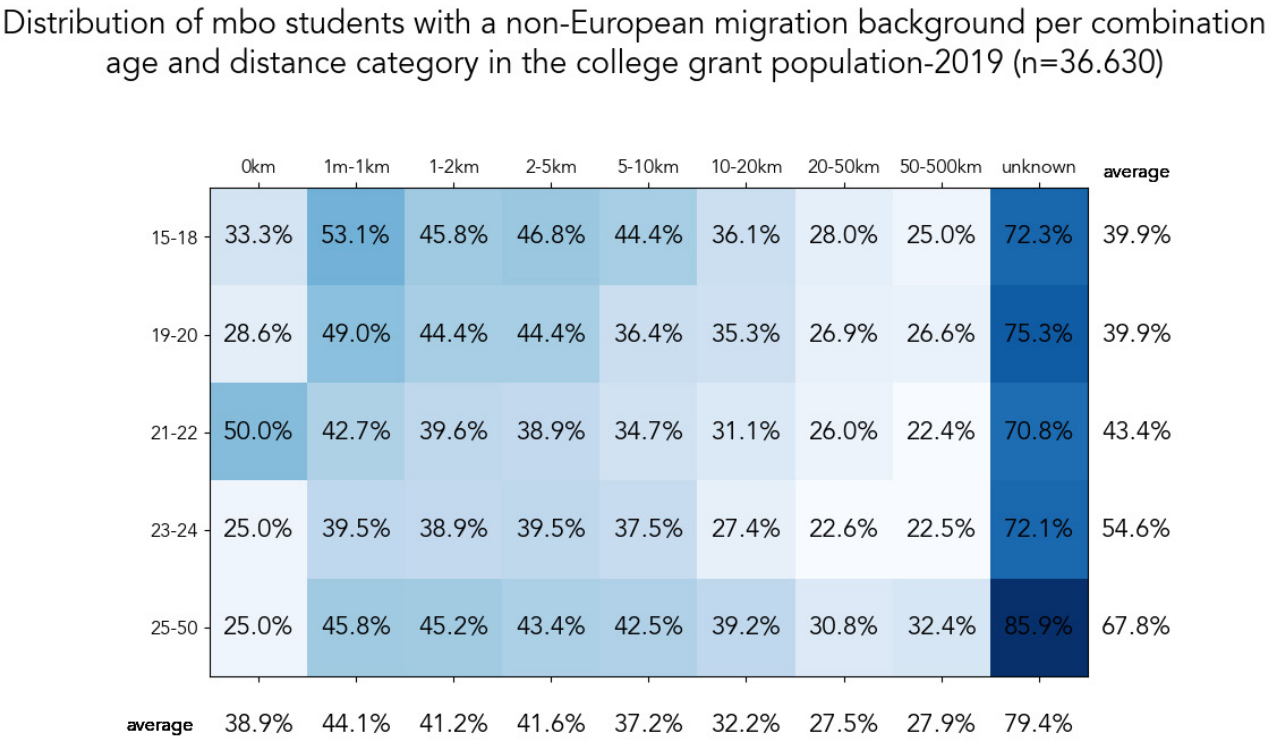

Figure 25 – Distribution of mbo students with a non-European migration background per combination of age and distance category in the college grant population-2019 (n=36.630)

4.4 Results of research question 3 (manual selection)

Research question 3

Were students with a migration background in 2014 and 2019 more often manually selected for a control procedure than students with Dutch origin?

Answer to research question 3

Yes. In 2014 and 2019, students with a non-European migration background were more often manually selected for checks than students with Dutch origin. In 2014, students with a non-European migration background were 6.2 times more likely to be manually selected for a check (step 3 of the CUB process) than students with Dutch origin. For 2019, this likelihood was 3.6 times higher.

Elaboration on research question 3

The results are explained based on the available CUB data for 2014 and 2019.

Figure 26 shows for the entire 2014 student population with a college grant which groups of students were manually selected for a check in step 3 of the CUB process. Of the 248.650 students who applied for a college grant in 2014, 2.810 students were selected for a check (1.1%). Of the 2.810 students selected for a check, 63.2% had a non-European migration background. This is a strong overrepresentation, both in relation to the initial population (see Research question 1) and in relation to the skewed ratios following from the application of the risk profile (see Research question 2). Of the 248.650 students not selected for a check, 21.5% had a non-European migration background.

In 2019, the distribution is similar.[24] Of the 740 students selected for a check, 69.1% had a non-European migration background (Figure 27). Of the 49.500 students not selected for a check, 38.4% had a non-European migration background. These figures are closer than in 2014 because a larger proportion of the entire college grant population had a non-European migration background due to the larger share of mbo students.

For both 2014 and 2019, the shifts in the proportion of students with a non-European migration background after manual selection for a check in step 3 of the CUB process are significant. Based on the 2014 CUB data, the proportion of students with a non-European migration background after step 1 of the CUB process (risk profile) shifts from 36.0% to 63.2%. For 2019, there is an increase from 37.7% to 69.1%.

2014

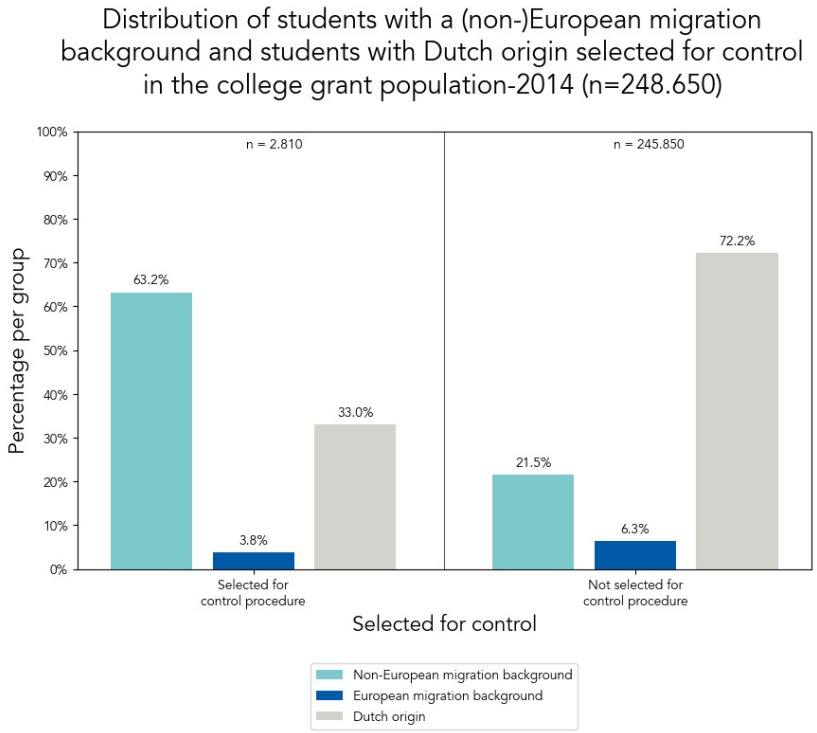

Figure 26 – Distribution of students with a (non-)European migration background and students with Dutch origin selected for control in the college grant population-2014 (n=248.650)

[24] For completeness, it is noted that these figures are for all students who received a college grant in 2019; not only the mbo students as in some other parts of this addendum, but also graduating wo and hbo students who were entitled to a college grant before the introduction of the loan-based system in 2015. Technically, mbo students could not easily be separated from the entire population for this step.

2019

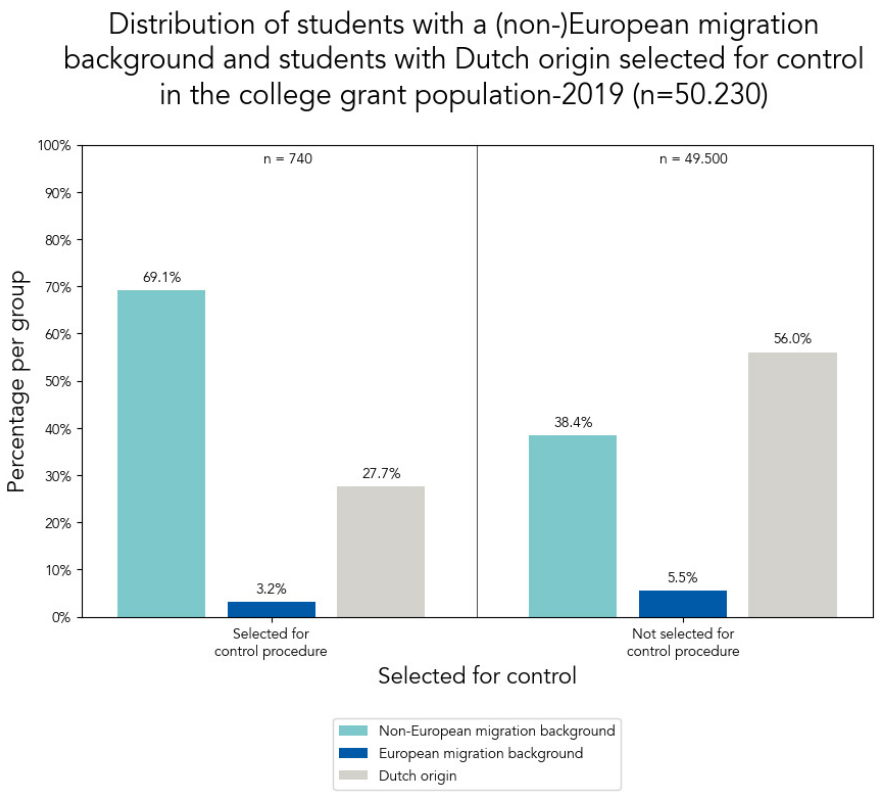

Figure 27 – Distribution of students with a (non-)European migration background and students with Dutch origin selected for control in the college grant population-2019 (n=50.230)

4.5 Results of research question 3a (extent to which algorithm prediction is followed)

Research question 3a

To what extent does the classification of the risk profile (high or low risk) align with manual selection for a control procedure, in particular for students with a migration background compared to students with Dutch origin?

