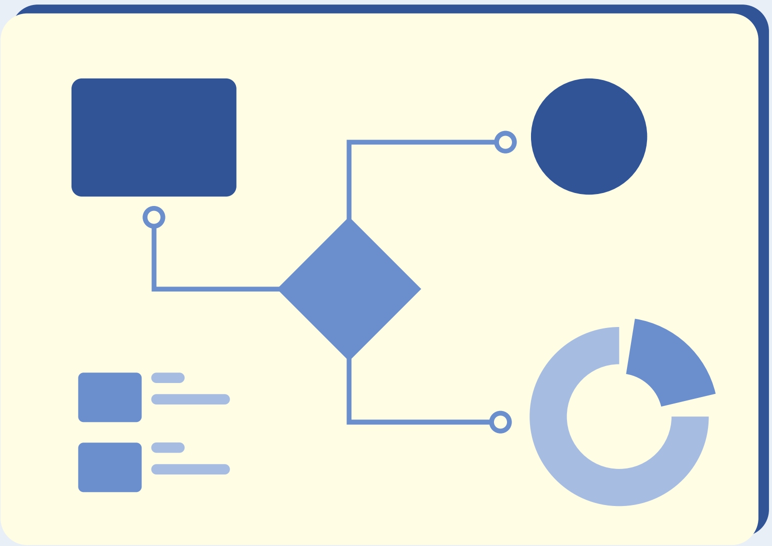
Algorithm Audit

Dienst Uitvoering Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap



Preventing prejudice

Recommendations for risk profiling in the College Grant Control process: a quantitative and qualitative analysis

February 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 a 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, such as our unsupervised bias detection tool and methods for 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 white papers

1. Summary

Below follows the management summary (§1.1), the findings (§1.2), the recommendations made to DUO (§1.3) and a reading guide (§1.4).

1.1 Management summary

In the period 2010-2023, the Dutch Education Executive Agency (Dienst Uitvoering Onderwijs – DUO) used an effective and simple risk profile to trace unduly allocated college grants.[1] Based on three selection criteria, DUO prioritized the use of enforcement and inspection resources within the College Grant Control (Controle Uitwonendenbeurs – CUB) process.[2] In 2023, the CUB process gained media attention. The reason was an investigation by Investico, NOS op 3 and the Higher Education Press Agency.[3] Journalists reported that, according to consulted lawyers, 97.6% (367 out of 376) of the students they assisted in cases where DUO imposed a fine for unduly use of the college grant have a migration background. To investigate possible bias in the CUB process, DUO has commissioned Algorithm Audit to 1) conduct quantitative research into occuring bias in the entire CUB process and 2) conduct qualitative research into the risk profile used in the CUB process. Other external parties investigate other aspects of the CUB process. This report contains the findings of the quantitative and qualitative research conducted by Algorithm Audit. An overview of the scope of this report is shown below.

What is this report?	What is this report not?
A description of the development of the risk profile and the way in which the profile was documented and evaluated in the period 2010-2023. A qualitative analysis of the development and use of the risk profile.	Research into whether the risk profile has lawfully been used. Therefore, there is: > No testing against standards from the period 2010-2022, but due to the future-oriented nature of this research against standards from 2023. > No discussion on possible justifications for the use of the risk profile, such as political-executive pressure in the period 2010-2022.

[1] The word effective is used here as DUO also used the term internally: better at detecting unduly use than random sampling. Figures for the effectiveness in the period 2011-2017 are given in 4. Results of quantitative analysis. The wording used internally by DUO is also used in this report for other terms used in the CUB process, such as 'parent(s)'.

[2] The three selection criteria are type of education, age and distance to parent(s). The exact composition of the risk profile is elaborated on in 2.3 Overview of CUB process.

[3] Published on NOS, Investico, De Groene Amsterdammer, Trouw, among others.

What is this report?	What is this report not?
A quantitative analysis of: > The risk scores assigned to students by the risk profile. > Manual selection of students for a control procedure by civil servants.	A quantitative analysis of bias with respect to migration background. This research has not been possible to date because the data required for it have not yet been made available by Statistics Netherlands (Centraal Bureau for Statistiek – CBS).
The report focuses on the future. It is intended to help DUO indicate how the risk profile holds up in the light of current standards for profiling. It is also intended to help DUO to deploy algorithms and risk profiles responsibly if it is decided to use them again in the CUB process. There is:	Investigation into whether the risk profile and/or the CUB process were discriminatory. This is a legal question, the answer to which may make use of the results of the quantitative analysis, but which is not answered as such in this report.
> Tested against standards that are *state-of-the-art* at the time of the publication of the report. > Advised to establish internal standards and set up processes for the use of algorithms.	A normative assessment of the risk profile and/or the CUB process. An exhaustive investigation into the origins and use of the CUB process. The research was conducted based on information made available by DUO.

Quantitative analysis CUB process

The CUB process has been analyzed quantitatively in two ways. For the first analysis, the results of the 2014 and 2017 random sample were examined. The purpose of this analysis is to test the selection criteria and associated subcategories used in the risk profile for statistical relevance.[4] The second analysis is a so-called bias measurement. This bias measurement is based on data about the course of the CUB process in 2014 and 2019 and aims to measure whether students in certain education, age and distance categories are selected excessively for a control procedure compared to the risk score assigned by the risk profile. In addition to the analysis of the samples and the bias measurement, an explanatory data analysis of the studied student populations from 2014 and 2019 is presented. The effectiveness of the CUB process compared to random samples to trace unduly use of the college grant is also examined.

Due to insufficient available data, Algorithm Audit has not yet been able to conduct quantitative research into the potential bias of the CUB process regarding students with a migration background. It has been verified that the used risk profile does not make any direct differentiation based on migration background (nor other protected grounds).

[4] The risk profile consists of the following criteria: education, age and distance to parent(s). The exact composition of the risk profile is explained in 2.3 Overview of CUB process.

The random samples from 2014 and 2017 provide insufficient statistical support for the selection criteria type of education and age as used in the risk profile. A consistent and statistically significant relationship has been found for various distance categories with unduly use of the college grant, albeit mainly for specific subcategories such as a small or very large distances to the address of parent(s). No support has been found for the usage of six different risk categories compared to a simple binary division into a low-risk and high-risk category. The results were obtained by performing one-sided statistical Z-tests on the unduly use percentages per education, age, and distance category. The results are confirmed by Fisher's exact test.

The bias measurement of the CUB process shows that students who are registered within a distance of 2 km from their parent(s) are manually selected for a home visit significantly more often than one would expect based on the risk scores assigned by the risk profile. This difference was measured by breaking down the percentage of students selected for a home visit by education, age, and distance category and comparing this with the assigned risk scores. It is likely that specific work instructions that encourage the manual selection of students who are registered close to the address of their parent(s) are the cause of this.

These findings lead to the recommendation to formulate more substantive rationales to use profiling criteria before the risk profile is possibly put back into use. In addition, further research must be conducted to determine whether there is a connection between the groups that are excessively manually selected for home visits and students with a migration background. In the future, if the CUB process is reintroduced, repetitive use of the same selection criterion in different steps of the CUB process must be prevented. Besides, an investigation and improvement trajectory must be initiated for the process of manual selection of students for a control procedure, including the work instructions that civil servants follow.

Qualitative analysis of risk profile

For the qualitative analysis, the design and deployment of the risk profile was tested against the standards and guidelines of 2023. These include the Dutch Fundamental Rights and Algorithms Impact Assessment (FRAIA) and the Algorithms Research Framework of the Dutch Government Audit Agency (ADR). The actual situation of the CUB process has been constructed through document analysis and in-person workshops. It is impossible to determine whether the researchers of Algorithm Audit received all available documents. It cannot be ruled out that there are unseen documents that shed a different light on the findings in this report. DUO points out that the profile was developed 13 years ago, which means that not all relevant documents are still available.

The following conclusions follow from the qualitative analysis.

Between 2010 and 2022, insufficient attention was paid within DUO to possible side effects of usage of the risk profile. There has been no evidence that, when drawing up and using the risk profile, it was investigated whether certain demographic groups are over- or under-represented in the higher risk categories. It is worth mentioning that this is unintentional bias: the investigation did not reveal any intentional bias or discrimination. After the extent of the childcare benefit scandal came to light, DUO investigated in 2019 whether the risk model directly discriminates. It was concluded that this was not the case. Indirect discrimination does not appear to have been investigated or discussed at the time.

The selection criteria of the risk profile have been public since 2012. These criteria are stated in letters to parliament [A.19] and in case law.[5] Nevertheless, the investigation did not show that other parties have pointed out to DUO the risk of bias when using these selection criteria. In any case, this concerns the House of Representatives, the Ministry of Education, Culture and Science (Ministerie van Onderwijs, Cultuur en Wetenschap – OCW), the Dutch Government Audit Agency (Audit Dienst Rijk – ADR), journalism, the legal profession and civil society.

> Testing past facts against standards from 2023 may seem too stringent. That would be the case for answering normative questions such as the question of guilt. However, these are the relevant standards for political-administrative interpretation in 2023. DUO needs to take into account these standards and processes to use algorithms and risk profiling in the future.

Findings and recommendations

Based on this analysis, actions can be identified to improve and prevent bias in the CUB process in the future. This report makes a start with formulating seven findings and four recommendations.

Findings

Finding 1 – A rule-based algorithm has been used for years, which assigned a risk score to students who received a college grant based on type of education, age and the distance between the student's address and the address of his/her parent(s). No self-learning algorithm or artificial intelligence system has been used. The use of risk profiling has proven to be effective.

Finding 2 – The random samples from 2014 and 2017 show insufficient statistical relationship between the selection criteria type of education and age, and unduly use of the college grant. A statistical relationship exists between specific categories within the selection criterion ‘distance to parent(s)’ and unduly use of the college grant. Insufficient statistical support has been found for the division into six risk categories compared to a binary risk classification.

[5] Including District Court Limburg December 13, 2013, ECLI:NL:RBLIM:2013:11417, District Court Rotterdam July 14, 2014, ECLI:NL:RBROT:20145684.

Finding 3 – Students who are registered within a distance of 2 km from their parent(s) are manually selected for a home visit significantly more often than one would expect based on the risk scores assigned by the risk profile. It is likely that specific work instructions, which encourage the manual selection of students who are registered near their parental address, are the cause of this.

Finding 4 – No direct differentiation has occurred on the basis of migration background (nor based on other protected grounds) in the risk profile. Due to insufficient available data, it has not yet been possible to conduct quantitative research into indirect bias towards students with a migration background.

Finding 5 – According to current standards, a well-motivated rationale for used selection criteria and risk categories in the risk profile are lacking, especially regarding possible bias.

Finding 6 – There has been no internal research into possible biases in the risk profile, neither during development and deployment of the risk profile.

Finding 7 – There are no internal standards for responsible use of algorithms.

These findings are further explained in 1.2 Findings.

Recommendations

Recommendation 1 – Provide a well-motivated rationale for usage of risk profiling and specific criteria before profiling is potentially used again, among others with help of a normative framework.

Recommendation 2 – Further research should be conducted to determine whether there is a link between the groups that are excessively manually selected for home visits and students with a migration background.

Recommendation 3 – Avoid using the same selection criterion in different steps of the CUB process. Set up an investigation and improvement trajectory for the manual selection process, including redesign of the work instructions.

Recommendation 4 – Establish an organization-wide algorithm management policy to reduce the risks associated with the use of algorithms.

These recommendations are further explained in 1.3 Recommendations.

1.2 Findings

This section explains the findings mentioned. The findings arise from the quantitative and qualitative analysis of the risk profile and how this profile is developed and deployed within DUO.

Finding 1 – A rule-based algorithm has been used for years, which assigned a risk score to students who received a college grant based on type of education, age and the distance between the student's address and the address of his/her parent(s). No self-learning algorithm or artificial intelligence systems has been used. The use of risk profiling has proven to be effective.

Between 2012 and 2023, a risk profile was used in the CUB process to automatically assign risk scores to all students who received a college grant from DUO. The risk profile is a linear model that assigns students a risk score based on three criteria. These criteria are type of education, age and the distance between the address where they were registered and the address of their parent(s). Each criterion is divided into categories, for example students who live between 1m-1km from their parent(s), students who live 50-500km from their parent(s), are following vocational training (mbo 1-2) etc. After applying predetermined weighting factors per category, a student's profile leads to a risk score. The risk score can be increased if the age of the student known to DUO differs from the age of the student in the General Registration of Persons (BRP). All students have been assigned a risk score. Applying the CUB process to trace unduly usage of the college grant, of which the risk profile is an important part, has proven to be effective. The effectiveness is evident from the fact that more unduly use of the college grant has been identified with application of the CUB process than with random samples. This concerns 3.6% and 3.8% effectiveness respectively in the random samples of 2014 and 2017 and 38.9% and 35.3% effectiveness respectively when applying the CUB process in 2014 and 2019 in which the risk profile was used.

More information about the risk profile and the CUB process is provided in 2.3 Overview of CUB process. More information about the effectiveness figures can be found in 3.1 Quantitative analysis.

Finding 2 – The random samples from 2014 and 2017 show insufficient statistical relationship between the selection criteria type of education and age, and unduly use of the college grant. A statistical relationship exists between specific categories within the selection criterion 'distance to parent(s)' and unduly use of the college allowances. Insufficient statistical support has been found for the division into six risk categories compared to a binary risk classification.

Per education, age and distance category, it has been counted how often students that were selected for a control procedure in the random sample in 2014 (n=387) and 2017 (n=293) unlawfully used the college grant. Based on these frequencies, it was tested, using a one-sided

Z-test and Fisher's exact test, whether the differences in unlawfulness percentages for the categories used in the risk profile are statistically significant. For example: is there a statistically significant difference between percentages of unduly allocation of college grants between the distance category 2-5km and the category 5-10km? For the profiling criterion age, insufficient evidence was found for statistically significant differences between the used categories. For the criterion type of education, one significant difference was found in the 2014 sample but no significant differences in the 2017 sample. Algorithm Audit considers this to be an insufficiently consistent signal on which risk profiling should be based. A new frequency count on a larger random sample could change this. For distance to parent(s), evidence is found for statistically significant differences between unduly usage percentages in the 1m-1km, 2-5km and 50-500km categories. This provides support for the use of distance as a selection criterion for risk profiling, although the binning thresholds for categorization should be determined more precisely. In addition, there is no quantitative support to divide the assigned risk scores into six risk categories. There is statistical support for a binary risk classification, and this simplification is preferred.

More information about the methodology of the statistical tests can be found in 3.1 Quantitative analysis. The results are discussed in 4.1 Results of random sample 2014 and 2017.

Finding 3 – Students who are registered within a distance of 2 km from their parent(s) are manually selected for a home visit significantly more often than one would expect based on the risk scores assigned by the risk profile. It is likely that specific work instructions, which encourage the manual selection of students who are registered near their parental address, are the cause of this.

In the bias measurement, a strong overrepresentation was observed of students who are registered within 2 km of their parent(s) during manual selection for home visits. This overrepresentation does not stem from the risk profile used.

In 2014, for students who are registered 0km from their parent(s), there is a disproportionate ratio between the probability of being selected for a home visit (8.3x the average of all categories) and the assigned risk score (3.7x the average). This means that students from this category are more than twice as likely to be selected for a home visit than would be expected based on their assigned risk score. For the 1m-1km category, this ratio between the probability of a home visit and risk score (both compared to their averages) is 6.8x : 3.0x and for the 1-2km category 3.5x : 2.7x. This means that the group with a small distance to parent(s) is structurally overrepresented in the manual selection for home visits, compared to the risk scores assigned to them.

This overrepresentation can probably be explained by the presence of specific work instructions that instruct employees to specifically select students with specific living conditions – such as a combination of young age, living close to their parent(s) and/or living with family members – for a home visit. The overrepresentation of students (such as those registered at a small distance from their parent(s)) in the manual selection for home visits could potentially cause an overrepresentation of other characteristics, such as their migration background. Further research is therefore needed to determine whether there is a connection between the groups that are excessively manually selected for home visits and students with a migration background. See Recommendation 2. In addition, an investigation and improvement trajectory must be set up for the manual selection process, including a review of the work instructions. See Recommendation 3.

The results of the bias measurement, including the figures mentioned, can be found in Figure 8 and in 4.2 Results of bias measurement 2014 and 2019.

Finding 4 – No direct differentiation has occurred on the basis of migration background (nor based on other protected grounds) in the risk profile. Due to insufficient available data, it has not yet been possible to conduct quantitative research into indirect bias towards students with a migration background.

DUO has requested Algorithm Audit to also carry out a bias measurement for possible bias regarding the protected attribute migration background. DUO itself has no data on the migration background of students. The Netherlands' national office of statistics (CBS) has therefore been asked to enrich the DUO data at group level with data on the migration background of students solely for the purpose of carrying out this bias measurement. To date, Algorithm Audit has been unable to obtain this data. In the future, this data may be made available for possible follow-up research. Algorithm Audit has considered alternative methods to measure indirect bias regarding migration background, such as using aggregation statistics per postal code area. However, according to statistical experts affiliated to Algorithm Audit this approach is insufficiently justified from a methodological point of view. At the time of publication of this report, Algorithm Audit was therefore unable to conduct quantitative research into indirect bias of the CUB process regarding students with a migration background. Possible indirect bias in the risk profile or in the further course of the CUB process cannot be ruled out.