Answer to research question 3a

Across all risk categories, students with a migration background are proportionally more likely to be selected for a home visit than students with Dutch origin. In 2014, students with a non-European migration background classified as high risk by the risk profile were 5.5 times more likely to be manually selected for a check than students with Dutch origin in the same risk category. Students with a non-European migration background classified as low risk by the risk profile were 1.8 times more likely to be manually selected for a check than students with Dutch origin in the same risk category. Students with a non-European migration background classified as unknown risk by the risk profile were 2.3 times more likely to be manually selected for a check than students with Dutch origin in the same risk category. For 2019, these figures were respectively 3.6 times (high risk), 4.2 times (low risk), and 1.0 times (unknown risk) higher than students with Dutch origin in the same risk category.

Elaboration on research question 3a

In 2014, 42.800 students were classified as high risk by the risk profile. Of those students, 36.0% had a non-European migration background. Not all students classified as high risk were selected for a check; 2.400 students were, of which 73.9% had a non-European migration background. 40.400 students classified as high risk were not selected for a check, of which 33.7% belonged to the non-European migration background group (Figure 28). Of the 171.810 students classified as low risk by the risk profile, 640 students were manually selected for a check and 171.180 students were not. Of these groups, respectively 25.0% and 15.3% belonged to the non-European migration background group (Figure 29). This is a considerably milder deviation than for the high-risk category. Of the 34.050 students classified as unknown risk by the risk profile, 140 students were manually selected for a check and 33.910 students were not. Of these groups, respectively 60.0% and 38.2% belonged to the non-European migration background group (Figure 30). This is a stronger deviation than observed for the low-risk category, but not as strong as the deviation observed for students with a non-European migration background classified as high risk by the risk profile.

For 2019, the deviation is also to the disadvantage of students with a non-European migration background during manual selection for both high and low-risk categories classified by the risk profile. Of the 14.850 students classified as high risk by the risk profile, 840 students were manually selected for a check and 14.010 students were not. Of these groups, respectively 67.9% and 35.9% belonged to the non-European migration background group (Figure 31). Of the 24.130 students classified as low risk by the risk profile, 60 students were manually selected for a check and 24.070 students were not. Of these groups, respectively 60.0% and 24.8% belonged to the non-European migration background group (Figure 32). The deviation for these groups is of the same magnitude. For students classified as unknown risk by the risk profile, a different picture emerges. Of the 11.260 students classified as unknown risk, 40 students were manually selected for a check and 11.220 students were not. Of these groups, respectively 75.0% and 70.7% belonged to the non-European migration background group (Figure 33).

By relating the above percentages to the average proportion of students with a non-European migration background per risk category, the increased odds for this group to be manually selected for a check can be determined. Since all these odds (as given above in the answer to the research question) are often significantly greater than 1.0x, it can be concluded that bias against students with a non-European migration background occurred during manual selection for checks.

The research did not find any evidence that this bias was due to individual officials' personal prejudices. Bias can also result from the nature of the work instructions, such as the exclusion of student housing from home visits, or other institutional causes. The question of how bias arose in the manual selection process falls outside the scope of this research.

2014

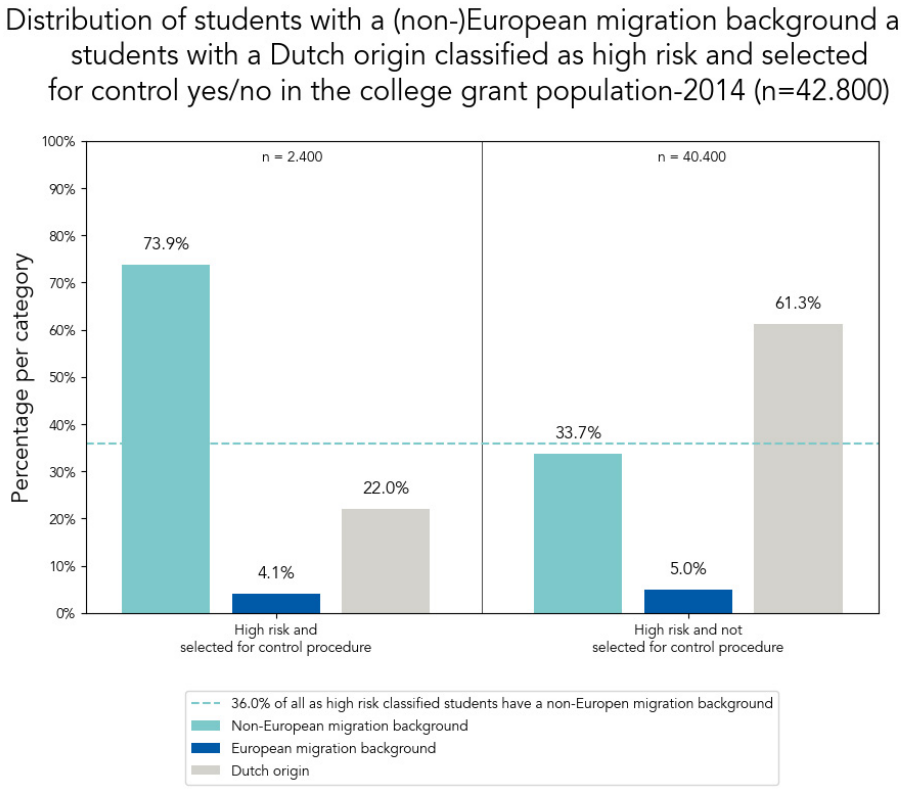

Figure 28 – Distribution of students with a (non-)European migration background and students with Dutch origin classified as high risk and selected for control yes/no in the college grant population-2014 (n=42.800)

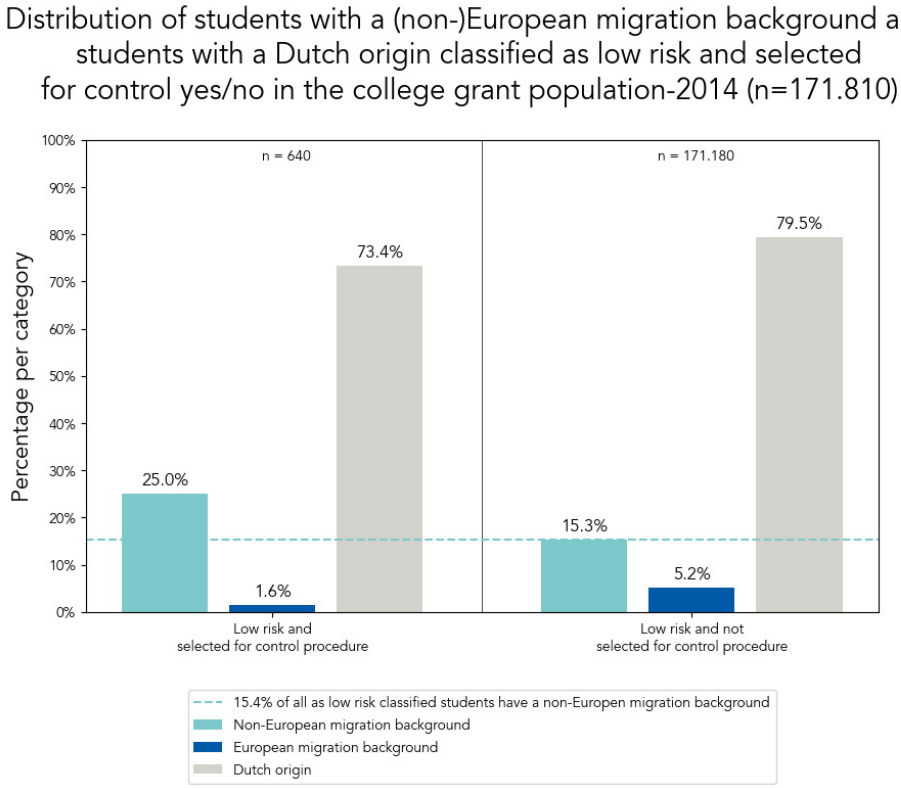

Figure 29 – Distribution of students with a (non-)European migration background and students with Dutch origin classified as low risk and selected for control yes/no in the college grant population-2014 (n=171.810)