How the bias measurement for migration background would have been carried out in the case of sufficient available data is explained in 3.1 Quantitative analysis.

Finding 5 – According to current standards, a well-motivated rationale for used selection criteria and risk categories in the risk profile are lacking, especially regarding possible bias.

The rationale for usage of criteria in the risk profile are largely based on personal experiences and so-called 'common sense' of employees. The suitability of these criteria for the aim pursued is not documented. The origins of the risk profile can be traced back to a series of workshops in 2010. Although subject matter expertise and common sense are important and useful, *self fulfilling prophecies*, *confirmation biases* and the unintentional use of proxy criteria (see Box 1) are a real risk here. The weighting factors used in the risk profile are based on a data study conducted in 2010. Documentation about the origin and methodology used to compose the control group in this data study is lacking. Without further research into the representativeness of this control group, certain groups may be over- or underrepresented in the population, specifically in case the group is manually sampled. Basing weighting factors on that population risks creating *negative feedback loops*. This means that groups that are overrepresented in the control group receive a higher weighting factor, as a result these groups are later systematically assigned an excessive risk score by the risk profile. Substantiating the choice of certain selection criteria and weighting factors and their possible influence on bias is prescribed by the currently applicable frameworks for the responsible use of algorithms such as the Fundamental Rights Algorithm Impact Assessment (FRAIA) and the Algorithms Research Framework of the ADR.

An analysis of the substantiation of the selection criteria and categories used in the risk profile is provided in 5.1 Qualitative analysis of risk profile.

Finding 6 – There has been no internal research into possible biases in the risk profile, neither during development and deployment of the risk profile.

Risk profiling is a means permitted by the Netherlands General Court of Appeals for increasing effectiveness, efficiency, and cost savings in combating unduly use of public allowances.[6] However, conditions apply to usage of risk profiling, including that made differentiation must be suitable, necessary, and proportionate. This must be guaranteed during the design and deployment of the risk profile. There has been no evidence of awareness within DUO (and/or the bodies involved such as the Ministry of Education, Culture and Science, the ADR, the House of Representatives after the introduction of CUB in 2012, etc.) that the use of the risk profile in the CUB process entails a risk of (indirect) unequal treatment. There has also been no evidence that checks and balances have been put in place to monitor or prevent such unequal treatment. No bias was measured, and no investigation or consideration was given to whether the criteria and weighting factors used may have been proxy characteristics for certain demographics.

An explanation of the lack of internal investigation into bias can be found in 5.2 Processes to address the risk of bias.

Finding 7 – There are no internal standards for responsible use of algorithms.

This study found no internal guidelines used by DUO to prevent bias or mitigate other risks when using algorithmic methods. It does not appear to have been tracked how DUO imple-

[6] Including CRvB 8 September 2015, ECLI:NL:CRVB:2015:3249, paragraph 4.5.; see also Assessment Framework for Discrimination through Risk Profiles of the Netherlands Institute for Human Rights (2021) Guideline 1.

Box 1

What are proxy attributes?

Proxy attributes refer to apparently neutral characteristics that are strongly correlated with protected grounds, such as gender or etnicity. Due to a strong correlation, differentiation based on proxy attributes can (possibly unintentionally) also affect the protected group. Membership in a female student association, for example, is a proxy for the personal characteristic 'woman'; owning a season ticket for a football club may be a proxy for the personal characteristic 'man'. Proxy effects must be considered when profiling is applied. It may be permitted to distinguish between citizens with andwithout a season ticket, but not between men and women.

ments general legal frameworks and guidelines from other authorities for the composition of risk profiles. If these guidelines do exist, not the entire organization is aware of them. Guidelines are lacking in at least two areas that are relevant to prevent (indirect) discrimination. Firstly, guidelines for which degree of correlation between risk profiling criteria and protected grounds indirect discrimination occurs. Secondly, to determine under what circumstances an objective justification exists for usage of profiling criteria, because the legal requirements of suitability, necessity and proportionality are met. In the period 2012-2022, annual monitoring reports, privacy audits and abuse and misuse checks (misbruik-en-oneigenlijk gebruik – M&O) took place. Random sampling took place in 2010, 2014 and 2017, primarily with the aim of determining the effectiveness of the CUB process and to estimate the financial 'residual risk' in the entire CUB population. The lack of guidelines for dealing with (algorithmic) risk profiling has contributed to the fact that the checks did not pay attention to possible bias.

An explanation of the lack of internal standards for the use of algorithms is given in 5.2 Processes to address the risk of bias.

1.3 Recommendations

The following recommendations follow from the findings mentioned.

Recommendation 1 – Provide a well-motivated rationale for usage of risk profiling and specific criteria before profiling is potentially used again, among others with help of a normative framework.

Inherent in the use of risk profiling is that differentiation is made between groups of students. This is partly the intention: not a *bug*, but a *feature*. However, treating groups of people differently can also exceed legal and social norms. This is, for example, the case when differentiation is not suitable, necessary, and proportionate. At the same time, differentiation that is lawful can still be considered socially and ethically undesirable. This could include differentiation that is not focused on a legally protected ground, but on different characteristics such as education level. It is recommended to devise an internal framework against which can be assessed which forms and to what extent differentiation of groups of people are considered (un)desirable. Such a framework specifically relates to testing criteria that precede usage of the risk profile, such as statistical hypothesis testing. The proposed Phase 2 study of Algorithm Audit provides starting points for making normative considerations to interpret quantitative testing results. Assessing a risk profile against a framework must also contain quantitative support of the selection criteria used, for which the analyzes in this report provides a preliminary step. To motivate the selection profile quantitatively, and not just measure the effectiveness of the CUB process, it is recommended to draw a larger random sample than the random samples from 2014 and 2017.

Recommendation 2 – Further research should be conducted to determine whether there is a link between the groups that are excessively manually selected for home visits and students with a migration background.

Whether groups that are selected excessively during manual selection also have a migration background is not clear and must be further investigated. The work instructions used, for example the guideline to further investigate details of the living situation (such as the combination of young people and living with family), can contribute to possible over-representation of demographic groups, including students with a migration background. In addition, the process of manual selection is susceptible to discrimination based on latent characteristics. At the lists that employees are presented with during manual selection, in addition to the criteria from the risk profile and the risk score, other student characteristics were also visible, such as name, address and date of birth. There is a risk of (unconscious) bias regarding migration background or other characteristics in manual selection. The influence that these and other characteristics play during the manual selection for home visits needs to be investigated. In addition, an improvement trajectory must be set up to improve work instructions for manual selection. See Recommendation 3.

Recommendation 3 – Avoid using the same selection criterion in different steps of the CUB process. Set up an investigation and improvement trajectory for the manual selection process, including redesign of the work instructions.

It has been established that students who are registered within a distance of 2 km from their parent(s) have been manually selected for a home visit significantly more often than would be expected based on the risk scores assigned by the risk profile. It is obvious that specific work instructions that encourage the manual selection of students who are registered near their parental address are the cause of this. These findings provide reason to examine the work instructions for employees in the selection process. If the risk score assigned by the risk profile already considers characteristics such as living nearby parent(s), further instruction to manually select the same characteristics more often may cause an overreaction. As a result, students with a short distance to their parent(s), for example, are wrongly selected for home visits much more often. The overall manual selection process needs further investigation to identify areas for improvement. The current process is opaque, because the procedure to manually include and exclude students from home visits based on their risk scores and relevant characteristics is not a clear protocol. It is difficult to find out exactly how employees work, and the process is therefore difficult to control. In addition, latent personal characteristics that are visible to employees (such as name, address, date of birth) can pose a risk of (unconscious) bias. Anonymizing students and exclusively showing only the relevant characteristics and the risk score would be a possibility. Further research should reveal what improvements can be made in the manual selection process.

Recommendation 4 – Establish an organization-wide algorithm management policy to reduce the risks associated with the use of algorithms.

Draw up a policy for consistent and responsible handling of algorithms and risk profiling. Such policies can be drawn up in different dimensions. At this point one could think of:

- Governance: centralization of key decision-making in committees, alignment of roles and responsibilities within the existing organizational structure, implementation of 3 lines-of-defense (3LoD) risk management framework.
- Documentation: drawing up standardized documentation requirements and work instructions.
- Algorithm inventory: central overview of all algorithms[7] within the organization for information requests, monitoring and evaluation.
- Processes: Establish evaluation and validation processes for each algorithm.
- Monitoring and reporting: Monitoring of reporting on algorithmic risks promotes risk-oriented practices.

Inspiration for a different division can also be drawn from existing frameworks for responsible handling of algorithms in the public sector.[8]

1.4 Reading guide

Chapter 2 explains the approach how facts were reconstructed, describes the chronology of the CUB process and discusses the various steps in the CUB process.

Chapter 3 introduces the research methodology and research questions for both the quantitative and qualitative analysis.

Chapter 4 discusses the results of the quantitative analysis.

Chapter 5 discusses the results of the qualitative analysis.

Chapter 6 contains disclaimers about this research.

Chapter 7 contains the conclusion in the form of findings and recommendations.

Appendix A refers to relevant literature.

Appendix B contains an example of a query to retrieve data from the DUO data warehouse.

Appendix C provides additional information about the statistical tests performed.

Appendix D contains the complete list of documents.

[7] For a definition of an algorithm see: https://algoritmes.overheid.nl/nl/footer/over-algoritmes.

[8] See, for example, the Implementation Framework 'Responsible Use of Algorithms' of the National Government https://www.rijksoverheid.nl/documenten/rapporten/2023/06/30/implementatiekader-verantwoorde-inzet-van-algoritmen, Algorithm Research Framework of the Dutch Government Audit Agency (ADR) https://www.rijksoverheid.nl/documenten/rapporten/2023/07/11/onderzoekskader-algoritmes-adr-2023 and the Algorithms Assessment Framework of the Court of Audit https://www.rekenkamer.nl/onderwerpen/algoritmes-digitaal-toetsingskader.

2. Background College Grant Control procedure: fact reconstruction

Below follows the approach how facts are reconstructed (§2.1), the chronology of the CUB process (§2.2) and an overview of the CUB process (§2.3).

2.1 Approach fact reconstruction

Fact reconstruction took place based on a predetermined approach and consisted of a desk research and workshops. DUO has shared documents about the CUB process. Algorithm Audit examined those documents and requested new documents as a result of investigation. New documents were subsequently provided by DUO, which were examined. These steps have been completed several times. In total, more than 125 documents were shared and examined. These documents are referred to using the abbreviations as maintained in Appendix D – Document list.

At the same time as this desk research, two workshops were organized with DUO employees who have been working with the CUB process in the past. During these workshops, Algorithm Audit raised questions and DUO provided background information about the CUB process. In consultation with DUO, it was decided not to mention in this report which employees were interviewed and not to quote from the workshops. This is to ensure that employees could speak freely. The advantages of such workshops are being able to receive information quickly, being able to link information that is in different parts of the organization and being able to more easily understand the broad outlines of the CUB process that do not always emerge from document analysis. Disadvantages are the risk of bias in the selection of the workshop group, the risk of hidden agendas of the participants and peer pressure. Because of this and other influences, incomplete or incorrect information may have been shared during the workshops. Because the insights on which this report is based were obtained from a combination of written documents and the workshops, these risks have been mitigated as much as possible.

In addition to the workshops, a focus group was held in which the necessary historical data was collected with data warehousing and CUB experts. Additional data was shared digitally after the focus group. Based on the initially shared data, an analysis was made, and an earlier version of this report was published. In February 2024, it turned out that the historical data shared by the data warehousing team was incorrect. The shared data that would only concern January 1, 2014, and January 1, 2019 also appeared to contain data from 2013 and 2018 respectively. The current version of this report contains the correct data, as far as Algorithm Audit is aware of and insofar as it can be checked.

An interim version of the factual investigation and the qualitative analysis were submitted to DUO on October 25 and December 14, 2023. Additions and comments were received which have been incorporated into this version.

Quality checks were carried out on the final report by Algorithm Audit employees and board member who were not otherwise involved in the project.

2.2 Chronology CUB

On October 31, 2007, the predecessor of DUO received signals from the Social Investigation Department Twente (Sociale Recherche Twente – SRT) about unduly use of college grants. The SRT informed the Minister that they regularly encountered such unduly use, but had no powers to take action against this [F.7,F.8]. At that time, unduly use with college grants was only checked administratively by comparing the address of the student registered as non-resident with the addresses of his/her parent(s) in the Municipal Personal Records Database (Gemeentelijke Basis Administratie – GBA). Home visits were not conducted and there was a heavy burden of proof on DUO to establish unlawful use [F.8]. In July 2009, members of the Dutch Parliament asked the Minister of Education, Culture and Science about the scale of unduly use of the college grant [A.16]. A report was subsequently published in October 2009 by the Social Intelligence and Investigation Service (Sociale Inlichtingen- en Opsporingsdienst – SIOD) [e.g. A.17]. Also, a workshop took place in October 2009 on tackling unduly use of the college grant [e.g. A.17]. Present were the predecessor of DUO, the SRT, the Public Prosecution Service, the Basic Administration for Personal Data and Travel Documents, the SIOD and the Ministry of Education, Culture and Science. During and after the workshop, various measures were conceptualised to prevent unduly use. Among other things, the Pilot Twente has been started [e.g. A.17].

In the Pilot Twente, the risk profile was drawn up in an iterative process between the predecessor of DUO and the SRT. DUO's predecessor came up with risk criteria. These were initially the distance between the student living away from home and his/her parent(s), the distance between the student living away from home and the educational institution, a large number of registrations at a certain address and certain moving times. The SRT conducted studies based on these risk criteria, after which the criteria were refined [A.17]. From conversations in 2023, it follows that the risk criteria were chosen based on experience, common sense and by manually searching for patterns in lists of students living away from home. This involved, for example, situations in which children from two or more families registered at the address of the other family, either crosswise or in a carousel.

In the first half of 2010, the Twente pilot was expanded with new participants (namely the Municipalities of Amsterdam, Rotterdam, The Hague, Utrecht, and the Social Investigation Department of North Holland) [F.11]. In March 2010, a risk profiling workshop was held in which various pilot municipalities, DUO, OCW and SIOD participated [e.g. E.1-2]. The 'decision tree' method [e.g. E.5]. was chosen to determine the risk profile. Estimations of unduly use are made based on subpopulations of students living not at their parent's place but on their own. Student data and risk profiles have also been exchanged between DUO and the pilot municipalities based on working agreements [E.5]. In 2010, DUO also set up a data warehouse containing data from all students living away from home, based on which an analysis of the selection criteria was made [F.11].

In 2010, the risk profile was refined by determining the weighting factors of three selection criteria and the mutual relationship between those weighting factors based on 934 students [A.47]. The substantiation is not documented.

In 2010, the ADR also carried out a random sample among all students who received a college grant from living away on their own [A.51].

In 2012 use of the risk profile started [A.19].

In 2012, a change in the law came into effect, giving DUO additional resources to combat unduly use of the college grant [e.g. A.19].

In 2013, 366 students living in student housing in Groningen were inspected. Given the practical and legal difficulties in inspecting student housing and the low amount of unduly usage found during the home visits in 2013, student housing has been exempt from inspections since 2013 [A.11].

In 2014, the ADR again carried out a random sample [A.40]. The entire population of students who received a college grant was examined. An additional sample was taken within the four lowest risk categories. After desk research and home visits, it was determined that 3.52 percent of the students surveyed were unlawfully receiving a non-resident student grant. This is statistically relevant with 95 percent certainty. This leads to the estimate that in 2010 there was a residual risk of EUR 12 to 22 million per year. DUO concludes that due to the CUB control procedure unduly use of the college grant decreased by EUR 30 million between 2011 and 2014.

In 2015, the loan system was introduced. The basic college grant – and thus the grant for people living away from their parent's place – has been abolished for students in higher education. Adjusting the risk profile is being considered, but this is not being executed [A.33].

In 2017, a random sample will be carried out again [A.15]. Given the abolition of the loan system, this sample exclusively concerns the mbo 1-2 and mbo 3-4 students. 3.65 percent of students have been found to be unlawfully receiving a non-resident student grant. This amounts to a residual risk of EUR 3.75 million per year for mbo students. Including higher education, the residual risk in 2017 is estimated at EUR 8 million.

In 2018, the risk profile was re-evaluated following the childcare allowance affair. DUO saw no reason to adjust the CUB process.

From 2021 to 2023, the Higher Education Press Agency asked DUO questions about the CUB process. A Woo request has also been made by Investico. These questions have been answered by DUO and the Woo request has been met.

In 2023, parliamentary questions were asked about the CUB process in response to news articles based on research by NOS op 3, Investico and the Higher Education Press Agency [e.g. A.42]. Selection based on the risk profile was then stopped.