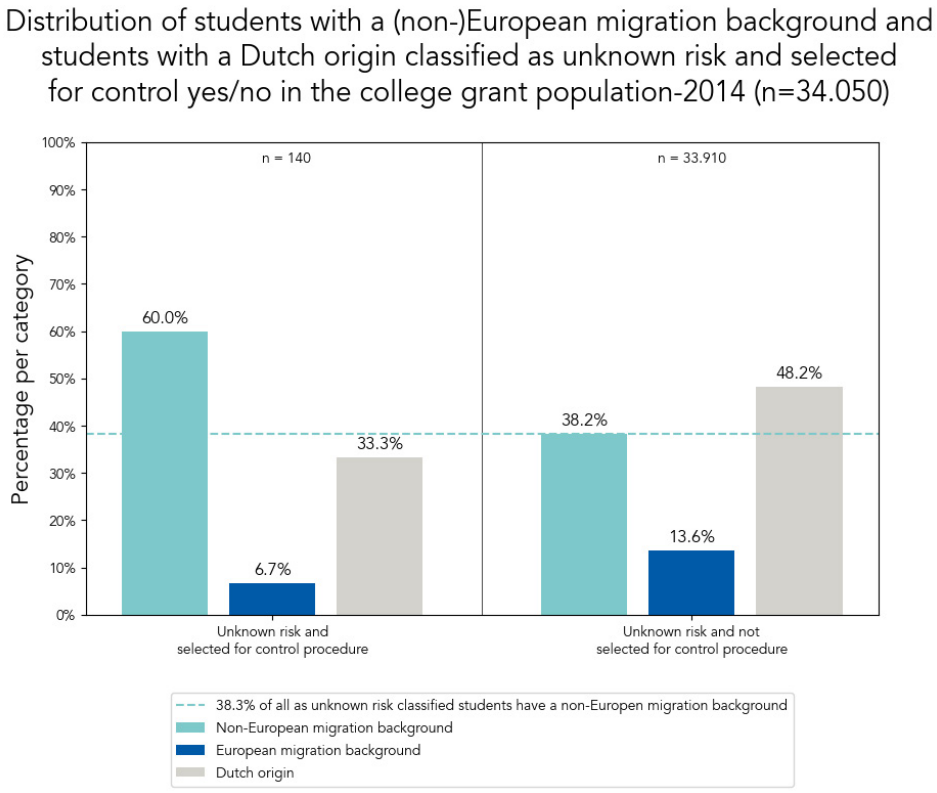

Figure 30 – Distribution of students with a (non-)European migration background and students with Dutch origin classified as unknown risk and selected for control yes/no in the college grant population-2014 (n=34.050)

2019

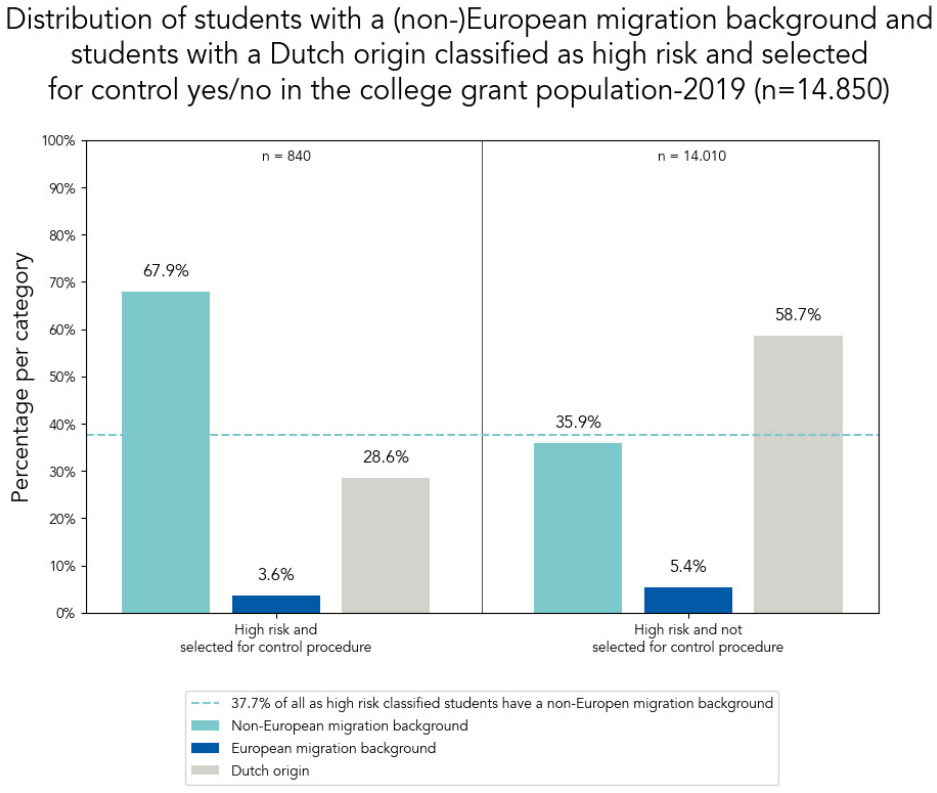

Figure 31 – Distribution of students with a (non-)European migration background and students with Dutch origin classified as high risk and selected for control yes/no in the college grant population-2019 (n=14.850)

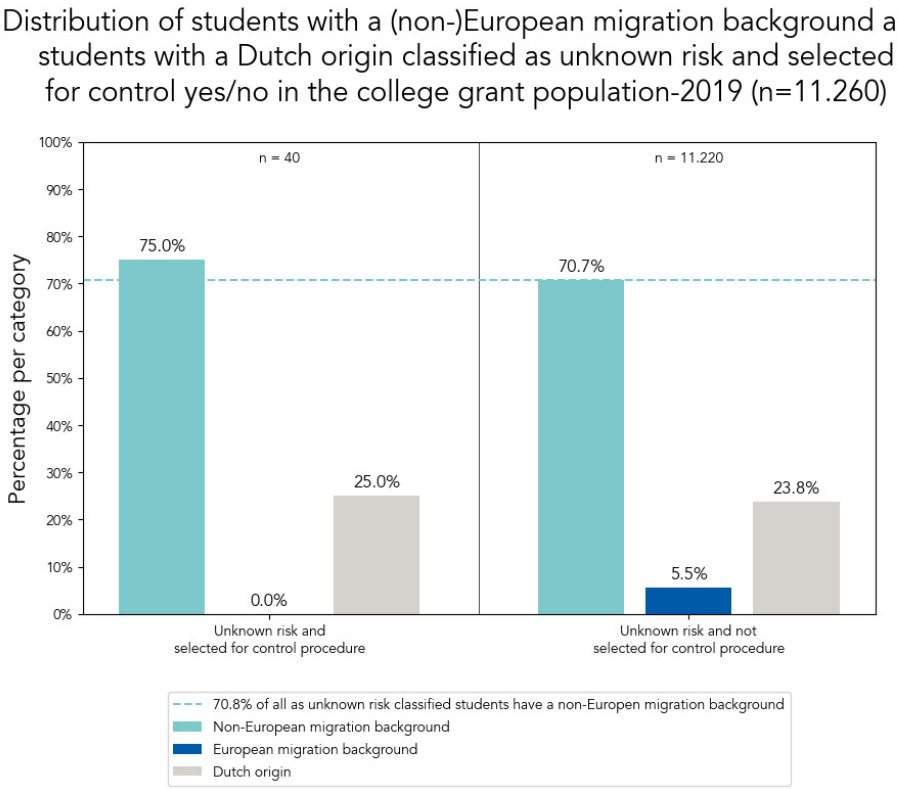

Figure 32 – Distribution of students with a (non-)European migration background and students with Dutch origin classified as low risk and selected for control yes/no in the college grant population-2019 (n=24.130)

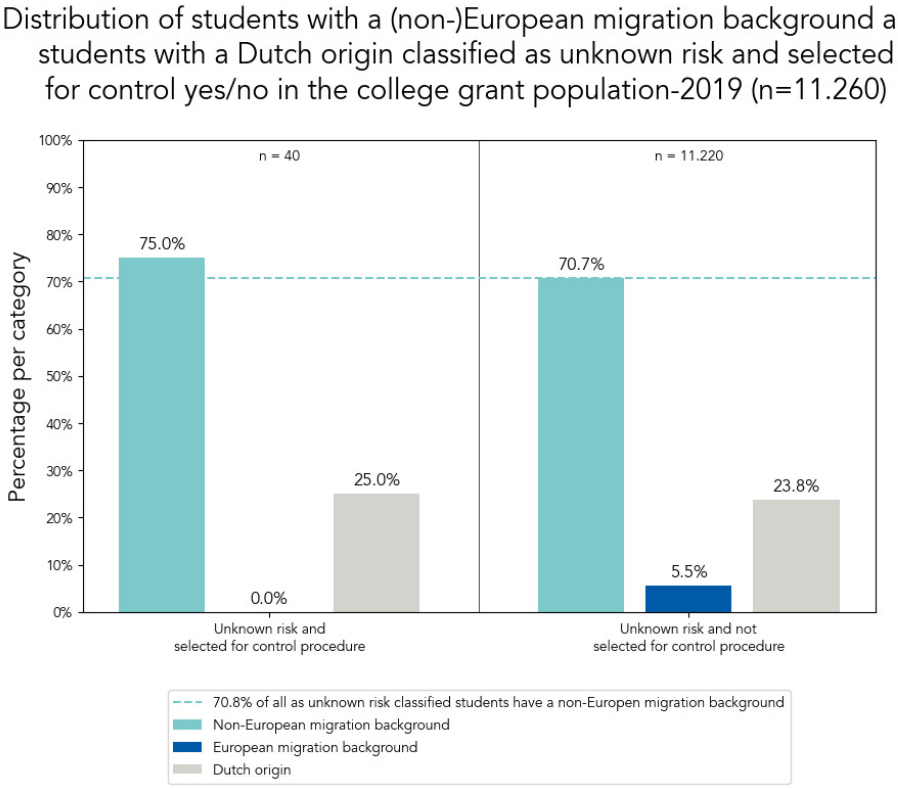

Figure 33 – Distribution of students with a (non-)European migration background and students with Dutch origin classified as unknown risk and selected for control yes/no in the college grant population-2019 (n=11.260)

4.6 Results of research question 4 (unduly use)

Research question 4

What was in 2014 and 2019 the distribution of students with a migration background in the group of students considered to have been receiving a grant unduly?

Answer to research question 4

Based on the college grant population-2014, 86.3% of students who according to DUO claimed the college grant unduly had a non-European migration background. Based on the college grant population-2019, this was 75.8%.

Based on the college grant population-2014, among students determined to duly claim the college grant after a home visit, 40.5% had a migration background. Based on the college grant population-2019, this proportion was 56.1%.