2.3 Overview of CUB process

The CUB process can be divided into seven steps:

- Step 0 – Source data: Determining all recipients of a college grant for a given reference date. This group is referred to as the *college grant population*.
- Step 1 – Risk profile: Assigning a risk score between 0-180 to all students in the college grant population based on the risk profile as shown in Table 1-2. The risk profile works as follows:
    - A weighting factor R_1 is assigned for a student's type of education (mbo 1-2: 1.2, mbo 3-4: 1.1, hbo: 1, wo: 0.8);
    - A risk score R_2 is assigned for the combination of a student's age[9] and the distance[10] between a student's zip code and the zip code of one of the parent(s). The risk factors for all combinations of age and distance categories are listed in Table 1. Note: 0km to parent(s) means that a student is registered in the same postal code area (4 numbers + 2 letters) as the parent(s), but not at the same address. An unknown address means that either the address of a student is not known, or the address of the parent(s) is not known. This may be the case when the parent(s) are unknown, are deceased or a hardship clause applies[11];
    - The risk factor is determined based on a possible deviation between the age of the student as known to DUO and the age known to the Municipal Personal Administration combined with information about how long the student has been living away from home R_3. The following ages are used to determine the risk factor R_3:
        - Current_age: age of student as known to DUO at the time of calculating the risk score,
        - Age_non-resident: age at the time the person became non-resident,
        - Age_current_GBA: age at the time of the effective date of the (at that time) current home address from the Municipal Personal Records Database.

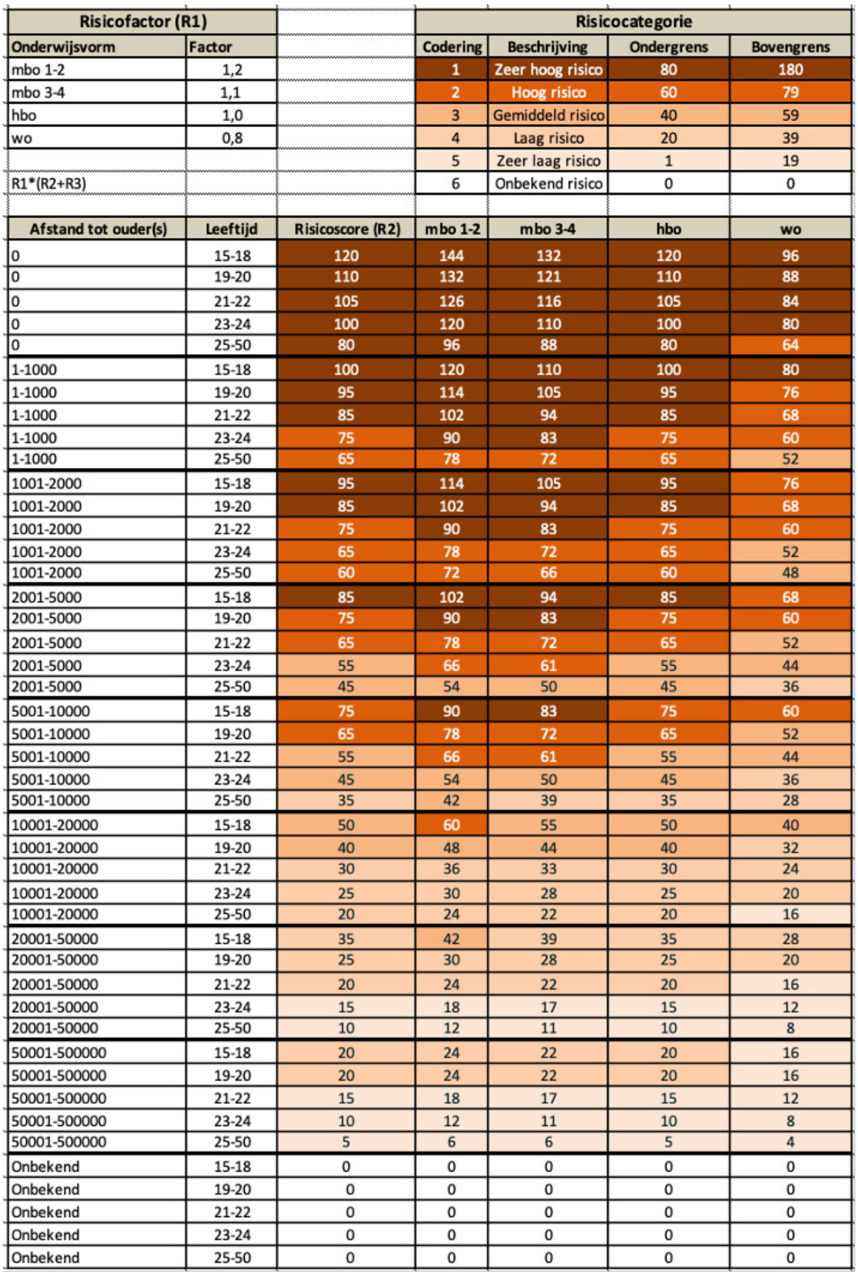
      The risk factor R_3 for all combinations of these age types is given in Table 2 [A.53].
    - The risk score is determined as follows:

$$Risk\ score = R_1 * (R_2 + R_3).$$

[9] The following age categories are used: 15-18, 19-20, 21-22, 23-24, 25-50.

[10] The following distance categories are used: 0km, 1m-1km, 1-2km, 2-5km, 5-10km, 10-20km, 20-50km, 50-500km, unknown. Note that the exact demarcation is specified in Table 1.

[11] There are also technical reasons why a student's address is not (yet) available on a reference date. For example, the student's address may not be known at the time of registration, but it may be known a few days, weeks or months later.

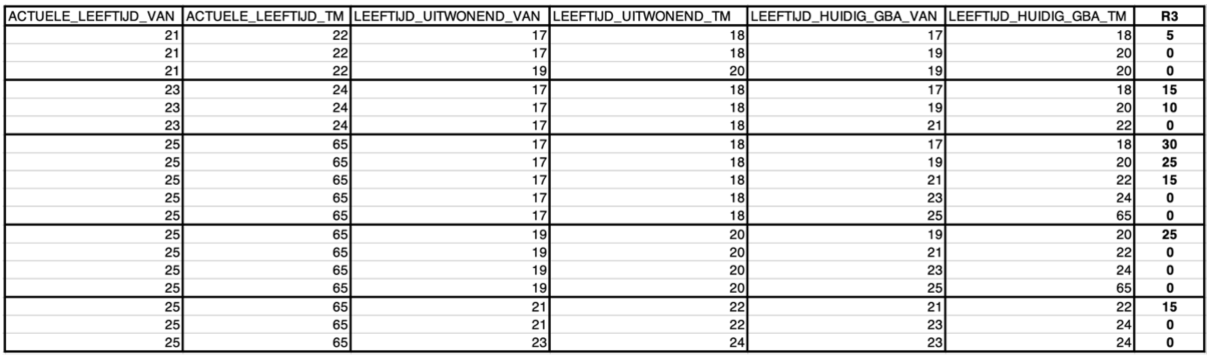


Risicofactor (R1)			Risicocategorie			
Onderwijsvorm	Factor		Codering	Beschrijving	Ondergrens	Bovengrens
mbo 1-2	1,2		1	Zeer hoog risico	80	180
mbo 3-4	1,1		2	Hoog risico	60	79
hbo	1,0		3	Gemiddeld risico	40	59
wo	0,8		4	Laag risico	20	39
			5	Zeer laag risico	1	19
R1*(R2+R3)			6	Onbekend risico	0	0
Afstand tot ouder(s)	Leeftijd	Risicoscore (R2)	mbo 1-2	mbo 3-4	hbo	wo
0	15-18	120	144	132	120	96
0	19-20	110	132	121	110	88
0	21-22	105	126	116	105	84
0	23-24	100	120	110	100	80
0	25-50	80	96	88	80	64
1-1000	15-18	100	120	110	100	80
1-1000	19-20	95	114	105	95	76
1-1000	21-22	85	102	94	85	68
1-1000	23-24	75	90	83	75	60
1-1000	25-50	65	78	72	65	52
1001-2000	15-18	95	114	105	95	76
1001-2000	19-20	85	102	94	85	68
1001-2000	21-22	75	90	83	75	60
1001-2000	23-24	65	78	72	65	52
1001-2000	25-50	60	72	66	60	48
2001-5000	15-18	85	102	94	85	68
2001-5000	19-20	75	90	83	75	60
2001-5000	21-22	65	78	72	65	52
2001-5000	23-24	55	66	61	55	44
2001-5000	25-50	45	54	50	45	36
5001-10000	15-18	75	90	83	75	60
5001-10000	19-20	65	78	72	65	52
5001-10000	21-22	55	66	61	55	44
5001-10000	23-24	45	54	50	45	36
5001-10000	25-50	35	42	39	35	28
10001-20000	15-18	50	60	55	50	40
10001-20000	19-20	40	48	44	40	32
10001-20000	21-22	30	36	33	30	24
10001-20000	23-24	25	30	28	25	20
10001-20000	25-50	20	24	22	20	16
20001-50000	15-18	35	42	39	35	28
20001-50000	19-20	25	30	28	25	20
20001-50000	21-22	20	24	22	20	16
20001-50000	23-24	15	18	17	15	12
20001-50000	25-50	10	12	11	10	8
50001-500000	15-18	20	24	22	20	16
50001-500000	19-20	20	24	22	20	16
50001-500000	21-22	15	18	17	15	12
50001-500000	23-24	10	12	11	10	8
50001-500000	25-50	5	6	6	5	4
Onbekend	15-18	0	0	0	0	0
Onbekend	19-20	0	0	0	0	0
Onbekend	21-22	0	0	0	0	0
Onbekend	23-24	0	0	0	0	0
Onbekend	25-50	0	0	0	0	0

Table 1 – Scoring table for R_1 and R_2. R_1 is the risk factor assigned to different forms of education. R_2 is the risk factor determined by the combination of the distance category and the age category of the student. In combination with R_3 – whether there is a deviation between the age of the student known to DUO and the age known to the Municipal Basic Administration, combined with information about how long the student has been living away from home – the final risk score is determined.

- > Step 2 – Division by region: Division of the college grant population, including risk score, by region. This group is referred to as the *risk score population by region*.
- > Step 3 – Manual selection by employee: A DUO employee goes through a *risk score population per region*. The risk score population per region is shared with the DUO employee in a document that contains a list of personal characteristics for every student in that region. This includes the risk score, living situation, address of the parent(s), the educational institution, and the student's type of education.[12] An

[12] [A.53] contains all the features shown. These are the correspondence number, the BSN number, the first name, the initials, the middle name, the surname, the gender, the date of birth, the street name of the GBA address, the



ACTUELE_LEEFTIJD_VAN	ACTUELE_LEEFTIJD_TM	LEEFTIJD_UITWONEND_VAN	LEEFTIJD_UITWONEND_TM	LEEFTIJD_HUIDIG_GBA_VAN	LEEFTIJD_HUIDIG_GBA_TM	R3
21	22	17	18	17	18	**5**
21	22	17	18	19	20	**0**
21	22	19	20	19	20	**0**
23	24	17	18	17	18	**15**
23	24	17	18	19	20	**10**
23	24	17	18	21	22	**0**
25	65	17	18	17	18	**30**
25	65	17	18	19	20	**25**
25	65	17	18	21	22	**15**
25	65	17	18	23	24	**0**
25	65	17	18	25	65	**0**
25	65	19	20	19	20	**25**
25	65	19	20	21	22	**0**
25	65	19	20	23	24	**0**
25	65	19	20	25	65	**0**
25	65	21	22	21	22	**15**
25	65	21	22	23	24	**0**
25	65	23	24	23	24	**0**

Table 2 – Scoring table for R_3. R_3 is the risk factor that is determined based on a possible deviation between the age of the student known to DUO and the age known to the Municipal Basic Administration, combined with information about how long the student has been living away from home. In combination with R_1 and R_2– where R_1 is the risk factor assigned to different forms of education and R_2 is the risk factor that is determined by the combination of the distance category and the age category of the student – the final risk factor is determined.

employee uses the grounds for exclusion to determine which students do not warrant external investigation.[13] Grounds for exclusion include couples (2 people who are simultaneously registered in the Basic Registration of Persons (Basis Registratie Personen – BRP) around the same age), living in a student house, living with the Salvation Army, or being a single parent. The list of students is then basically scrolled through in order from high to low-risk scores and selected for a home visit. This order (high to low) can be deviated from. For example, a student with a slightly lower risk score can be treated earlier than a student with a higher risk score because there are special living conditions (for instance: young age and living with family) [H.1 p.10]. When assessing the living conditions, the BRP is consulted for (family) relationships and the Basic Register of Addresses and Data (Basisregistratie Adressen en Gegevens – BAG) for the residential function and living area of a building. For a home visit, the following are specifically selected in accordance with the instructions to the employees:

> Students with a risk factor 1 to 3 with a registration near their parental address, a registration with a family with small children, a registration with the very elderly without a family relationship, a registration at an address without a residential function, a long travel time (>1 .5 hours each way) between the home address and the educational institution and no internship, registration at an address with more residents than 12m^2 per resident, or an illogical composition of residents such as a young person at the same address as several single older people in combination with a small living space.

> Request a further assessment from students for whom there is no link with their parents in the Study Financing System, but there is in the BRP. The same applies to

house number, the addition to the house number, the indication of the home address, the postal code, the place of residence, the effective date of the GBA address, the forms of student financing that the student received in recent months, the first month in which the student was registered as living away from home, whether the student is living at home or living away from home in the previous months, the type of study, the course of study, the institution where the student is studying, of both parents: all address details and whether the parent is unwilling or deceased, the risk scores and risk factors for the current and upcoming month.

[13] There is an exhaustive list of which students are excluded from the CUB process [H.1].

situations with a long travel time between the home address and the educational institution [H.1 p.10].

The final selection of students for home visits may, based on the above criteria, deviate from the sorted list on risk score from high to low. How many students are selected depends on the capacity agreements that have been made with external research agencies per region. The group selected for a home visit (taken together for all regions) is referred to as the *home visit population*.

> Step 4 – Home visit: External party carries out a home visit, collects data, if possible, about the legality of the claim for a college grant and reports results back to DUO.[14]

> Step 5 – Feedback, processing and next steps: Processing results. Determining lawful and unduly use. The population that has made unduly use of the college grant is referred to as the unduly use population.

> Step 6 – Appeal procedure: The unduly use population can object to DUO's decision. The students who submit an objection are referred to as the *appeal population*.

These steps are shown schematically in Figure 1. The qualitative part of this research focuses exclusively on step 1 – risk profile. The quantitative part of this research focuses on steps 0-6.

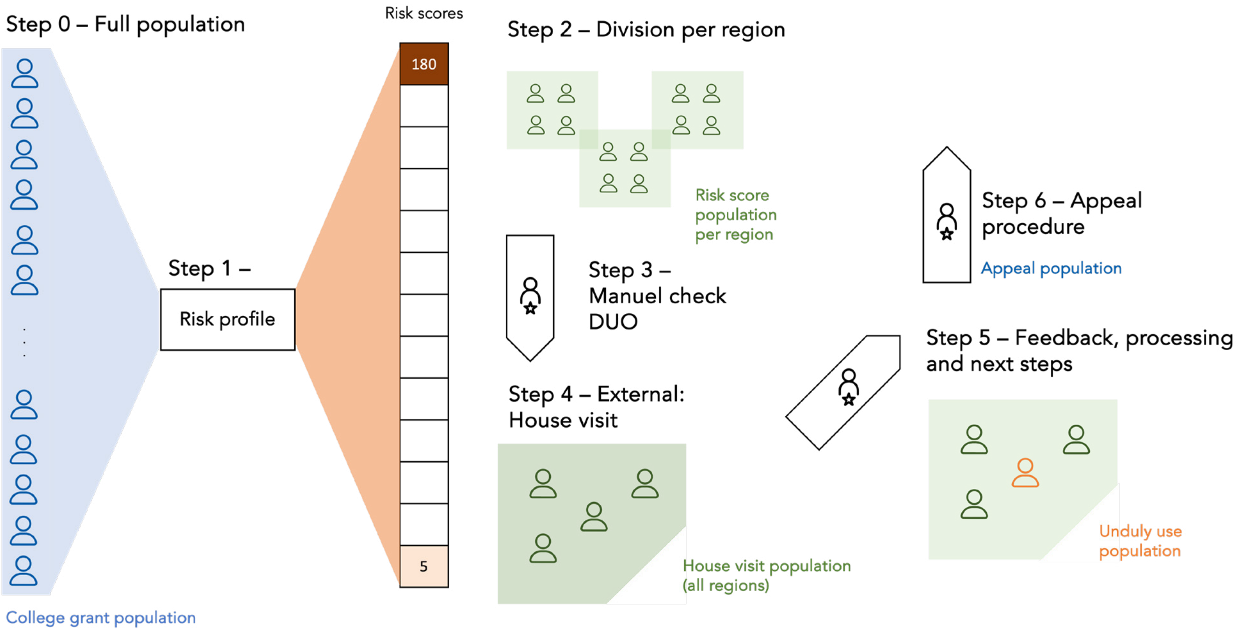

Step 0 – Full population
Risk scores
180
5
Step 1 –
Risk profile
College grant population
Step 2 – Division per region
Risk score population per region
Step 3 – Manuel check DUO
Step 4 – External: House visit
House visit population (all regions)
Step 5 – Feedback, processing and next steps
Unduly use population
Step 6 – Appeal procedure
Appeal population

Figure 1 – Schematic representation of the CUB process.

[14] If a home visit cannot take place, an alternative investigation will take place to determine whether a student from the home visit population belongs to the unduly use population. See also [A.3].

3. Research methodology

Below follow the methodology of the quantitative analysis (§3.1) and the method of the qualitative analysis (§3.2). The aim, research set-up and research questions of both type of analyses are discussed.

3.1 Quantitative analysis

First, the purpose of the quantitative analysis is described, followed by the methodology of this analysis, which is explained based on the random samples of 2014 and 2017 and the bias measurements of 2014 and 2019 and associated research questions.

Purpose of the analysis

The purpose of the quantitative analysis is to trace possible bias regarding the risk profiling criteria 1. type of education, 2. age, 3. distance from parent(s) or 4. migration background, in different steps of the CUB process. An overview of algorithmic actions (steps 1-2) and human actions (steps 3-5) in the CUB process is given in 2.3 Overview of CUB process. By carrying out a measurement per step, targeted follow-up research can be conducted into where possible (indirect) bias occurs in the CUB process. This quantitative analysis is also called a *bias measurement*.

Set-up of the analysis

To investigate possible bias in the CUB process, the bias measurement has been divided into three parts:

A. Testing the suitability of the risk profile for detecting unduly use – A first question about possible bias in the CUB process is whether there is a connection between unduly use of the college grant and the characteristics on which the risk profile differentiates: 1. type of education, 2. age and 3. distance from parent(s). This question can be answered by analyzing the *random samples*. The method of this analysis is explained below in A. Random sample 2014 and 2017.

B. Testing proportionality of the CUB risk profile – If the random sample indicates there is a statistical relationship between unduly use of the college grant and characteristics 1-3, the question is whether differentiation made in the risk profile on the basis of these characteristics is proportional. For example: the random sample may show that for students within a certain type of education the chance of unduly use is 10% greater, but that on average the risk model assigns students of this type of education a risk score which deviates much more than 10% compared to the average of students with other forms of education. This analysis is performed before step 1 (risk profile) and step 3 (manual selection) of the CUB process. An explanation of this proportionality test is given below in the B. Bias measurement 2014 and 2019.

C. Indirect distinction in CUB process regarding migration background – After analysis of the random sample and the proportionality test, the question arises whether distinction based on 1. type of education, 2. age and 3. distance from parent(s) indirectly disadvantages students with a migration background and, if so, in which phase of the CUB process this bias arises. To properly investigate indirect forms of bias, access to sensitive personal data is required, namely data on whether an individual student has a migration background or not. A request has been submitted to the CBS to obtain this type of aggregation statistics for groups larger than 10 people. To date, CBS has not provided this data. In the future, this data may be made available for possible follow-up research. Alternative methods to process data on migration background at the individual level are not available. As a result, it has not yet been possible to perform part C of the quantitative analysis.

It has been considered to use publicly accessible data on migration background aggregated per postal code area from the CBS as an alternative.[15] Algorithm Audit has refrained from this for the following reason: the publicly accessible data is based on the entire population per postal code area. It cannot be determined whether the CBS data is representative of the population of interest: students living on their own. Algorithm Audit believes that there are strong objections to assuming this representativeness. The residential behavior of students, such as housing type and distribution across the country, but also the socio-economic status of students living on their own, is so specific that these characteristics are most likely not to be similar as the Dutch population. This means that when *N* students are selected for a home visit in a postcode area with *x%* residents with a migration background, it is not clear that among these *N* students *x%* have the same migration background. Due to this limitation, no meaningful statements can be made based on this data about how students with a migration background were over- or underrepresented in the CUB process.

A. Random sample 2014 and 2017

The suitability of the risk profile used in the CUB process is examined based on the following three research questions:

Research question 1

Based on the random sample of 2014 and 2017, is there statistical evidence of a relationship between the characteristics used in the risk profile and unduly use of the college grant?

Research question 2

Based on the random sample of 2014 and 2017, is there statistical evidence of a relationship between the used risk categories 1-6 and unduly use of the college grant?

[15] See https://www.cbs.nl/nl-nl/cijfers/detail/83503NED.

Research question 3
How much unduly use has been detected in the random sample population of 2014 and 2017?

The methods used to answer these research questions are explained below.

In a random sample, individuals are randomly selected from the college grant population of students. Randomness is important because it makes it possible to draw conclusions about the entire population. For this reason, a random sample is used to measure the relationship between certain characteristics (for example type of education, age, and distance to parent(s)) and the risk of unduly use of the college grant. Such connections cannot be based on outcomes of the CUB process, as data was not collected from random individuals, but from students selected by DUO for a house visit. This group of students selected in the CUB process is not with certainty representative of the entire (starting) population, and therefore, for this data there is a chance that incorrect conclusions will be drawn regarding the relationship between the characteristics of the risk profile and the chance of unduly use.

In 2014 and 2017, DUO, in collaboration with the ADR, took random samples with the aim of determining the extent of unduly use of the college grant. Details about the samples drawn can be found in [A.40]. These samples can also be used to answer research questions 1-3. The results from the 2014 and 2017 samples are reported separately, because the populations between the two samples have changed due to the introduction of the loan system.

To investigate research questions 1 and 2, the difference in the unduly use rates of different education, age and distance categories is tested. Suppose that the unduly use percentage is for group A p_A and for group B p_B. It is then tested whether the difference between p_A and p_B is statistically significant. This can be used to determine whether students with a certain profile make unduly use of the college grant more often. Formally, the following null hypothesis H_0 en alternatieve hypothese H_1 are tested for a group A and B with unduly use rates p_A en p_B:

$$H_0: p_A = p_B$$
$$H_1: p_A > p_B.$$

p_A can be determined as follows. Let $N_{A,\ unduly}$ be the number of individuals for group A who unduly used the college grant. Let N_A be the total number of individuals belonging to group A in the random sample. Then:

$$p_A = N_{A,\ unduly} / N_A.$$

p_B can be determined in the same way. Then, H_0 is subjected to a one-sided Z test (see Box 2). Given the possible small group size, Fisher's one-sided test is performed as a check (see Box 2). Further explanation of these statistical tests can be found in Appendix C – Additional information statistical analysis.

This statistical test is applied to test all education, age and distance categories against each other. So, for example, for type of education, the following scenarios are subjected to the Z-test:

I. A=mbo 1-2, B=mbo 3-4
II. A=mbo 1-2, B=hbo
III. A=mbo 1-2, B=wo
IV. A=mbo 3-4, B=hbo
V. A=mbo 3-4, B=wo
VI. A=hbo, B=wo.

This step-by-step comparison was chosen because it follows the logic of the CUB risk profile: different forms of education are gradually given a different risk score. Based on Table 3, p_A and p_B for scenarios I-VI are computed.

	Education			
	mbo 1-2	mbo 3-4	hbo	wo
Unduly use rate	$N_{mbo\ 1\text{-}2,\ unduly}$ / $N_{mbo\ 1\text{-}2}$	$N_{mbo\ 3\text{-}4,\ unduly}$ / $N_{mbo\ 3\text{-}4}$	$N_{hbo,\ unduly}$ / N_{hbo}	$N_{wo,\ unduly}$ / N_{wo}

Table 3 – Required information for testing statistical significance between the unduly use percentages across types of educational.

The same method is used for the distinction made based on the age, distance and risk categories used. This analysis answers research questions 1-3. The results are presented in 4.1 Results of random sample 2014 and 2017.

Note that an attempt was made in 2010 to prepare Table 3 based on n=934 studentst [A.47]. However, the results of this analysis are insufficiently carefully documented to be able to be reused or analyzed for this investigationt.[16] See also 5.1 Qualitative analysis of risk profile.

B. Bias measurement 2014 and 2019

The over- or under-reaction of the risk profile (step 1) and manual selection by employees (step 3) regarding characteristics 1. type of education, 2. age, 3. distance from parent(s) is

[16] It is unknown how the 934 students were selected. It cannot be determined whether this was done randomly. Formulas used in the available Excel file have also been omitted, making it impossible to trace how certain figures were arrived at, which complicates the reconstruction of the analysis.

Box 2

Z-test and Fisher's exact test

A Z-test is a statistical test used to compare the proportions (e.g., unduly use rate) between two independent groups. The test is often applied for binary outcomes (for example duly or unduly use). The test is used to assess whether the observed differences in proportions between the groups are statistically significant.

Fisher's exact test is another statistical test that can also be used to compare the proportions (e.g., unduly use rate) between two groups. Fisher's exact test is especially suitable when the sample size is small.

examined using a bias measurement. Two college grant populations are relevant for this bias measurement.

The first college grant population includes students who received a college grant before the introduction of the loan system. This population, consisting of mbo (vocational training), hbo (applied sciences) and wo (university) students, is referred to as the *college grant population-2014*. 01-02-2014 was chosen as the reference date for determining this population. This reference date has been chosen because most CUB studies are completed per calendar year. Results of the CUB process in calendar year 2014 were collected for all students who would receive a grant for living away from home for February 2014 on 01-02-2014. The query that was performed in the SQL database to retrieve the 2014 college grant population is shown in Appendix B – Datawarehouse query.

The second college grant population concerns students who received a college grant after the introduction of the loan system. This population, which consists of mbo students and phasing-out hbo and wo students, is referred to as the *college grant population-2019*. 01-02-2019 has been chosen as the reference date for determining this population. Calendar year 2019 was chosen for two reasons. First, because in 2019 as many hbo and wo students who would still receive a college grant – because they started studying in the period before the basic grant system was abolished – graduated. Secondly, because 2019 is the last year in which CUB control procedure took place in its regularly form before the process was adjusted due to measures as an effect of the covid-19 pandemic in 2020 and 2021.

As explained in 2.3 Overview of CUB process, the college grant population-2014 and -2019 are followed through the various steps in the CUB process. The research questions for the proportionality test are formulated using two core concepts from the literature surrounding fair algorithms that are introduced first.

1. **Demographic parity –** When the action between two steps in the CUB process is independent of 1. type of education, 2. age or 3. distance from parent(s). For example: the average score assigned by the risk profile to students with education form A is the same as the average score assigned to students with education form B. In the case of step 1 - risk profile per type of education implies:

average(risk score | education) = average(risk score).

For step 3 – chance of manual selection by employee per type of educational implies:

p(selection for home visit | education) = p(selection for home visit).

2. **Conditional demographic parity –** Also referred to as *equalized odds*. When the action between two steps in the CUB process is independent of 1. type of education, 2. age or 3. distance from parent(s), given that the college grant has unduly been used. For example: the average score assigned by the risk profile students with type of education A compared to the degree of unduly use in the subpopulation of students with type of education B is the same as the average score assigned to students with type of education B compared to the degree of unduly use in the subpopulation of students with type of education B. In the case of step 1 – risk profile per type of education implies that:

   average(risk score | education, unduly use) =
   average(risk score | unduly use).

   Note that measuring conditional parity for step 3 – manual selection by employee is not useful, as the probability of being selected by an employee given that a college loan has unduly been used is always 1.

Demographic parity and conditional demographic parity are both measured to investigate possible bias in the CUB process in the most complete way possible. The following research questions are formulated based on these concepts.

Research question 4
Has demographic parity been satisfied in the risk score population (after step 1 – risk profile) and the home visit population (after step 3 – manual selection by employee)?

Note: Research question 4 attempts to measure whether deviating patterns occur when assigning risk scores based on the risk profile or when manually selecting students for a home visit by employees.

Research question 5
Has conditional demographic parity been satisfied in the risk score population (after step 1 – risk profile)?

Note: Research question 5 attempts to measure which risk scores have been assigned to students who were found to have unlawfully used the college grant. Based on this, gaps in the risk model can be found.

Research question 6
How much unduly use is observed in the population selected by the CUB process in 2014 and 2019?

Building on the notation introduced in A. random sample 2014 and 2017 sub-questions 4-6 regarding the type of education can be answered by Table 4.

Note that $N_{\textit{college grant population-X}} = N_{\textit{risk score population-X}}$ for X = 2014 en 2019 since every student who receives a college grant receives as well a risk score.

#	Selection for home visit	Type of education	Unduly use
1	Yes	A	Yes
2	No	B	No
…			
N	Yes	C	No

Table 4 – Information required to determine the (conditional) demographic parity of step 3 – manual selection of students. To determine the (conditional) demographic parity for step 1 – risk profile, the assigned risk score per student (#) is also required.

Based on the data from Table 4 p(selection home visit), p(selection home visit l education) and p(selection home visit l education, unduly) can be determined:

p(selection home visit) = $N_{\textit{selection for home visit = Yes}} / N$

p(selection home visit l education) =

$$N_{\textit{selection for house visit = Yes, education = A}} / N_{\textit{ education = A}}$$

p(selection home visit l education, unduly use) =

$$N_{\textit{selection home visit = Yes, education = A, unduly = Yes}} / N_{\textit{ education = A, unduly = Yes}}.$$

The (conditional) demographic parity for step 1 – risk profile can be determined by determining the average risk score for the above groups.

Limitations of chosen research methodology

There are some limitations to the chosen research methodology to quantitatively measure bias in the CUB process.

Firstly, by only analyzing the selected data from DUO's data warehouse (1. type of education, 2. age, and 3. distance from parent(s)), bias of the CUB process regarding migration background can only be investigated to a very limited extent. Deviations in the data on the CUB process can only indicate a possible connection with bias regarding migration background. This connection cannot be confirmed.

The bias measurement only measures deviations regarding the characteristics used in the risk profile (type of education, age, and distance to parent(s)). Other variables that are not included in the risk profile but may have predictive value for detecting unduly use of the college grant, are not involved in this study.

3.2 Qualitative analysis

Goal of the analysis

In the qualitative analysis, the suitability of the risk profile in the CUB process is tested against the qualitative standards that apply at the time of publication of this report. The aim is to reflect on the risk profile and to provide DUO with tools for effectively combating unduly use of the college grant in the future.

Analysis set-up

The research questions for the qualitative analysis are:

Research question 7

Did the risk profile used in the CUB process meet the standards that are currently set for algorithms used by Dutch public sector organisations?

Research question 8

If the answer to research question 7 is negative, what is needed to use the risk profile responsibly in the future?

To answer these research questions, the IRAC method is used – an acronym for *issue, rule, analysis, conclusion*.

Issue. First, the research questions and framework of the study were determined. Subsequently, there is actual research about the risk profile done. See 2.2 Chronology CUB.

Rule. This report does not aim to provide a normative or legal answer to the question of whether DUO acted right or wrong. Nevertheless, Algorithm Audit considers it relevant to assess the risk profile according to normative standards. These standards are explained below.

Analysis. The analysis was primarily carried out by comparing the established facts to the standards found. Where the established facts do not meet the standards, this is noted and explained where possible.

Conclusion. Findings follow from the analysis. These are listed in the conclusion in combination with recommendations for next steps.

Motivation for research approach

The CUB process is no longer used at the time of publishing this report. The purpose of this report is to assist DUO in its reconsideration of the use of risk profiling. Particularly, in mitigating unduly use of the college grant. To this end, it is important to determine how the CUB process was established and used and to analyze those findings (see also 2. Background College Grant Control procedure: fact reconstruction).

The qualitative analysis is done based on the standards that apply at the time of publication of this report. After all, these are the standards that are relevant for new policy. In particular, for the risk profiles used in the future or for stopping risk profiling.

The downside of this method is that no normative conclusions can be drawn from the qualitative analysis made in this report about the use of the CUB process in the past. To achieve this, the CUB process would have to be based on the standards that applied at the time the CUB process was used. That is the relevant test for the question of whether DUO has been guilty of something in the past. Answering the question of guilt is expressly not the purpose of this report and the report should not be read or used as such. The possible Phase 2 research is suitable for a normative judgment. The option has been agreed with DUO to continue in Phase 2 after publication, in which a normative judgment can be made about the CUB process (and possible future processes) based on the findings of this report.

Standards and guidelines

Various standards and guidelines apply to the use of risk profiling by organizations. In Box 3 an overview is given.

The standards and guidelines are divided into legal standards that DUO must adhere to when detecting unduly use, *soft law*, and internal standards.

Legal standards

In the context of this report, three categories of legal standards are relevant: the general standards for equal treatment, the General Principles of Good Administration as included in the General Administrative Law Act (Algemene wet bestuursrecht – Awb) and standards from the General Data Protection Regulation (GDPR). These three categories are briefly discussed below.

Firstly, the general standards for equal treatment from Dutch legislation and international instruments apply.[17] Article 1 of the Constitution states the prohibition of discrimination. The General Equal Treatment Act (Algemene Wet Gelijke Behandeling – AWGB) provides

[17] First, there are the prohibitions on discrimination under Article 14 ECHR and Art. 1 twelfth ECHR protocol. In addition, there are several equal treatment directives under Union law, which have been implemented in Dutch legislation, namely the Anti-Racial Discrimination Directive (2000/43/EC), the Framework Directive on Equal Treatment in Employment and Occupation (2000/78/EC), the Equal Treatment Directive gender in access to and supply of goods and services (2004/113/EC), and the recast Equal Treatment for Men and Women Directive (2006/54/EC).

substance to this. For example, Article 7a paragraph 1 AWGB stipulates that discrimination based on race is prohibited in social protection (including the awarding of the college grant). Internationally, Article 14 of the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and Article 1 of Protocol 12 to the ECHR apply.