Elaboration on research question 4

Based on the available CUB data from 2014 and 2019 (Figure 34 and Figure 35), it can be concluded that a higher proportion of students with a non-European migration background were found by DUO to have unduly claimed the college grant after a home visit. However, this does not necessarily mean that students with a migration background are more likely to misuse the college grant. The data available do not allow for such a conclusion. Throughout the steps of the CUB process, there is a magnifying effect concerning the group with a non-European migration background. The potential higher unduly use among students with a non-European migration background cannot be isolated from the biases in the CUB process at step 1 (risk profile) and step 3 (manual selection). Additionally, the random sample size is too small to measure this independently of the CUB process. Whether there is further bias in the process of home visits itself, beyond the magnifying effect, could not be examined either.

In addition to the above findings, Figure 36 and Figure 37 provide the distribution of home visit outcomes per migrant group for 2014 and 2019. In 2014, 2.810 students received a home visit. Of these, it was determined that 1.570 students had duly claimed the college grant. Of this group, 45.5% had a non-European migration background. This indicates that students with a non-European migration background and Dutch origin did not have an equal probability of an unjustified home visit, as the ratios from the base population should be maintained (22% vs. 71.7%).[25] For the 1.240 students found to have unduly claimed the college grant, 86.3% had a non-European migration background.

In 2019, 740 students received a home visit. Based on these visits, it was determined that 410 students had duly claimed the college grant. Of this group, 59.1% had a non-European migration background. Of the 330 students who according to DUO unduly claimed the college grant, 75.8% had a migration background. Again, in 2019, students with a non-European migration background and Dutch origin did not have an equal probability of an unjustified home visit, as the ratios from the base population should have been maintained (22% vs. 71.7%).

[25] 'Unjustified home visit' is a working title for a home visit where later no unduly use was found

2014

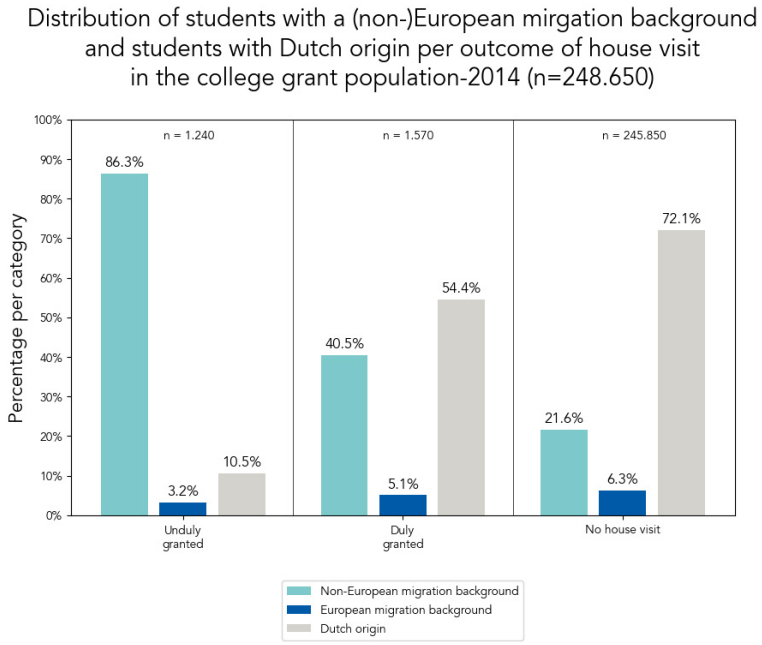

Figure 34 – Distribution of students with a (non-)European migration background and students with Dutch origin per outcome of house visit in the college grant population-2014 (n=248.650)

2019

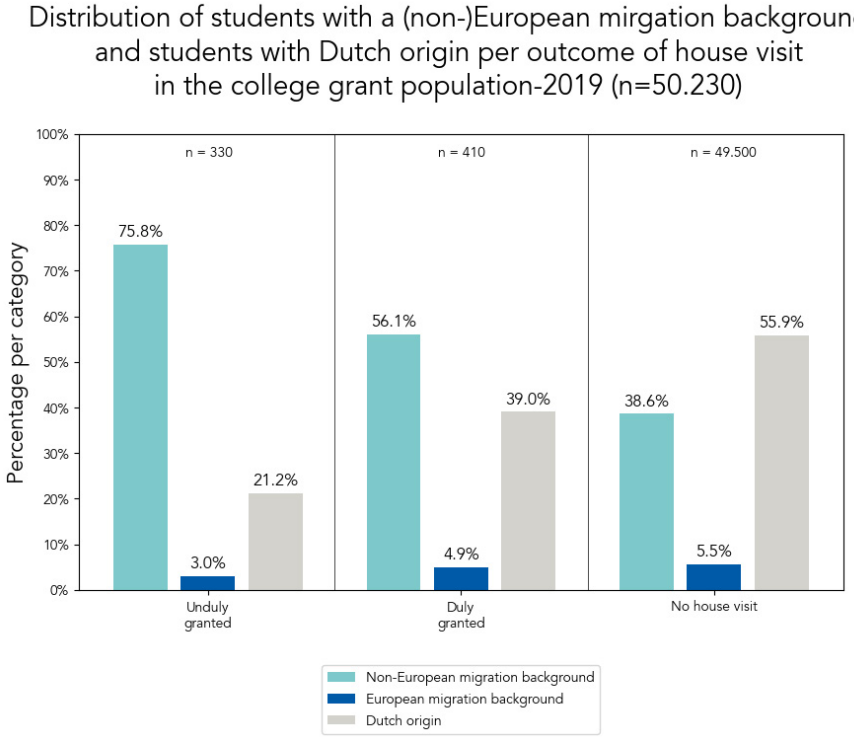

Figure 35 – Distribution of students with a (non-)European migration background and students with Dutch origin per outcome of house visit in the college grant population-2019 (n=50.230)

Which groups received a justified/unjustified home visit?

2014

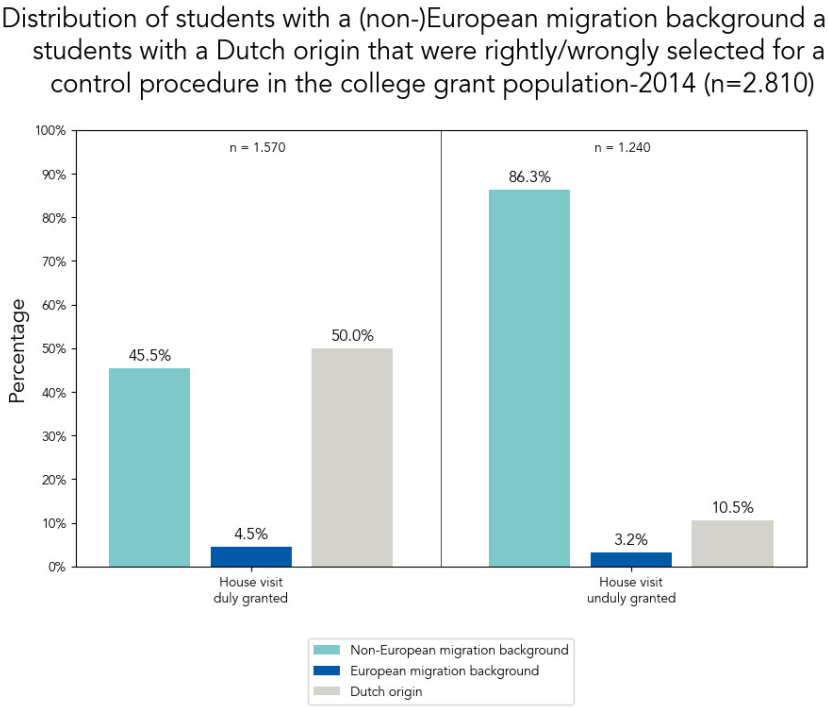

Figure 36 – Distribution of students with a (non-)European migration background and students with Dutch origin that were rightly/wrongly selected for a control procedure in the college grant population-2014 (n=2.810)

2019

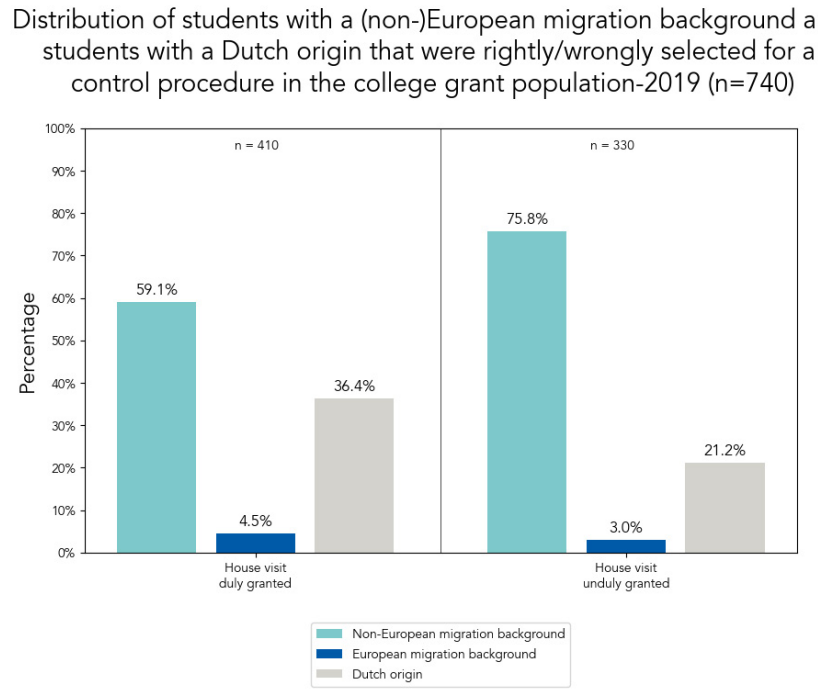

Figure 37 – Distribution of students with a (non-)European migration background and students with Dutch origin that were rightly/wrongly selected for a control procedure in the college grant population-2019 (n=740)

4.7 Results of research question 5 (appeal populations)

Research question 5

What was the distribution of students with a migration background who appealed DUO's decision in 2014, 2019, 2021 and 2022?