The General Court of Appeal (Centrale Raad van Beroep – CRvB) has commented on risk profiling in the context of the above-mentioned instruments.[18] The CRvB allows risk profiling under certain conditions to increase the effectiveness, efficiency, and cost savings of government actions and because of the importance of combating incorrect use of public allowances. This case law of the CRvB concerned the foreign assets of welfare recipients. The CRvB considered:

> *As the Council considered in those rulings, experiences with groups of welfare recipients and criteria that can objectively form a risk profile for unreported assets abroad may justify the use of the general investigative power regarding welfare recipients of non-Dutch origin, a certain age and a certain holiday behavior, and not with regard to other welfare recipients. However, if such a risk profile, as is currently in dispute, is not aimed at all welfare recipients of non-Dutch origin but is only aimed at welfare recipients from a specific country of birth, then this may involve a distinction that, according to the Council in its case law mentioned in 4.5, is regarded as "suspicious". Such a distinction must be justified by very important reasons.*

Secondly, the principles of good administration as part of Dutch Public Administrative Law apply to the use of risk profiling.[19] The principle of equality and the duty of care are particularly important for risk profiling.

> Principle of equality – Equal cases should be treated equally. Unequal cases should be treated unequally according to the extent to which they differ. The principle of equality can be compromised in risk profiling, among other things, if the profiling leads to (indirect) discrimination.

> Duty of care – This principle aims to create the circumstances in which an administrative body can make a correct decision. An administrative body must inform itself of

[18] ECLI:NL:CRVB:2015:3249, ECLI:NL:CRVB:2018:1541, and ECLI:NL:CRVB:2020:1664 as applied in lower case law, see, among others, ECLI:NL:RBLIM:2013:11417, ECLI :NL:RBOBR:2013:BZ6037, ECLI:NL:RBROT:2014:5684, ECLI:NL:RBROT:2013:2359, ECLI:NL:RBAMS:2012:BV6364, ECLI:NL:RBAMS:2018:9659 and ECLI :NL:RBLIM:2023:1325.

[19] These principles are the prohibition of bias (Article 2:4 General Administrative Law Act), the principle of due care (Article 3:2 General Administrative Law Act), the prohibition of arbitrariness (Article 3:4 paragraph 1 General Administrative Law Act), the principle of proportionality (Article 3:4 paragraph 2 General Administrative Law Act) , the prohibition of *détournement de pouvoir* (Article 3:3 General Administrative Law Act) and the principle of motivation (Articles 3:46 and 3:47 General Administrative Law Act). Applicable standards from unwritten administrative law are the principle of trust, the principle of equality, and the principle of legal certainty. See, for example, Barkhuysen et al., 'Administrative law in the AWB era', p. 110 e.v.

the relevant facts and interests to be weighed (Article 3:2 Dutch Public Administrative Law). An appropriate method of balancing interests must be used, and a full balance of interests must be carried out.[20] The duty of care in risk profiling can be compromised, among other things, if the data used is incomplete or incorrect and if the risk profile does not include all relevant facts or interests.

In 2018, the Advisory Division of the Council of State recommended a stricter interpretation of the principles of sound administration, and in particular the principle to give reason and the duty of care, in automated decision-making by public sector organisations.[21] According to this advice, this could mean, among other things, explaining to citizens which decision rules have been used. When using profiling, the advice also explicitly states that there is a chance that citizens will be confronted with a "*statistical reality that deviates from the concrete facts*".

The SyRI ruling of the District Court of The Hague is also relevant to the exchange of data between government agencies to combat unduly use.[22]

Thirdly, for the sake of completeness, it is noted that the GDPR also contains relevant standards for the risk profile. The data subject whose data is processed has the right to be informed about the existence of (semi-)automated decision-making, such as risk profi-

[20] Chris Adriaansz, The legality of algorithmic decision-making in the light of the duty of care and the principle to give reason, NTB 2020/100.

[21] Parliamentary Papers II 2017/2018, 26 643, 557.

[22] District Court The Hague February 20, 2020, ECLI:NL:RBDHA:2020:865.

Box 3

Standards, legal framework and algorithm defintion

Standards

In this report, the following rules, guidelines, etc. are referred to as 'standards'. To be clear, this is not just about enforceable standards. These are standards in the sense of rules of conduct that are more or less taken for granted by the members of a community and against which behavior can be assessed. When establishing the standards, it is assumed that the law, assessment frameworks drawn up by the government and the EU and guidance from science and the Netherlands Institute of Human Rights (College voor de Rechten van de Mens) contain the standards that are taken for granted in the community.

Legal framework

A student must live on their own to be eligible for the college grant (Article 1.5 of the Student Finance Act 2000 Wet studiefinanciering – Wsf). Supervision of this obligation has been assigned to the Enforcement and Inspection department of DUO (Article 9.1a WSF and Decree on designation of persons to monitor compliance ex Article 1.5 WSF). A student who does not live on their own, but has applied for a college grant, may be fined. Too much student financing received can be reclaimed (Article 9.9 Wsf).

Algorithm definition

In this study, the risk profile is classified as an algorithm. The risk profile meets the definition of an algorithm used by, among others, the Netherlands Court of Auditors (Algemene Rekenkamer) and the Algorithm Register (het Algoritmeregister): “A set of rules and instructions that a computer follows automatically when making calculations to solve a problem or answer a question”. See for example: https://algoritmes.overheid.nl/nl/footer/over-algoritmes.

ling, and "useful information about the logic involved", and the importance and expected consequences of this processing for the data subject (Article 15 GDPR paragraph 1(h)). However, assessment of the CUB profile against the GDPR falls outside the scope of this report. Article 22 of the GDPR prohibits fully automated decision-making. This is not at stake within the CUB process.

Instruments of soft law

There are frameworks for the responsible use of algorithms by public sector organizations. These frameworks provide guidance on issues regarding equal treatment when using algorithms. Authoritative frameworks in the Netherlands are:

- The Government-wide Implementation Framework for *Responsible Use of Algorithms*[23] (2023);
- The Algorithms Research Framework[24] (2023) of the ADR ('ADR Research Framework');
- The Algorithms Assessment Framework from the Dutch Court of Auditors;
- The *Fundamental Rights Algorithms Impact Assessment Algoritmes*[25] (FRAIA) of the Ministry of the Interior and Kingdom Relations (2021);
- The *Insructions for Non-Discrimination by Design*[26] (2021) from the Ministry of the Interior and Kingdom Relations.

The Implementation Framework 'Responsible use of algorithms' deserves special attention. This document is not used in the analysis because a final version is not yet available. Once the Implementation Framework has been published in final form, it will be useful in the use of algorithms by implementation organizations. For now, the following quote suffices, which also applies to the other *soft law* standards:

> *"The implementation framework is not a 'checklist'. Taking measures in this context often does not directly lead to compliance with an obligation and not all measures in this context are mandatory. Users of this implementation framework will therefore make their own decision. Where necessary, they will also have to ask for additional advice, for example in the event of value tensions and ethical dilemmas."*

Furthermore, relevant principles are contained in frameworks drawn up outside the Dutch government. These are, among others:

- The *Assessment Framework for Discrimination through Risk Profiles* of the Netherlands Institute for Human Rights (2021);
- The report '*De mens in de Machine*' by WAAG (February 2020);
- MJ Vetzo, JH Gerards and R. Nehmelman, '*Algoritmes en Grondrechten*' (2018);
- The European Commission (2019), *High-level Expert Group on Artificial Intelligence,*

[23] https://www.rijksoverheid.nl/documenten/rapporten/2023/06/30/implementatiekader-verantwoorde-inzet-van-algoritmen

[24] https://www.rijksoverheid.nl/documenten/rapporten/2023/07/11/onderzoekskader-algoritmes-adr-2023

[25] https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes

[26] https://www.rijksoverheid.nl/documenten/rapporten/2021/06/10/handreiking-non-discriminatie-by-design

Ethics Guidelines for Trustworthy AI;
- European Commission (2020), *White Paper on Artificial Intelligence: a European approach to excellence and trust*.

The Netherlands Court of Auditors and the ADR largely based their recommendations on the latter two documents.

Internal standards

It has been investigated whether DUO has internal standards for the use of risk profiles, for investigating possible bias and/or for the further implementation of administrative law standards. There has been no evidence that these standards exist.

It is therefore recommended to draw up more concrete standards for the use of risk profiling within DUO. The ADR Research Framework also prescribes this: "*the algorithm complies with the internally established Policy Frameworks*" (SV.8) and "*the roles and responsibilities in the development and use of the algorithm have been assigned*" (SV.9).

There are methods for drawing up such a normative framework. These methods fall outside the scope of this research, but further advice can be provided if necessary.

Selection of standards against which the risk profile is tested in this study

Not all standards mentioned above are suitable for answering the research question. The legal standards provide a relevant outer limit but are not sufficiently specific to be relevant for an analysis of the risk profile. Since internal standards are lacking, no guidance is found here either.

Algorithm Audit considers the Dutch FRAIA and the Algorithms Research Framework to be the most suitable standards for this research. These frameworks have been specifically drawn up for responsible use of algorithms by public sector organizations. These frameworks will be used for the qualitative analysis. Specific reference will be made to Chapter 1 Control & Accountability (Sturing & Verantwoording – SV) and Chapter 3 Data & Model (DM) from the Algorithms Research Framework.

3.3 Composition of research team

The research was conducted by researchers affiliated with Algorithm Audit. Diversity of the research team has been taken into account, among others by involving researchers with different gender, cultural backgrounds, and professional expertise. The research team was diverse and multidisciplinary.

4. Results of quantitative analysis

Below the results are presented of the analysis of the random samples from 2014 and 2017 (§4.1) and the results of the bias measurements from 2014 and 2019 (§4.2). Research questions 1-6 are answered based on these results.

4.1 Results of random sample 2014 and 2017

In 2014 and 2017, 387 and 293 students with a college grant were randomly selected respectively for a control procedure [A.15,A.40]. These students were examined to determine whether they duly received the college grant. In consultation with the ADR, DUO has computed the size of the sample population to determine unduly use of the college grant with 95% certainty in the full population. The size of the entire student population, consisting of mbo, hboand wo students, at the time the random sample took place in March 2014 was 248.649 [A.40]. The size of the entire student population, consisting only of mbo students, at the time the random sample took place in October 2017 was 50.233 [A.15].

The files with the random samples to which the researchers were given access consists of 387 and 293 students for the years 2014 and 2017 respectively [G.1,G.2].[27] The results from these samples were used to answer research questions 1-3. The relevant characteristics for the risk model (type of education, age, and distance to parent(s)) have been coded for all these students.

Results of research question 1

Research question 1:
Based on the random sample of 2014 and 2017, is there statistical evidence of a relationship between the characteristics used in the risk profile and unduly use of the college grant?

Results research question 1:
The random samples from 2014 and 2017 show no evidence for a relationship between most of the characteristics used for the risk profile and unduly use of the college grant. There is a clear indication that, based on the random sample from 2014, students who live 1m-1km and 2-5km from their parent(s) more often and students who live 50-500 km from their parent(s) less often make unduly use of the college grant. In addition, the random sample from 2014 shows that mbo 3-4 students (vocational training) made unduly use of the college grant statistically significantly more often than wo students (university). The random sample from 2017 shows that students within age category 15-18 are statistically significantly more likely to unduly use the college grant than students within age category 21-22. However, these results do not support the granular differences used in the risk profile.

[27] The difference with the numbers stated in [A.15] and [A.40] cannot be explained.

Motivation research question 1:
Research question 1 is answered per characteristic of educational form, age and distance to parent(s).

As explained in 3.1 Quantitative analysis two random samples are used: the 2014 (n=387) and 2017 (n=293) samples, the latter of which only includes students having mbo type of education. This is because at the time of the sample in 2017, the college grant was awarded exclusively to mbo students.

Type of education

Tables 5-6 show the unduly use percentage per type of educational for students drawn for the random sample from 2014 and 2017.

Sample 2014	Group size	# unduly grants	Percentage
mbo 1-2	15	1	6,7%
mbo 3-4	53	4	7,5%
hbo	150	5	3,3%
wo	169	4	2,4%
Total	387	14	3,6%

Table 5 – Overview of group size and unduly use percentage per type of education in the random sample 2014 (n=387).

Sample 2017	Group size	# unduly grants	Percentage
mbo 1-2	53	4	7,5%
mbo 3-4	240	7	2,9%
Total	293	11	3,8%

Table 6 – Overview of group size and unduly use percentage per type of education in the random sample 2017 (n=293).

The results of the one-sided statistical Z-test for the different unduly use percentages per educational form are shown in Figure 2. A description of the statistical tests performed can be found in 3.1 Quantitative analysis.

Figure 2 shows that only the difference between unduly use percentages of mbo 3-4 and wo students is statistically significant. Algorithm Audit considers this to be an insufficiently consistent statistical signal on which risk profiling can be based. This means that, on the basis of the 2014 and 2017 random samples, there is no statistical basis for making a distinction based on the types of education mbo, hbo and wo. From the 2014 random sample, the pattern is observable that the % of unduly use decreases as the level of education 'increases'. However, these differences are not statistically significant. Note that the relatively limited size of subpopulations in the 2014 sample, for example, 15 selected mbo 1-2 students, may cloud the results. A larger sample is needed to confirm whether there are indications of differences in unduly use percentage per type of education.

Fisher's exact test – a method suitable for small sample sizes – confirms the above results. Additional information about Fisher's exact test is provided in Appendix C – Additional information statistical analysis.

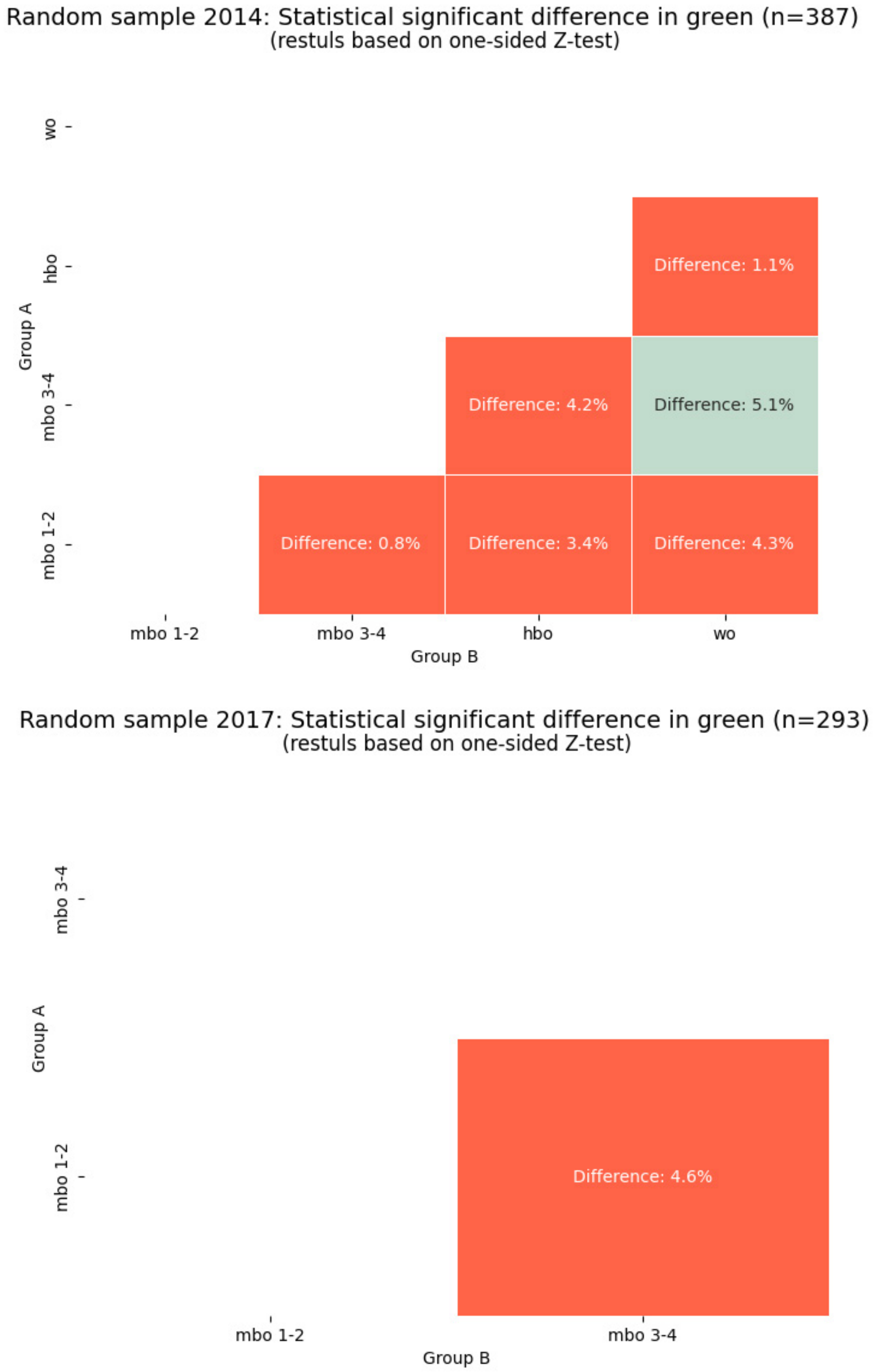


Figure 2 – Results of one-sided Z-tests, based on a significane level of 5%, for difference in unduly use percentage per educational form (Group A vs Group B) based on the random sample from 2014 (n=387) and the random sample from 2017 (n=293). Statistically significant distinction between two unduly use percentages per education category in green. Statistically insignificant distinction between two unduly use percentages per education category in red. A positive difference means that Group A has a higher unduly use percentage than Group B. A negative difference means that Group B has a higher unduly use percentage than Group A.

Age

Table 7-8 show the unduly use rate by age category for students drawn for the 2014 and 2017 random sample.

Sample 2014	Size of group	# unduly grants	Percentage
15-18	24	0	0%
19-20	115	1	0,9%
21-22	149	8	5,4%
23-24	62	3	4,8%
25-50	37	2	5,4%
Totaal	387	14	3,6%

Table 7 – Overview of group size and unduly use percentage per age category in the random sample 2014 (n=387).

Sample 2017	Size of group	# unduly grants	Percentage
15-18	24	2	8,3%
19-20	105	5	4,8%
21-22	76	1	1,3%
23-24	43	2	4,7%
25-50	45	1	2,2%
Totaal	293	11	3,8%

Table 8 – Overview of group size and unduly use percentage per age category in the random sample 2017 (n=293).

The results of the one-sided statistical Z-test for the different unduly use percentages per age category are shown in Figure 3. A description of the statistical tests performed can be found in 3.1 Quantitative analysis.

Figure 3 shows that only the difference between unduly use percentages of 15-18 year-olds and 21-22 year-olds in 2017 is statistically significant. Algorithm Audit considers this to be an insufficiently consistent statistical signal on which risk profiling should be based. The percentages from Table 7 even point to a pattern that goes against the logic of the CUB risk profile. A higher unduly use rate is observed for older students. While the CUB profile assumes that young students have a higher risk of making unduly use of the college grant. The logic of the risk profile is reflected in the results of the random sample from 2017 (Table 8). Note that the 2017 sample only concerns mbo students, and it is therefore not clear whether this finding generalizes to hbo and wo students.

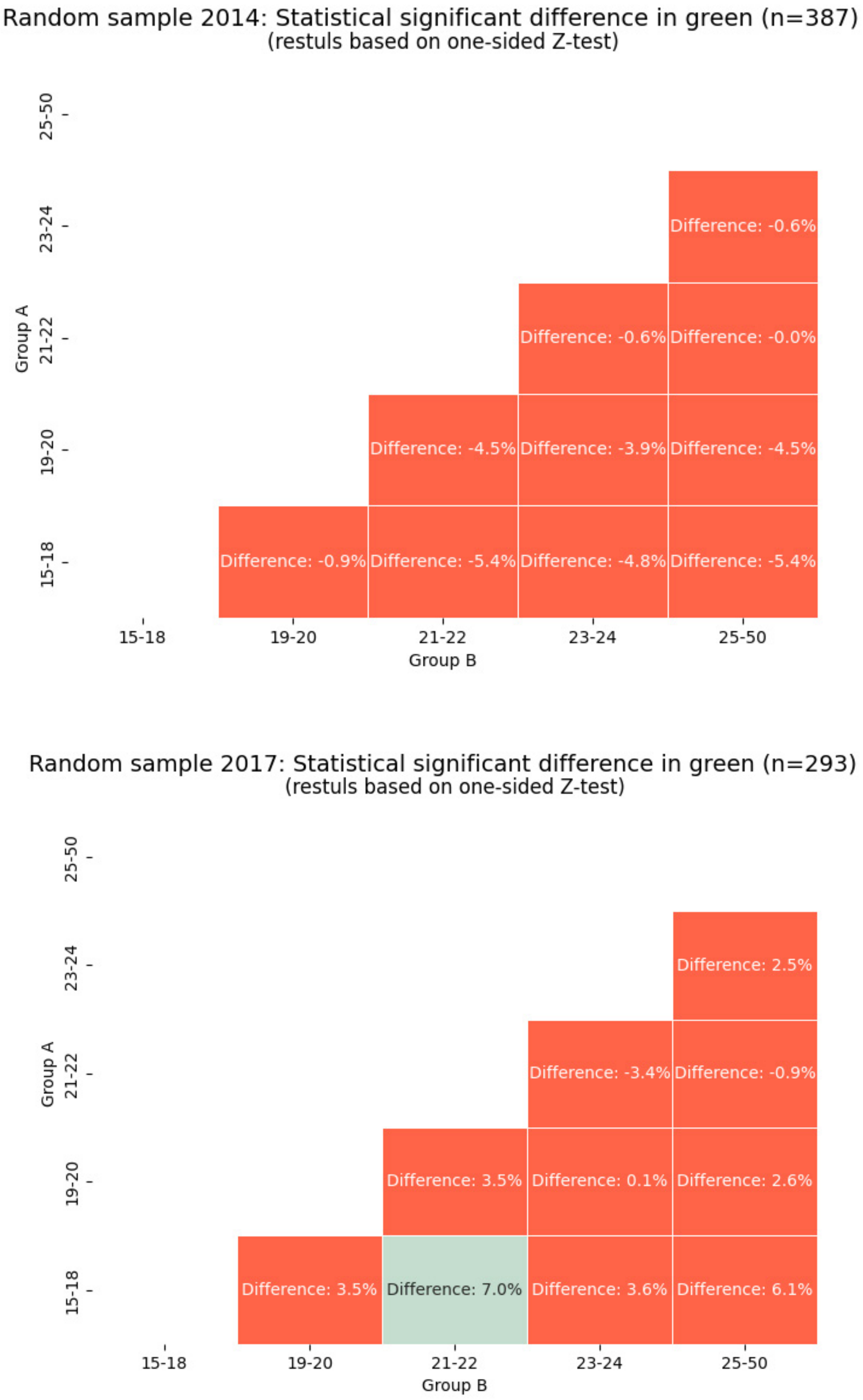


Figure 3 – Results of one-sided Z-tests, based on a significane level of 5%, for difference in unduly use percentage per age category (Group A vs Group B) based on the random sample from 2014 (n=387) and the random sample from 2017 (n=293). Statistically significant distinction between two unduly use percentages per age category in green. Statistically insignificant distinction between two unduly use percentages per education category in red. A positive difference means that Group A has a higher unduly use percentage than Group B. A negative difference means that Group B has a higher unduly use percentage than Group A.

Distance to parent(s)

Table 9-10 show the unduly use rate by distance category for students drawn for the 2014 and 2017 random samples.

Sample 2014	Size of group	# unduly grants	Percentage
0km	8	0	0%
1m-1km	21	5	23,8%
1-2km	11	0	0%
2-5km	31	5	16,1%
5-10km	24	1	4,2%
10-20km	31	1	3,2%
20-50km	58	1	1,7%
50-500km	137	0	0%
onbekend	66	1	1,5%
Totaal	387	14	3,6%

Table 9 – Overview of group size and unduly use percentage per distance category in the 2014 random sample (n=387).

Sample 2017	Size of group	# unduly grants	Percentage
0km	4	0	0%
1m-1km	24	0	0%
1-2km	24	0	0%
2-5km	40	1	2,5%
5-10km	29	1	3,4%
10-20km	38	4	10,5%
20-50km	28	1	3,6%
50-500km	45	1	2,2%
onbekend	61	3	4,9%
Totaal	293	11	3,8%

Table 10 – Overview of group size and unduly use percentage per distance category in the 2017 random sample (n=293).

The results of the one-sided statistical Z-tests for the different unduly use percentages per distance category are shown in Figure 4. A description of the statistical tests performed can be found in 3.1 Quantitative analysis.

Figure 4 shows that there are many statistically significant differences between the distance categories used based on the 2014 sample. There is sufficient statistical support for a distinction based on 1m-1km, 2-5km and 50-500km. However, the clear differences that follow from the 2014 sample are not reflected in the 2017 sample. For the random sample from 2017, none of the differences in unduly use percentage per distance category are statistically significant. These results indicate that the statistically significant signals found in the random sample from 2014 cannot simply be generalized to other years.

A possible explanation for the difference between 2014 and 2017 is that the 2017 sample almost exclusively contains students from mbo education. To verify this, the relationship between distance from parent(s) and unduly use percentage for mbo students in the 2014 sample is checked. This shows that the statistically significant differences found in the entire 2014 sample no longer apply when only the mbo students selected for the 2014 sample are analyzed. Based on the random sample from 2014, the predictive value of the 1m-1km, 2-5km and 50-500km distance categories only seems to be relevant for hbo and wo students.

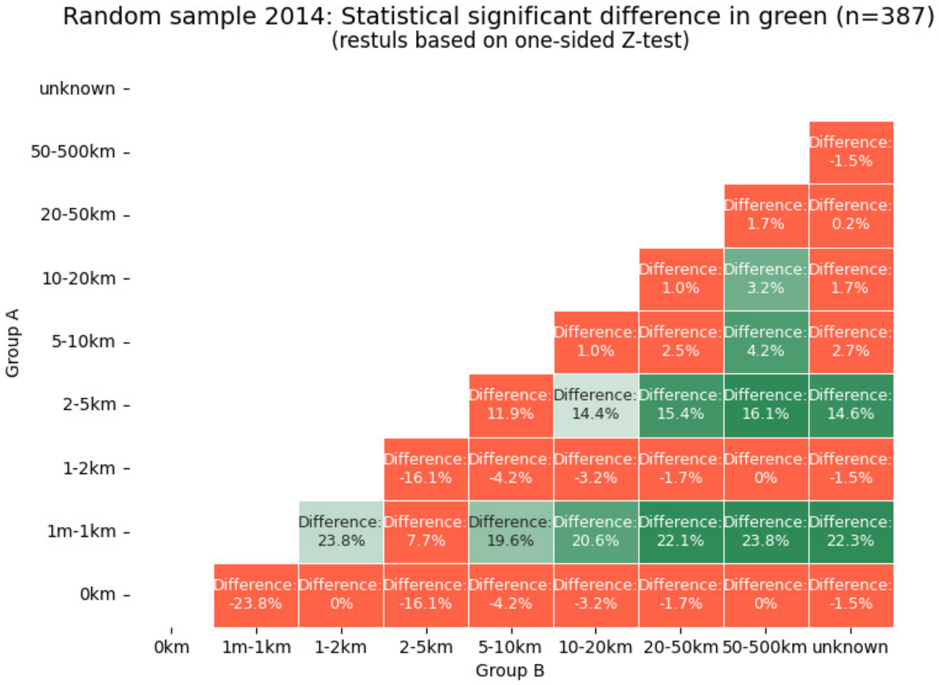
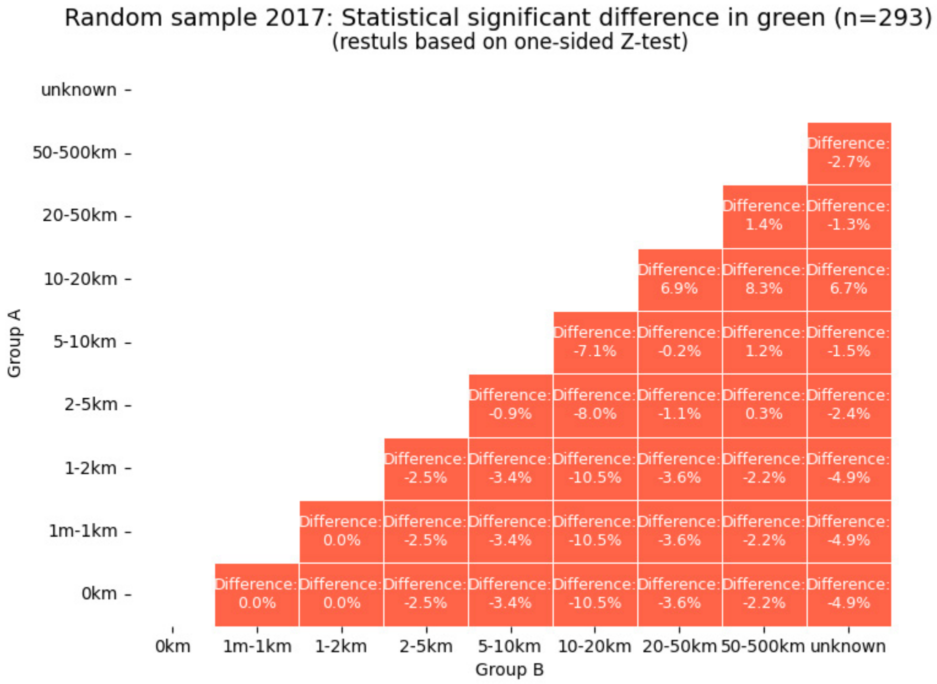


Figure 4 – Results of one-sided Z-tests, based on a significane level of 5%, for difference in unduly use percentage per distance category (Group A vs Group B) based on the random sample 2014 (n=387) and the random sample 2017 (n=293). Statistically significant distinction between two unduly use percentages per distance category in green. Statistically insignificant distinction between two unduly use percentages per education category in red. A positive difference means that Group A has a higher unduly use percentage than Group B. A negative difference means that Group B has a higher unduly use percentage than Group A.



Results of research question 2

Research question 2:
Based on the random sample of 2014 and 2017, is there statistical evidence of a relationship between the used risk categories 1-6 and unduly use of the college grant?

Answer research question 2:
There is insufficient statistical evidence to divide the assigned risk scores into six risk categories. There is statistical support for a binary risk classification, given it is applied to hbo and wo students. A simplification of the risk classes used is preferable.

Motivation research question 2:
The risk of unduly use of the college grant has been classified into different 'risk groups', ranging from 'Very high (1)' to 'Very low (5)', or 'Unknown (6)'. In 2014, a sample was taken to '*make a statement about the unduly use in the lower risk categories 3 to 6 as a whole*' [A.40]. Based on this analysis, it was concluded at the time: '*Based on the results in categories 3 to 6, it seems that DUO rightly pays less attention to these groups*'.

Based on the 2014 random sample, this conclusion is partly supported. As shown in Table 11, the unduly use percentage is higher for the group 'Very high (1) and 'High (2)' than for the group 'Medium (3)', 'Low (4)' and 'Very low (5)', possibly supplemented with 'Unknown (6)'. The one-sided Z-test, the application of which is explained in 3.1 Quantitative analysis, considers the difference between the group 1+2 and the group 3+4+5 plus possibly 6 to be statistically significant (see the green percentages in Table 11). However, the difference in unduly use percentages between individual groups 1-6 is not statistically significant (see red percentages in Table 11). In summary, the 2014 and 2017 random samples provide statistical support for a binary high and low risk classification. However, there are no clear reasons for choosing risk categories 1-6, as used in the CUB process.

Group (random sample 2014)	% unduly use	# unduly grants	Total number
1 - Very high	10,0%	2	20
2 - High	17,2%	5	29
3 - Medium	13,2%	5	38
4 - Low	0,0%	0	106
5 - Very low	0,8%	1	125
6 - Unknown	1,4%	1	69
Very high (1) + High (2)	14,3%	7	49
Medium (3) + Low (4) + Very low (5)	2,2%	6	269
Medium (3) + Low (4) + Very low (5) + Unknown (6)	2,1%	7	338
Difference Group 1 + 2 & Group 3 + 4 + 5	**12,1%**	-	-
Difference Group 1 + 2 & Group 3 + 4 + 5 + 6	**12,2%**		
Difference Very high (1) & high (2)	**-7,2%**	-	-
Difference High (2) & Medium (3)	**4,1%**	-	-
Difference Medium (3) & Low (4)	**13,2%**	-	-
Difference Low (4) & Very low (5)	**-0,8%**	-	-

Table 11 – Analysis of risk categories used in the CUB process based on the results of the random sample from 2014 (n=387). All differences in green are statistically significant according to a one-sided Z-test, based on a 95% confidence level. All differences in red are statistically insignificant based on a significane level of 5%.

The random sample from 2017 provides a different view on the risk categories than the above analysis based on the random sample from 2014. As shown in Table 12, there are no statistically significant differences between the groups. In fact, the percentage of unduly use goes against the logic of the CUB profile – on several occasions the unduly use percentage is higher for lower risk categories. It appears that the risk profile is not suitable for the population studied based on the random sample from 2017, which consists exclusively of mbo students. A new, larger random sample should provide a definitive answer to whether, and if so how, mbo students can be subjected to risk profiling in a statistically substantiated manner.

Group (random sample 2017)	% unduly use	# unduly grants	Total number
1 - Very high	3,4%	2	59
2 - High	0,0%	0	52
3 - Medium	6,9%	2	29
4 - Low	3,3%	2	60
5 - Very low	4,1%	2	49
6 - Unknown	6,8%	3	44
Very high (1) + High (2)	1,8%	2	111
Medium (3) + Low (4) + Very low (5)	4,3%	6	138
Medium (3) + Low (4) + Very low (5) + Unknown (6)	4,9%	9	182
Difference Group 1 + 2 & Group 3 + 4 + 5	**-2,5%**	-	-
Difference Group 1 + 2 & Group 3 + 4 + 5 + 6	**-3,1%**		
Difference Very high (1) & high (2)	**3,4%**	-	-
Difference High (2) & Medium (3)	**-6,9%**	-	-
Difference Medium (3) & Low (4)	**3,6%**	-	-
Difference Low (4) & Very low (5)	**-0,7%**	-	-

Table 12 – Analysis of risk categories used in the CUB process based on the results of the random sample from 2017 (n=293). All differences in green are statistically significant according to a one-sided Z-test, based on a 95% confidence level. All differences in red are statistically insignificant based on a significane level of 5%.