Answer to research question 5

For the years 2014, 2019, 2021, and 2022, respectively, 85.2%, 79.3%, 80.0%, and 82.6% of students in the appeal population belong to the group of students with a non-European migration background.

Elaboration on research question 5

The results are based on CBS data regarding the appeal population-2014, -2019, -2021, and -2022. The results are explained for each year.

From Figure 38 it follows that 85.2% of students in the appeal population-2014 belong to the group of students with a non-European migration background. From Figure 39, it follows that this 85.2% consists of 70.3% children of non-European migrants and 14.8% non-European migrants.
From Figure 40, it follows that 79.3% of students in the appeal population of 2019 belong to the group of students with a non-European migration background. From Figure 41, it follows that this 79.3% comprises 62.1% children of non-European migrants and 17.2% non-European migrants.

From Figure 42, it follows that 80.0% of students in the appeal population of 2021 belong to the group of students with a non-European migration background. From Figure 43, it follows that this 80.0% comprises 65.0% children of non-European migrants and 15.0% non-European migrants.

From Figure 44, it follows that 82.6% of students in the appeal population of 2022 belong to the group of students with a non-European migration background. From Figure 45, it follows that this 82.6% comprises 60.9% children of non-European migrants and 21.7% non-European migrants.

The appeal populations thus contain a disproportionately high share of children of non-European migrants compared to the initial population (22.0% of students with a non-European migration background), which can be explained by bias towards this group in various steps of the CUB process.

2014

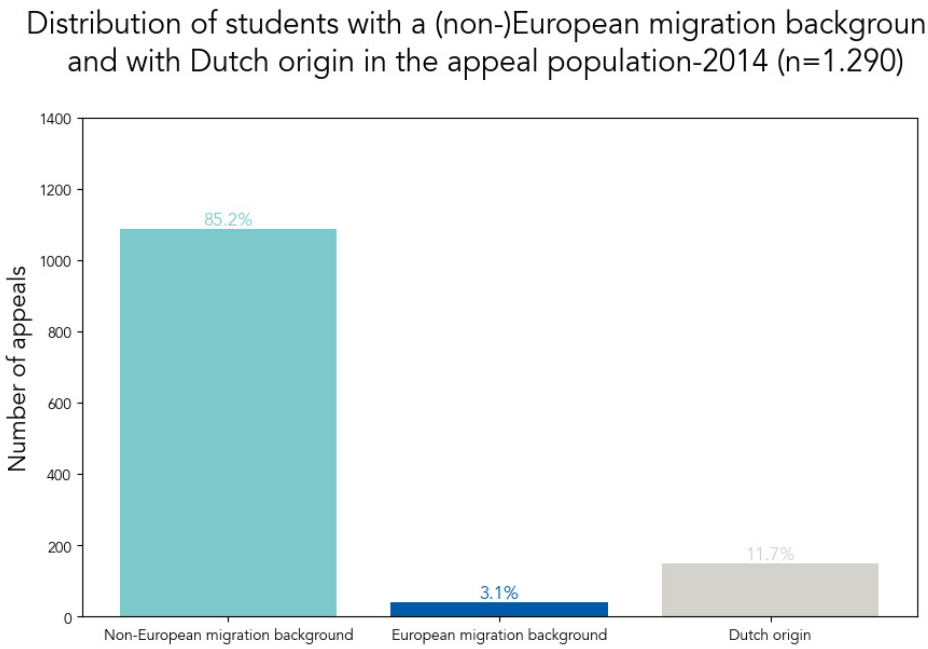

Figure 38 – Distribution of students with a (non)-European migration background and with Dutch origin in the appeal population-2014 (n=1.290)

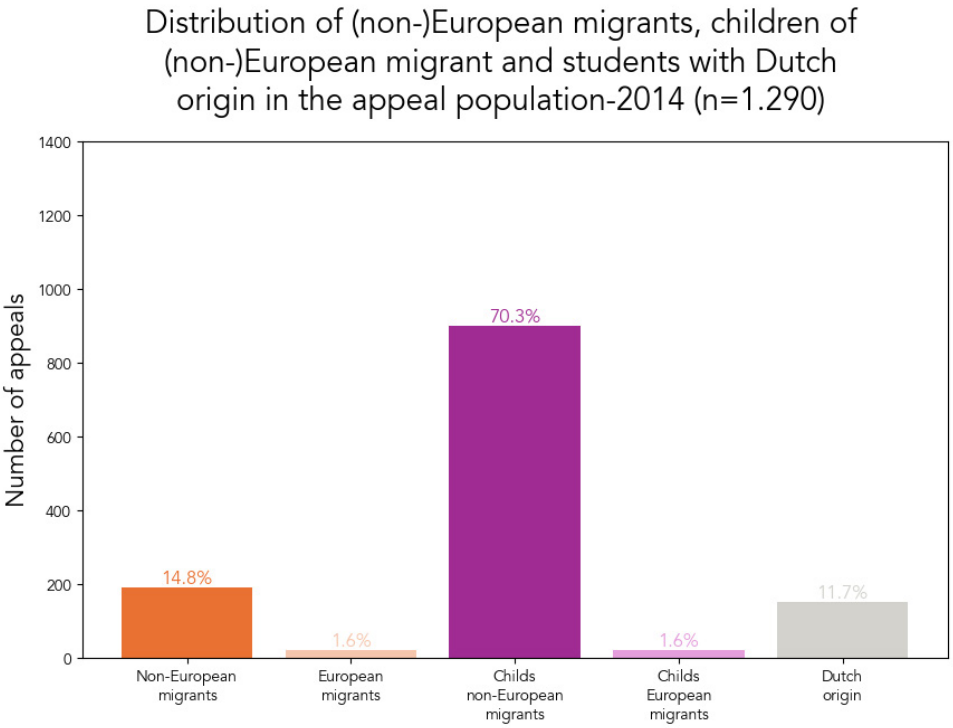

Figure 39 – Distribution of (non-)European migrants, children of (non-)European migrants, and students with Dutch origin in the appeal population-2014 (n=1.290)

2019

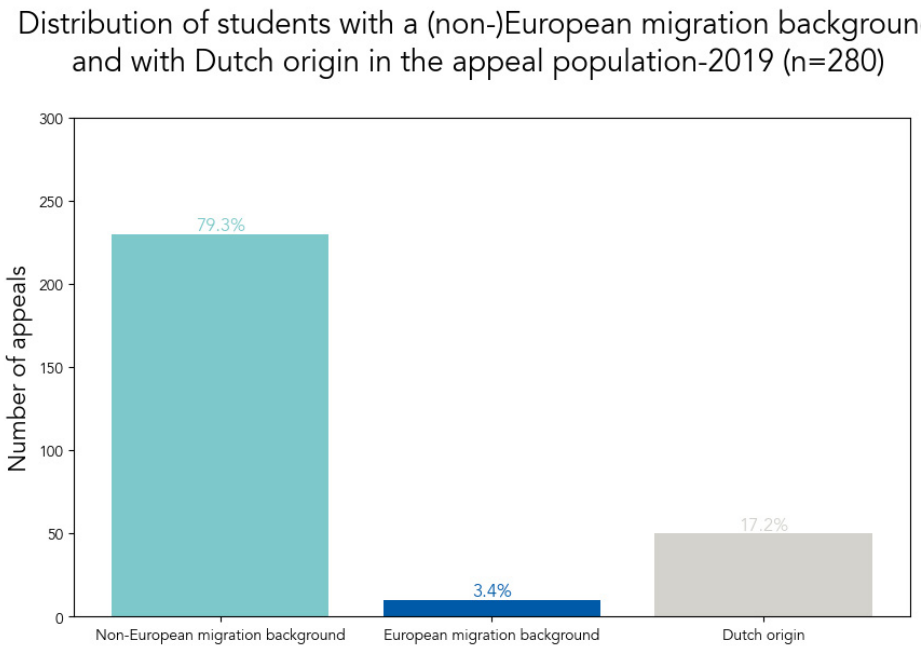

Figure 40 – Distribution of students with a (non)-European migration background and with Dutch origin in the appeal population-2019 (n=280)

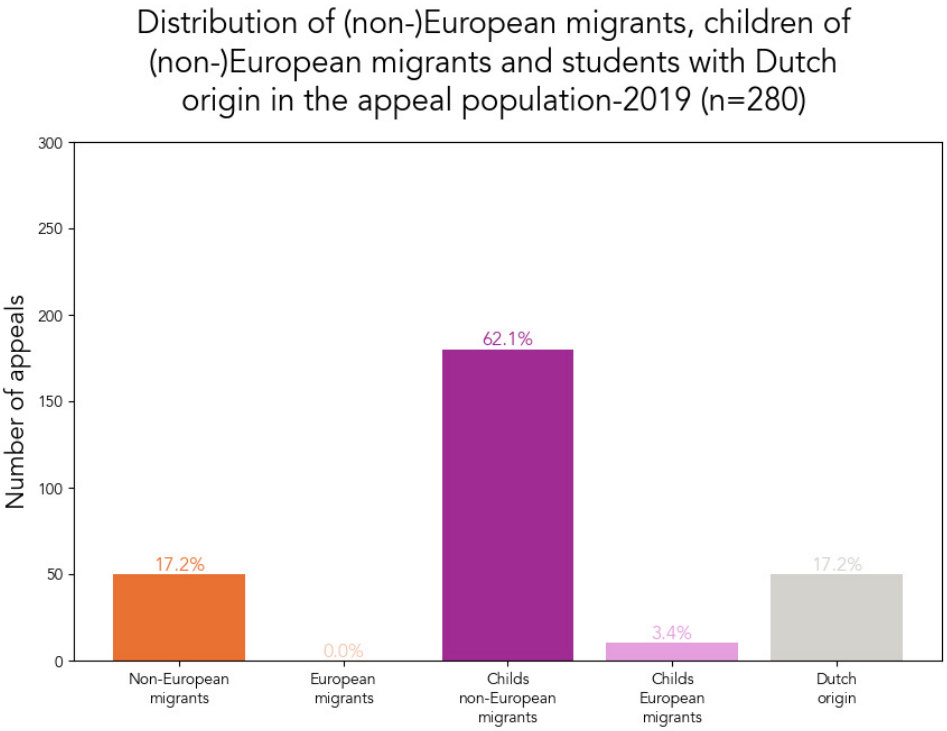

Figure 41 – Distribution of (non-)European migrants, children of (non-)European migrants, and students with Dutch origin in the appeal population-2019 (n=280)

2021

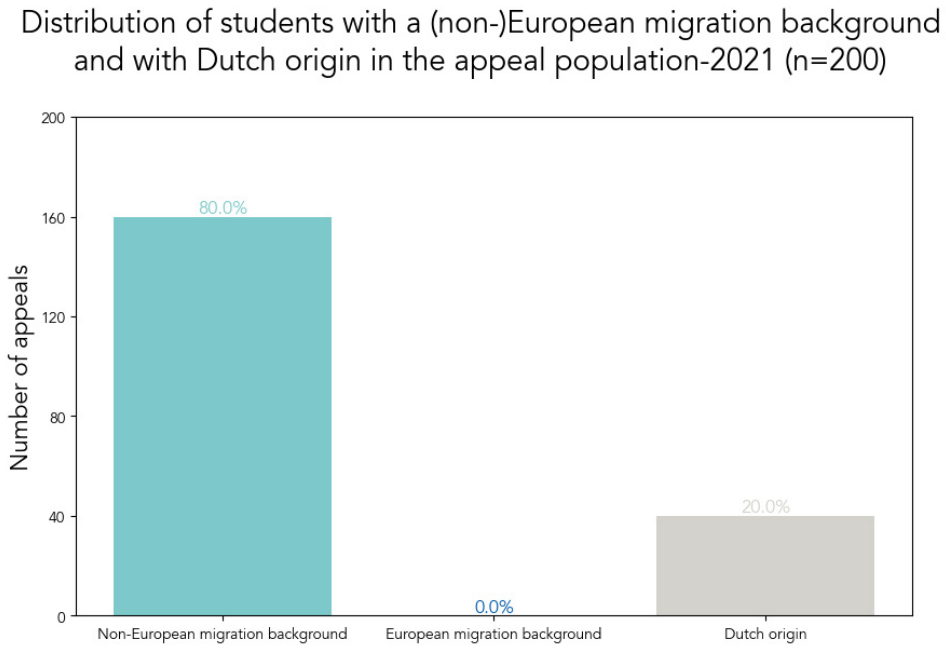

Figure 42 – Distribution of students with a (non)-European migration background and with Dutch origin in the appeal population-2021 (n=200)

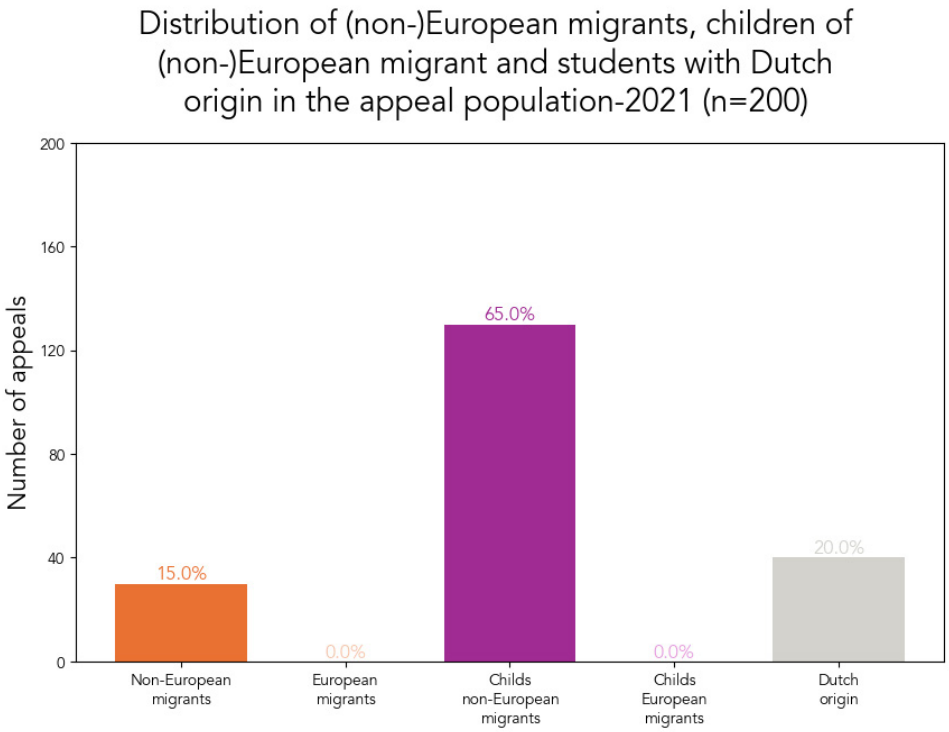

Figure 43 – Distribution of (non-)European migrants, children of (non-)European migrants, and students with Dutch origin in the appeal population-2021 (n=200)

2022

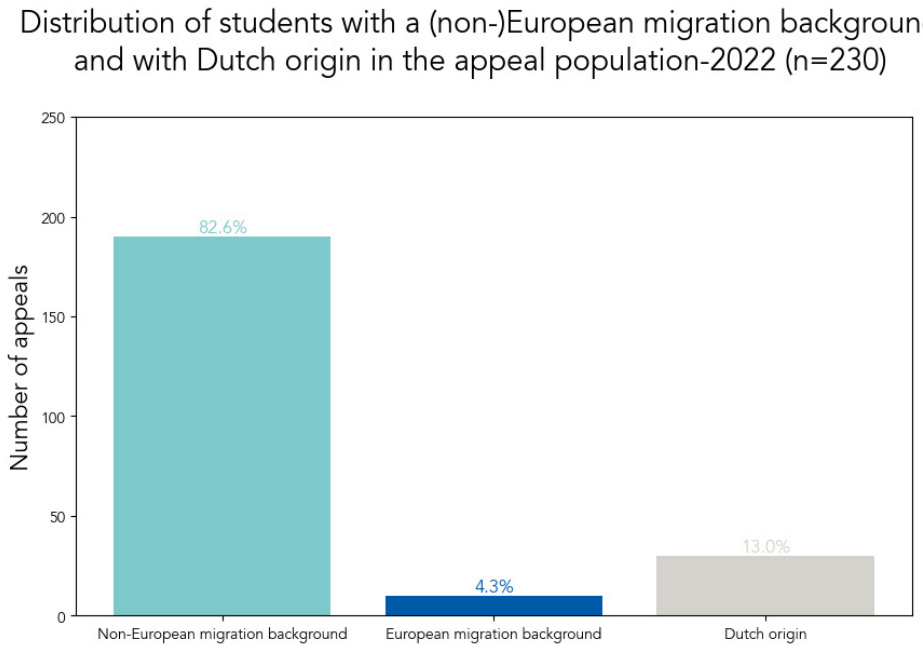

Figure 44 – Distribution of students with a (non)-European migration background and with Dutch origin in the appeal population-2022 (n=230)

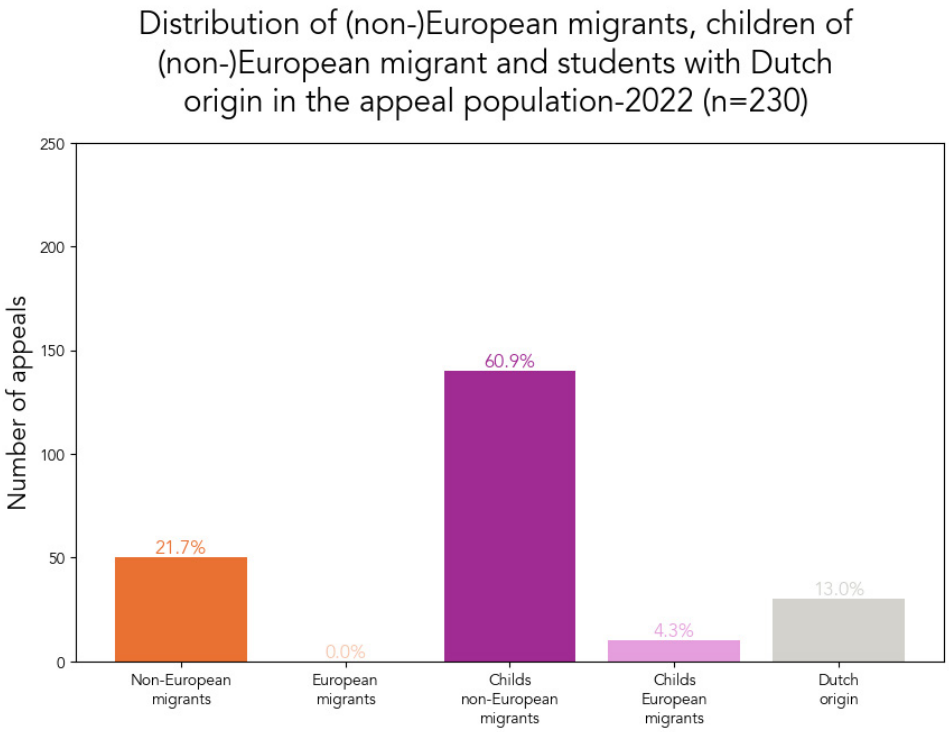

Figure 45 – Distribution of (non-)European migrants, children of (non-)European migrants, and students with Dutch origin in the appeal population-2022 (n=230)

4.8 Overview of the CUB process as a whole

In Figure 46 and Figure 47, the above results for 2014 and 2019 are summarized in a single figure (funnel visualization). This funnel illustrates the ratios of students with a migration background (European and non-European) compared to students with Dutch origin throughout the entire CUB process. The specific numbers are discussed in the above sections. Findings regarding the CUB process as a whole are shared in 6. Conclusion.