Results of research question 3

Research question 3:
How much unduly use has been detected in the random sample population of 2014 and 2017?

Answer research question 3:
Table 5-6 show that, based on the random sample from 2014, there is an unduly use percentage of 3.6%. For the random sample from 2017, an unduly use percentage of 3.8% follows.

4.2 Results of bias measurement 2014 and 2019

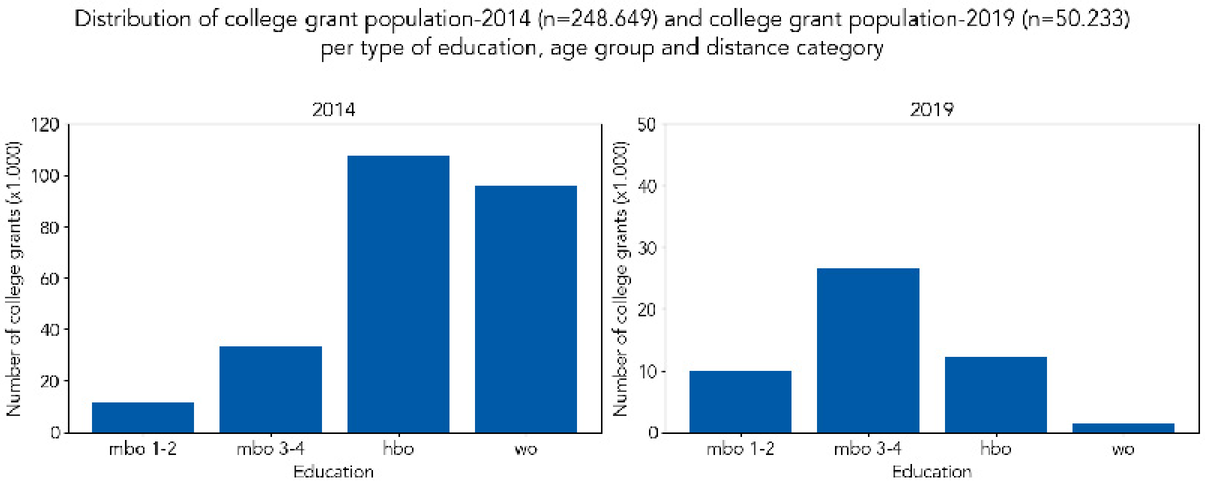
As in 3.1 Quantitative analysis has been explained, the bias measurement is carried out for the college grant population-2014 and -2019. The characteristics of these populations are first presented. Thereafter, it is discussed how the data on the college grant populations were obtained from the DUO data warehouse. The section concludes with answers to research questions 4-6.

Data exploration of college grant population-2014 and -2019

The size of the 2014 college grant population, determined based on the reference date 01-02-2014, is 248.649. The size of the 2019 college grant population, determined based on the reference date 01-02-2019, is 50.233. These populations have been traced in the DUO data warehouse [G.3,G.4]. An example of a query to track these populations is given in Appendix B – Datawarehouse query. This section explores the properties of these populations

The distribution of the number of students with a college grant per type of education, age category and distance category in the college grant population-2014 and the college grant population-2019 is shown in Figure 5. Note that in 2014, all wo, hbo and mbo students living away from their parent(s) home are entitled to a college grant. This is different in 2019 due to the introduction of the college loan system. That year, only mbo students are entitled to a college grant. Hbo and wo students who still receive a college grant in 2019 are phasing-out students who already applied for a college grant before the introduction of the loan system in 2015. In 2019, almost exclusively mbo students were selected for home visits in the CUB process.

In October 2014, 479.800, 446.434 and 255.661[28] students were registered at mbo, hbo and wo education institutions respectively. In February 2014, approximately 15%, 35% and 40% of these students received a college grant. These figures largely correspond with the National Student Housing Monitor 2014.[29]
Note that the distance category 'unknown' is relevant for students for whom the address of the parent(s) is unknown, the parent(s) are unknown, are deceased or where a hardship clause applies.[11]



[28] Netherlands Youth Institute – Figures on higher https://www.nji.nl/cijfers/hoger-onderwijs-ho

[29] National Student Housing Monitor (2014) https://www.kences.nl/publicaties/landelijke-monitor-studentenhuisvesting-2014/

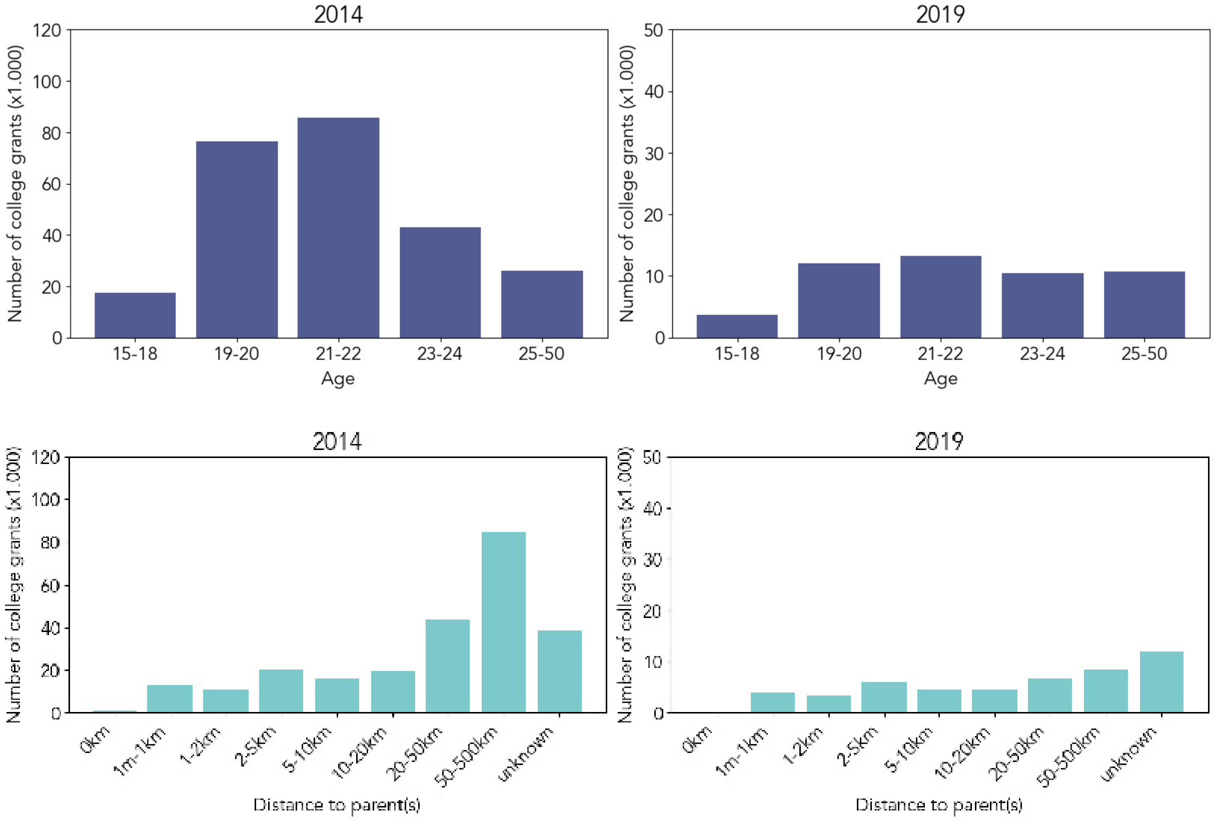


Figure 5 – Distribution of type of education, age group and distance to parent(s) in college grant population-2014 (n=248.649) and college grant population-2019 (n=50.233).

There are respectively 3.179 and 934 students in the college grant population-2014 and -2019 who have been selected for a control procedure. The result of an control can have three outcomes: the college grant is duly allocated (duly), the college grant was unduly allocated (unduly) and the outcome of the control is unknown (unknown). The outcome of a control procedure may be unknown, for example because a home visit is refused by the student, or a student is not at home. The results of checks carried out in 2014 and 2019 are shown in Table 13-14.

CUB checks 2014	Number	Percentage
Duly	1.566	49,2%
Unduly	1.238	38,9%
Unknown	375	11,9%
Total	3.179	100%

Table 13 – Results of CUB checks in 2014

CUB checks 2019	Number	Percentage
Duly	406	43,5%
Unduly	330	35,3%
Unknown	198	21,2%
Total	934	100%

Table 14 – Results of CUB checks in 2019.

The results of the checks from 2014 and 2019 in terms of type of education, age groups and distance to parent(s) categories of the students are shown in Figure 6.

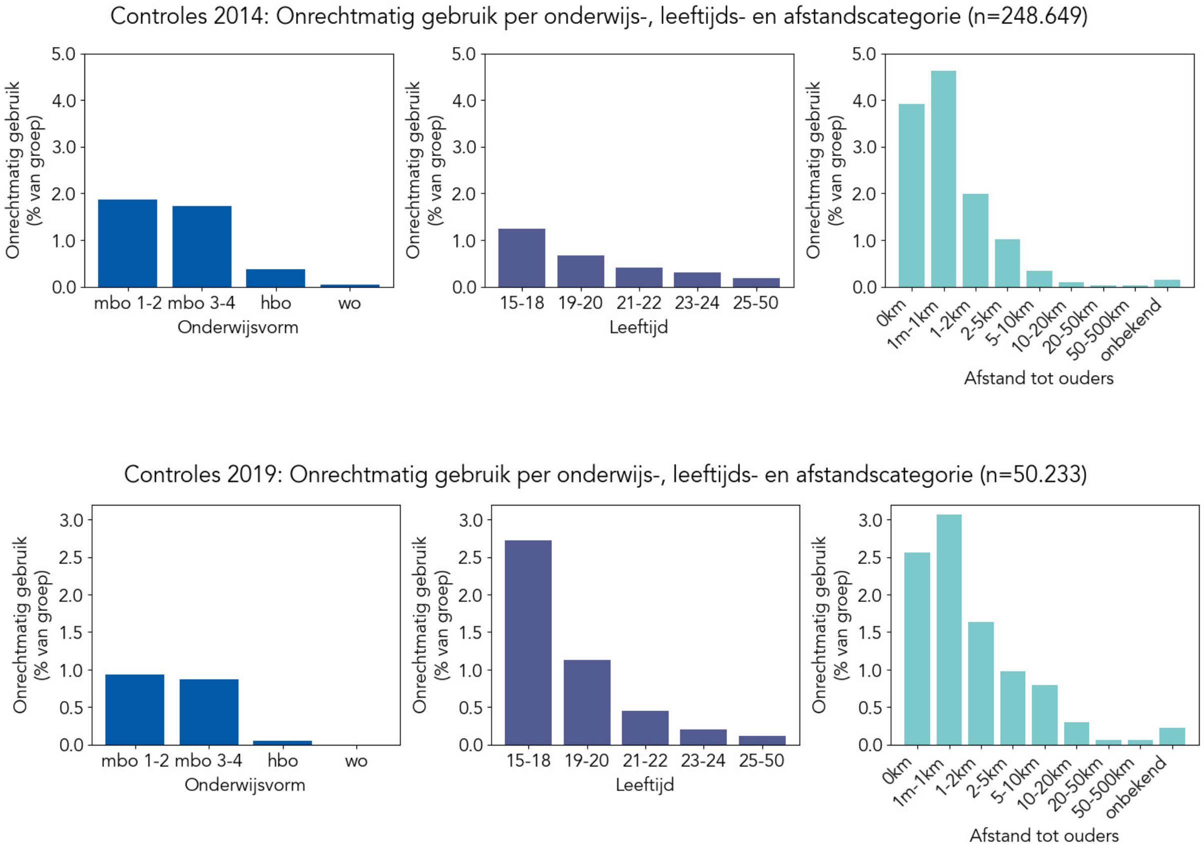


Figure 6 – Unduly use of the college grant, broken down by education, age groups and distance categories based on college grant population-2014 (n=248.649) and college grant population-2019 (n=50.233).

The results that can be derived from Figure 6 are in line with the logic of the risk profile: students are more likely to make unduly use of the college grant if they have a vocational type of education (mbo), are young and are registered close to their parent(s). However, these differences are observed at the end of the CUB process; the assigned risk scores in step 1 of the CUB process and the manual selection of students in step 3 of the CUB process have influenced this. This raises the question what the influence os of the algorithm (step 1) and manual selection (step 3) on the outcome of the CUB process. This question plays a central role in the following sections of this study.

Results of research question 4

Research question 4 is formulated based on two key concepts from scientific literature surrounding fair algorithms (conditional demographic parity and demographic parity). These concepts are explained in 3.1 Quantitative analysis.

Research question 4:

Has demographic parity been satisfied in the risk score population (after step 1 – risk profile) and the home visit population (after step 3 – manual selection by employee)?

Answer research question 4:

For both step 1 (risk profile) and step 3 (manual selection by employee), demographic parity is violated in 2014 and 2019. In practice, pure demographic parity is seldom realized. Demographic parity mainly serves as a reference point against which to compare algorithmic of human behavior. Based on the below analysis, it can been observed that students who are registered close to their parent(s), especially in the 0km, 1m-1km and 1-2km distance categories, are exceptionally often manually selected for a home visit in step 3 of the CUB process. It is obvious that specific work instructions, which encourage the manual selection of students who are registered near their parental address, are the cause of this.

Motivation research question 4:

Figure 7 shows the average risk score assigned by the risk profile per education, age and distance category. The gray bars show the expected score as can be determined using the risk profile given in Table 1-2. This is the score that the risk profile gives to a given group, assuming that all other characteristics are chosen at random.

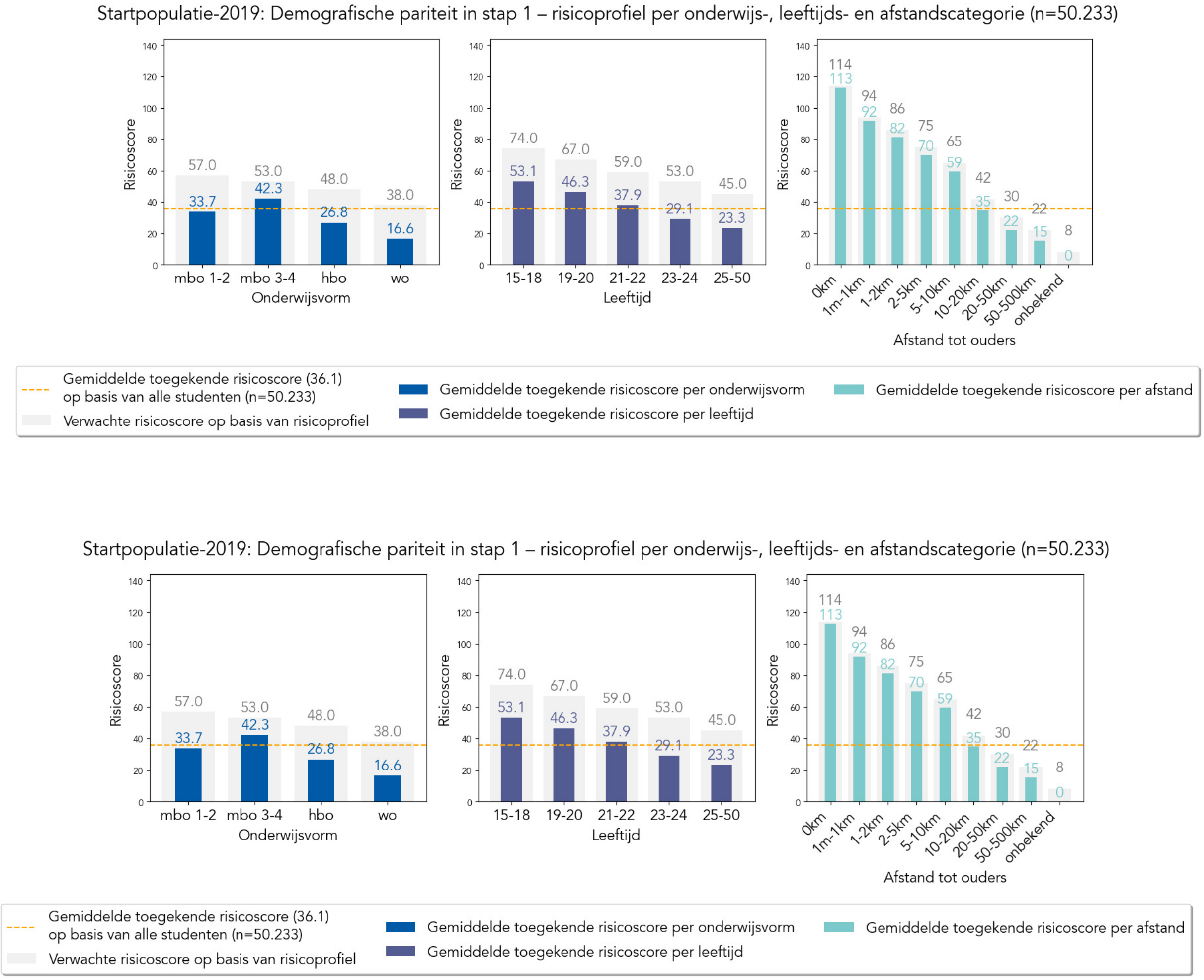


Figure 7 – Demographic parity step 1 – risk profile by education, age, and distance category for college grant population-2014 (n=248.649) and college grant population-2019 (n=50.233).