2014

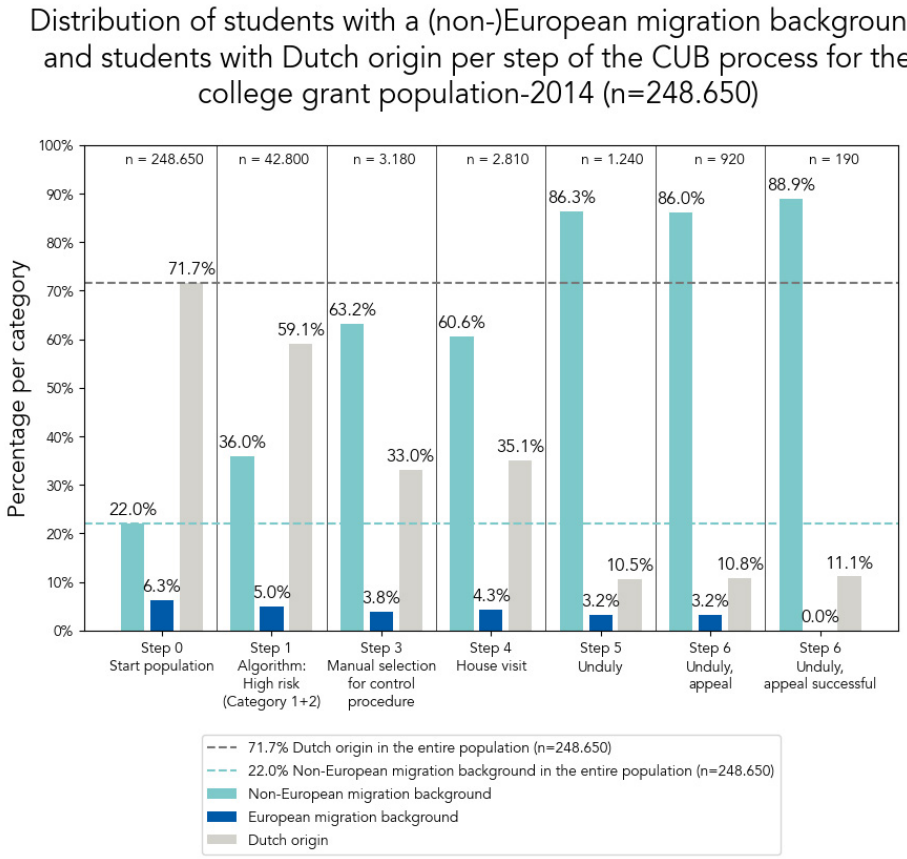

Figure 46 – Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)

2019

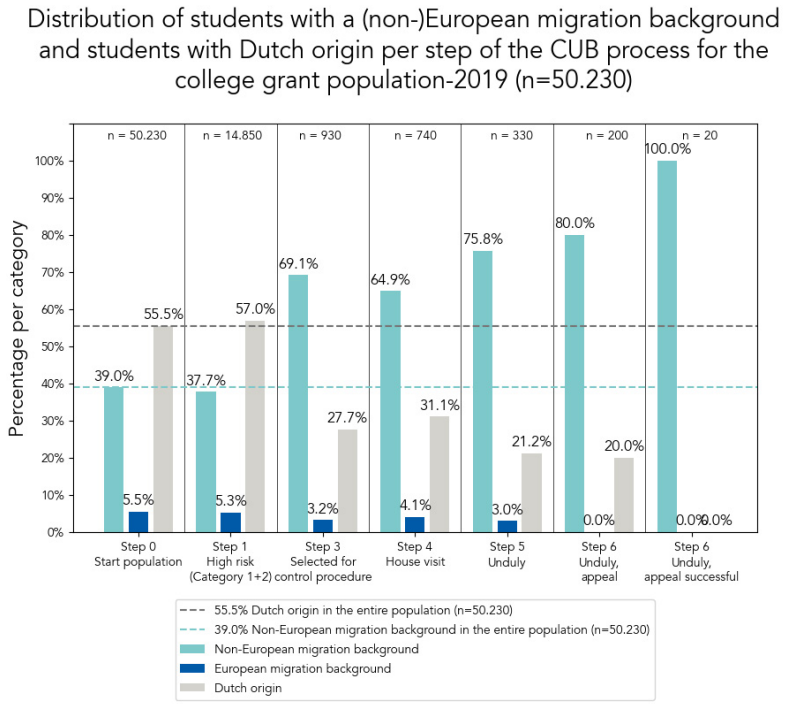

Figure 47 – Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2019 (n=50.230)

5. Disclaimers

Several disclaimers apply.

The customized tables from the CBS analyzed in this report are based on data provided by DUO to the CBS. Quality controls have been conducted on the accuracy of the data. However, Algorithm Audit cannot guarantee that all associated *queries* and/or underlying data structures at DUO and the CBS are complete and entirely free from errors or imperfections.

Algorithm Audit is not responsible for decisions made based on this report.

In drafting this report, the protection of personal data has been taken into account.

> The internal processes of Algorithm Audit comply with the AVG;
> DUO shared its documents via its own platform.

Algorithm Audit has never had access to documents originating from DUO. Upon the removal of access to DUO's systems, Algorithm Audit definitively no longer has access to the documents.

6. Conclusion

Five findings emerge from the bias analysis outlined above.

Finding 1 – The CUB process as a whole has been biased towards students with a non-European migration background.

In the CUB process, there was a strong bias towards students with a non-European migration background. Students with a non-European migration background were classified as high risk by the risk profile 2 times more often than students with Dutch origin. The group was manually selected for a home visit 6.2x more often. Ultimately, students with a non-European migration background had a 3.0x greater probability of receiving an unjustified home visit than students with Dutch origin.

Taking into account the substantially different populations in different years (largely due to the introduction of the loan-based system[26]), there is a clear structural trend in bias towards students with a non-European migration background.

Figure 48 provides an overview of the distribution of students from different origins per step of the CUB process.

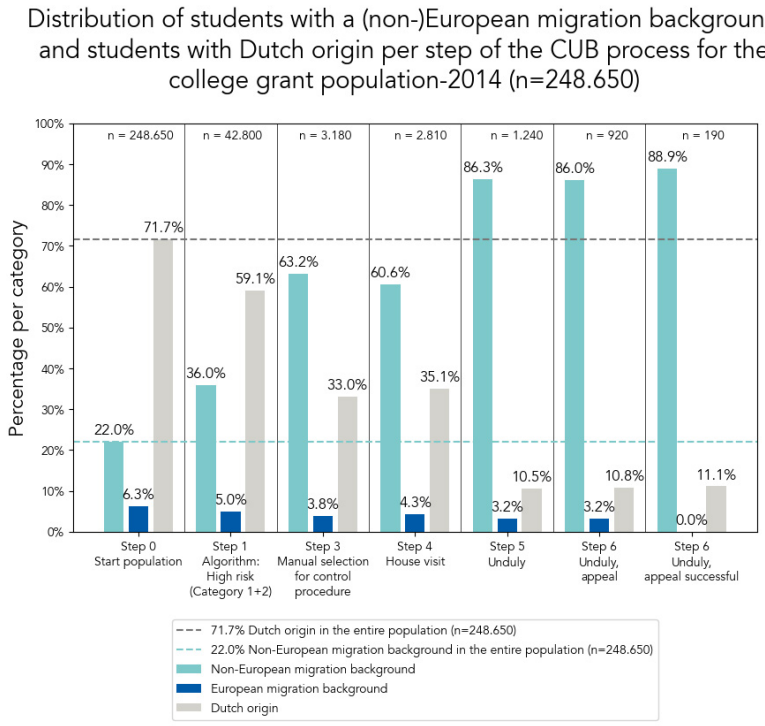

Figure 48 – Distribution of students with (non-)European migration background and students with Dutch origin per step of the CUB process in the college grant population-2014 (n=248.650)[27]

[26] In the period 2015-2023, only students enrolling in vocational education (mbo) were entitled to a college grant. Further details can be found in 3.2 Research populations.

[27] Since the aggregation statistics compiled by CBS are rounded to the nearest ten, percentages may not add up exactly to 100%. The same applies to the sum of counts.

Finding 2 – The risk profile used in the CUB process was biased towards students with a non-European migration background. The reason for this was the assignment of a higher risk to mbo students and to students who were registered close to their parental address.

In the utilized risk profile (which utilizes characteristics such as education type, age, and distance between the registered address and the parental address), bias occurred towards migration background. Students with a non-European migration background were 2.0 times more often classified as high risk by the risk profile than students with Dutch origin. This can be attributed to the proxy characteristics, particularly education type and distance to parent(s). Characteristics such as mbo 1-2 and a short distance to parent(s) correlate strongly with the group of students with a non-European migration background. Profiling based on these characteristics resulted in bias towards this demographic. An *unsupervised bias detection method* confirms these results.

The proxy analysis is further explained in 4.2 Results of research question 2 (risk profile) and 4.3 Results of research question 2a (proxy analysis). The results of the *unsupervised bias detection method* are explained in Appendix A – Unsupervised bias detection.

Finding 3 – Manual selection reinforced the bias of the CUB process.

After classification into risk categories by the risk profile, students were manually selected for inspection and home visits. In that manual selection process, written research, work instructions, certain exclusion criteria, and other procedural aspects played a role. The bias introduced by the risk profile is reinforced in this step. The study did not reveal any indications that this bias is due to personal prejudices of individual officials. Bias can also arise from the nature of work instructions, such as the exclusion of student housing from home visits and care facilities. Selection criteria within the group selected for conducting a home visit included factors such as the ratio between area and number of residents at the registered address, registration with family, and the distance from the registration to the parents' address. The question of how the bias in manual selection arose is beyond the scope of this study.

Students with a non-European migration background classified as high risk by the risk profile were 5.5 times more likely to be selected for inspection than students with Dutch origin in the same risk category. Students with a non-European migration background who were classified as low risk by the risk profile were 1.8 times more likely to be selected for inspection than students without a migration background in the same risk category. It should be noted that the assigned risk scores themselves are already biased.

The analysis of the manual selection of students is further explained in 4.4 Results of research question 3 (manual selection) and 4.5 Results of research question 3a (extent to which algorithm prediction is followed).

Finding 4 – Due to the bias in the CUB process towards students with a non-European migration background, a considerable amount of unduly use has been identified in this group. This is largely attributed to excessive scrutiny of this demographic.

Due to the bias of the entire CUB process, students with a non-European migration background were relatively more often selected for a home visit than students with Dutch origin. In the case of this group, this selection was also more often unjustified: students with a non-European migration background were 3.0 times more likely to receive a home visit that later did not reveal unduly use than students with Dutch origin.

Even among the group of students where unduly use is eventually determined, the proportion of students with a non-European migration background is large. Throughout the steps of the CUB process, a magnifying effect emerges concerning the group with a non-European migration background. Students with Dutch origin were relatively less frequently inspected, and unduly use is less frequently detected in this group.

Whether students with a migration background also make unduly use of the college grant more frequently cannot be determined based on the available data. This is because this ratio cannot be isolated from the bias of the CUB process, and because the random sample is too small to measure this independently of the CUB process. Additionally, it could not be investigated whether there is any bias in the home visit process itself beyond the mentioned magnifying effect.

The overrepresentation of this group of students in the unduly use population is further explained in 4.6 Results of research question 4 (unduly use).

Finding 5 – The group of students who appeal to a determination of unduly use consists largely of students with a non-European migration background. No bias has been identified in the appeal process itself.

From the various reference years (2014, 2019, 2021, and 2022), a consistent picture emerges that the appeal population consists of 79-85% students with a non-European migration background. This broadly confirms the image presented in investigative journalism regarding the strong overrepresentation of students with a migration background who appeal to a decision by DUO.[28]

On the other hand, no bias is identified in the appeal step of the CUB process. Appeals are equally likely to be successful for all students regardless of their origin.

The overrepresentation of this group of students in the appeal populations and equal treatment during the appeal process is further explained in 4.7 Results of research question 5 (appeal populations).

[28] B. Belleman, B. Heilbron & A. Kootstra. De discriminerende fraudecontroles van Duo. Investico Onderzoeksjournalisten, 2023

Appendix A – Unsupervised bias detection

The bias analysis in this report is based on aggregate statistics about the migration background of groups of students as provided by CBS. This type of bias analysis is referred to as supervised bias detection. 'Supervised' refers to the fact that access is available to special personal data (albeit at the aggregate level). However, for many organizations, such as businesses, it is impossible to measure supervised bias because special personal data (ethnicity, migration background, etc.) is often not available to these organizations under the General Data Protection Regulation (GDPR). However, bias also occurs in algorithms and data in private organizations, but how do you measure it?

In this appendix, we briefly study an alternative when access to special personal data would not have been available for the bias analysis of the CUB process. Such a measurement without access to special personal characteristics is called unsupervised bias testing. An unsupervised bias detection tool attempts to say something about groups of students that are (potentially) treated unequally in the process, without using any information about the migration background of students living away from home. The groups of students are classified into groups (also known as clusters) based on a fairness metric, such as 'classified as high risk by the risk profile' or 'selected for inspection.' The deviant clusters can be manually inspected for proxy characteristics, such as whether a deviant cluster (in terms of the fairness metric) primarily consists of young university students who live close to their parents, or conversely whether it consists of mbo 1-2 students in the age group 25-50 years.

For this experiment, we used the so-called Hierarchical Bias-Aware Clustering (HBAC) algorithm, as described in scientific literature.[29] This algorithm is applied to the college grant population of 2014. As the fairness metric 'High risk classification (category 1 or 2) by the risk profile' is chosen. All cases where the risk category is unknown are removed. The total population thus consists of 214.599 students living away from home. The HBAC algorithm takes as input the education type, age, and distance to parent(s) of a student. For this experiment, the HBAC algorithm uses the k-modes clustering algorithm, as this clustering algorithm is suitable for categorical data. The HBAC algorithm is tuned so that it produces a result if a deviant cluster consists of at least 25.000 students. The implementation of the HBAC algorithm is open-source and can be found on GitHub.[30]

In this experiment, three clusters are found that differ significantly in terms of the chosen fairness metric. In cluster 1, 3% of the 58.362 students are classified as 'high risk'; in cluster 2, this is 21% of the 129.041 students; and in cluster 3, this is 53% of the 27.196 students. Cluster 1, the cluster with the lowest degree of bias, mainly consists of mainly university students (96%) and many students who live far from their parents (91% lives more than 20km from their parent(s)). In cluster 3, the cluster with the highest degree of bias, exclusively mbo 3-4 students are present, and many students who live close to their parent(s) (1m to 5km) (44%). See Figure 49.

[29] Joanna Misztal-Radecka, Bipin Indurkhya, Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems, Information Processing & Management, Volume 58, Issue 3, 2021.

[30] https://github.com/NGO-Algorithm-Audit/unsupervised-bias-detection

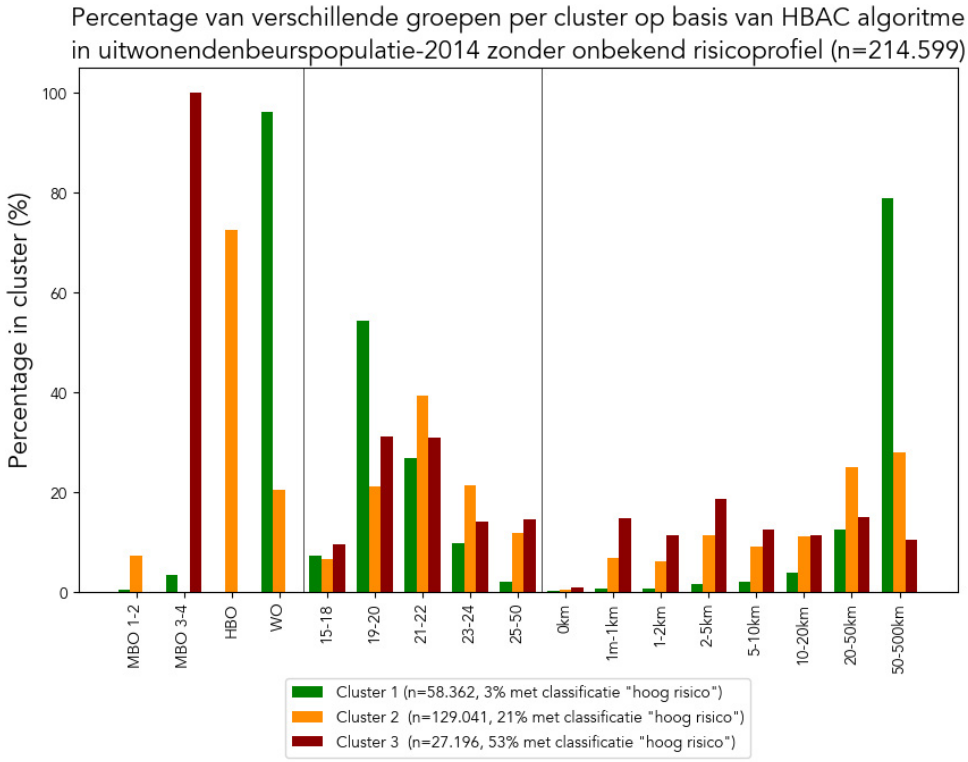

Figure 49 – Percentage of different groups per cluster based on HBAC algorithm in the college grant population-2014, without unknown risk profile (n=214.599)

In summary, the results of the unsupervised bias detection tool broadly correspond to the results of the supervised bias analysis. Note that in this case, the HBAC algorithm does not have access to the logic used in the risk profile, but nevertheless finds (some) comparable groups that are classified as high risk. Such application of unsupervised bias detection could thus provide a signal to human experts that deviations occur in a process and that the cause of this should be further investigated.

Unsupervised bias detection is therefore a promising method that should be extensively tested and further developed in the near future. It enables public and private organizations to detect and mitigate (undesirable) bias in algorithms without access to special personal data.

Algorithm Audit

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Stichting Algorithm Audit is registered as a non-profit organisation at
the Dutch Chambre of Commerce under license number 83979212