Observations regarding demographic parity after step 1 (risk profile) are discussed per selection criterion.

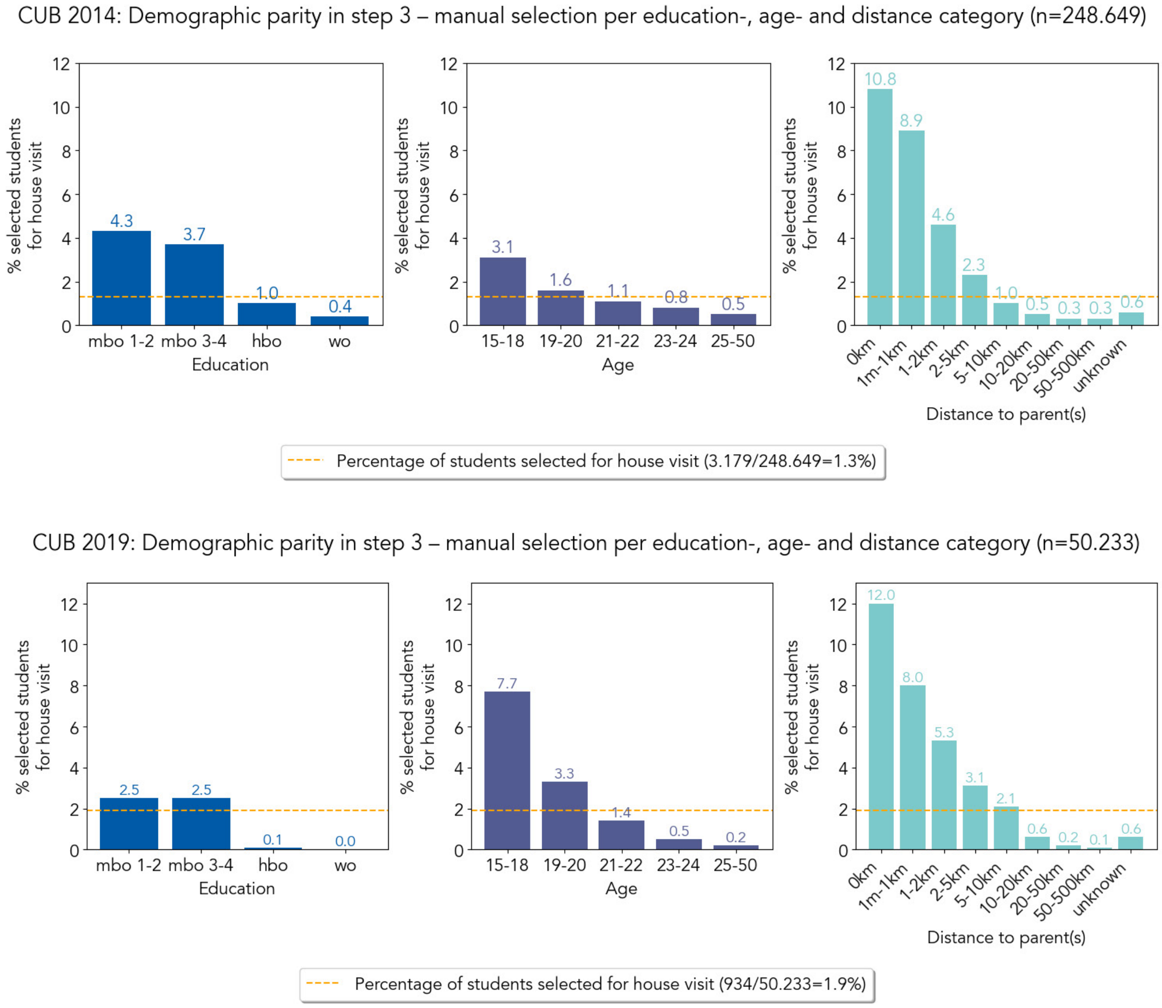
> Type of education – The risk model makes a strong distinction based on the type of education. In 2014, the risk score for mbo 1-2 and mbo 3-4 students was 1.6x higher than the average assigned risk score (45.8/28.2 and 46.5/28.2 respectively). It is striking that this pattern is not visible in 2019, where less distinction is made based on the type of education – for mbo 1-2 the score is 0.9x the average (33.7/36.1) and for mbo 3-4 only 1.2x (42.3/36.1).
> Age – The risk model makes a distinction based on age, but in 2014 this distinction appears to be less significant than with the characteristics of type of education and distance to parent(s). For example, in 2014 the risk score for someone aged 15-18 is 1.1x the average risk score (32.4/28.2). However, this pattern appears different for 2019, where a higher factor is observed based on age. In 2019, the risk score for someone aged 15-18 is 1.5x the average risk score (53.1/36.1).
> Distance to parent(s) – The risk model makes a very strong distinction regarding the distance to parent(s). In 2014, the risk score assigned to someone who lives 0km from his/her parent(s) is 3.7x higher than the average risk score assigned for all distance categories (104/28.2). For the 1m-1km category this is 3.0x higher (86/28.2) and for the 1-2km category this is 2.7x higher (75/28.2). We observe a similar pattern for 2019.

In addition, the differences in assigned and expected risk scores per selection criterion are discussed:

> Type of education – The average risk scores per type of education (blue bars) are well below the expected risk scores (gray bars). This can be explained by the group of students with unknown distance to parent(s) who are assigned a risk score of 0, which significantly reduces the average scores. In 2014, there were 38.657 recipients of the college grant with an unknown address. In 2019 this was 12.059.
> Age – The average risk scores per age category (purple bars) fall below the expected risk scores (gray bars). This can also be explained by the group of students with unknown distance to parent(s) who are assigned a risk score of 0, which significantly reduces the average scores.
> Distance to parent(s) – The average risk scores per distance category (green bars) for 2014 and 2019 are largely in line with the expected risk scores (gray bars).

Based on the above observations, it is concluded that the risk profile has assigned risk scores to students in accordance with expectations, and often even lower than expectations.

Figure 8 shows the chance of being selected for a home visit per education, age, and distance category in 2014 and 2019.


CUB 2014: Demographic parity in step 3 – manual selection per education-, age- and distance category (n=248.649)

CUB 2019: Demographic parity in step 3 – manual selection per education-, age- and distance category (n=50.233)

Figure 8 – Demographic parity step 3 – manual selection by DUO employee in CUB process 2014 (n=248.649) and CUB process 2019 (n=50.233).

Observations regarding demographic parity after step 3 (manual selection of employee) are discussed per selection criterion.

> Type of education – For 2014 it is striking that an excessive number of mbo 1-2 and mbo 3-4 students are selected. For mbo 1-2 students this is 3.3 times as often as the average (4.3%/1.3%). For mbo 3-4 students this is 2.8x as often as the average (3.7% / 1.3%). This is remarkable since the risk profile for these groups assigns a risk score that is 0.9x and 1.3x higher than the average, respectively.

> Age – For 2019, it is also striking that an excessive number of 15-18 year-olds are selected: 4.1x more than the average (7.7%/1.9%). The risk profile assigns a 1.5x higher risk score than the average for this group.

> Distance to parent(s) – For both 2014 and 2019, the chance of being selected for a home visit if students are registered close to their parent(s) (0km, 1m-1km, 1-2km) is exceptionally high. For those who live 0km away in 2014, the chance of a home visit (10.8%) is 8.3x greater than the average for all categories (1.3%). For the 1m-1km category this is 6.8x higher (8.9%/1.3%) and for the 1-2km category this is 3.5x higher (4.6%/1.3%). A similar pattern is observed in the 2019 CUB process data. This is in stark contrast to the scores assigned by the risk profile – for example, someone living 0km away in 2014 only has a 3.7x higher score than the average, which is relatively less of a

difference than the chance of a home visit (8.3x higher than the average). This difference indicates a strong overreaction of employees based on distance near the parental address during the manual selection of students for a home visit.

Results of research question 5

Note that the two core concepts from scientific literature surrounding fair algorithms (conditional demographic parity and demographic parity), on the basis of which research question 4 was formulated, are introduced in 3.1 Quantitative analysis.

Research question 5:

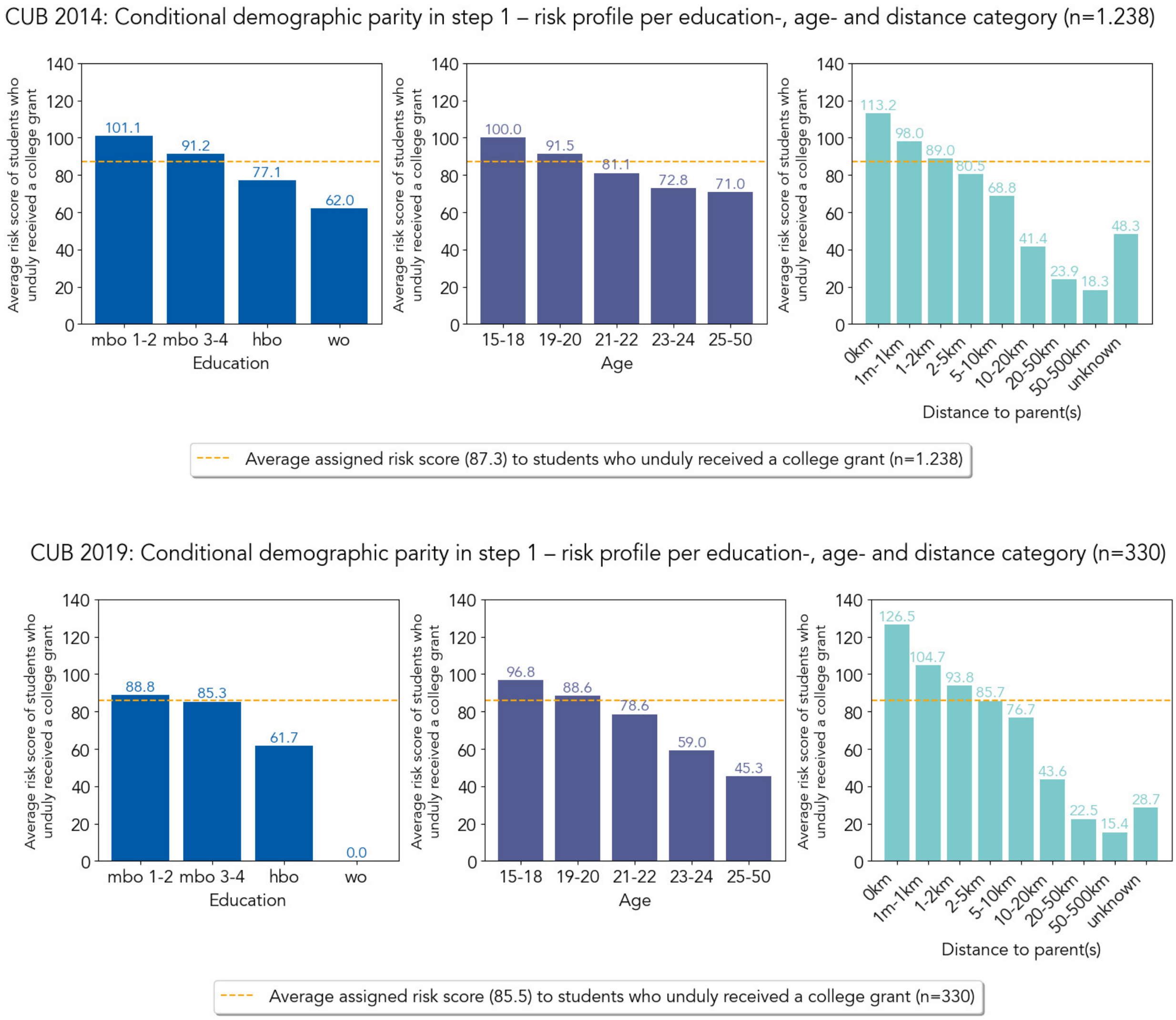
Has conditional demographic parity been satisfied in the risk score population (after step 1 – risk profile)?

Answer research question 5:

In both 2014 and 2019, conditional demographic parity for step 1 (risk profile) of the CUB process was not satisfied. Pure conditional demographic parity (as defined in 3.1 Quantitative analysis) rarely occurs. Measuring conditional demographic parity mainly serves as a reference point against which to compare the behavior of a risk profile or human actions. Based on this analysis, it is observed that some groups that are registered far from their parent(s) receive a lower risk score than would be expected based on the results of the control procedures. This observation confirms the impression that the risk profile does not function optimally.

Motivation research question 5:

Differences noted in the previous section in terms of demographic parity may be justified because students with certain characteristics more often make unduly use of the college grant. Conditional demographic parity compensates for this by showing relative differences in the group of students who have unduly used the college grant. Figure 9 shows that based on the results of the CUB process in 2014, in a certain sense, there is conditional demographic parity for type of education and age groups. Only university (wo) students in 2014 were awarded a lower score than would be expected for students who unduly used the college grant. In addition, the discrepancy is striking for higher distance categories. This means that there are relatively many students with a low-risk score (partly explained by the fact that they are registered far from their parent(s)) who have nevertheless made unduly use of the college grant. Figure 9 shows the same pattern for the outcomes of the CUB process in 2019. Except that age categories including older students are below the average compared to 2014. These results confirm the impression that the risk model and associated risk classification do not perform optimally.



CUB 2014: Conditional demographic parity in step 1 – risk profile per education-, age- and distance category (n=1.238)

Average assigned risk score (87.3) to students who unduly received a college grant (n=1.238)

CUB 2019: Conditional demographic parity in step 1 – risk profile per education-, age- and distance category (n=330)

Average assigned risk score (85.5) to students who unduly received a college grant (n=330)

Figure 9 – Average risk scores by risk profile (step 1) assigned to students who unduly used the college grant, broken down by type of education, age groups and distance categories in 2014 (n=1.238) and 2019 (n=330).

Results of research question 6

Research question 6:

How much unduly use is observed in the population selected by the CUB process in 2014 and 2019?

Answer research question 6:

Table 12-13 indicate that the effectiveness of the CUB process in 2014 and 2019 was found to be 38.9% and 35.3% respectively (n=3.179 and n=934).

5. Results of qualitative analysis

The results of the qualitative research are discussed below. The development of the risk profile is discussed in §5.1 Processes to mitigate bias are discussed in §5.2. Research questions 7 and 8 are answered based on these results.

5.1 Qualitative analysis of risk profile

Research question 7:
Did the risk profile used in the CUB process meet the standards that are currently set for algorithms used by Dutch public sector organisations?

The research question is answered based on two parts. Part 1: Mitigating bias when developing the risk profile. And part 2: Processes to counter the risk of bias during deployment.

Part 1: Mitigating bias when developing the risk profile

The risk profile consists of two parts: the selection criteria and the weighting factors.

The selection criteria were drawn up based on a manual analysis of lists of students living on their own in the pilot regions in combination with experience data and what was called 'common sense' in the workshops [A.46, A.47].

The Netherlands Institute for Human Rights warns against the use of experiential data:

> *Although experiential data can contribute to the efficient exercise of supervisory and enforcement powers, they can also further encourage stigmatization and discrimination against certain population groups. Experiential data is not always as neutral or objective as it seems: a correlation between factor X and behavior Y is not the same as a causal relationship between them and this is often confused. As a result, prejudices and assumptions are often reflected in experiential data. Moreover, when a certain population group is monitored more often, people from this group will also be exposed more often as perpetrators or offenders, which in turn leads to more intensive monitoring of this population group and to confirmation of the need for this. In this way, an overemphasis on experiential data encourages a self-fulfilling prophecy. Partly for this reason, experiential data and crime statistics must be handled with extreme caution: in many cases their compilation is anything but objective.*[30]

Risk profiling can also be a method of selection based on more objective criteria than experiential data. A risk profile based on objective data is to a certain extent more neutral than a manual selection. However, this objectivity can also be apparently neutral. Risk

[30] Netherlands Institute for Human Rights, *Discrimination through risk profiling. A human right assessment framework*, 2021, p.14.

profiles that are not based on experiential data are also not necessarily fair. The present risk profile is an example of this. Although the selection criteria in themselves are objective, the choice of these selection criteria shows how the designers of the risk profile think about unduly use of the college grants. The selection criteria were not 'found' by an analysis of many neutral selection criteria; they have been selected by the designers of the risk profile on the basis of common sense and subsequently tested. It is for these reasons that normative considerations underlying the composition of a risk profile must be carefully documented. See also *Algorithms and Fundamental Rights* (Utrecht University) p.49 with further references. Safeguards for the preparation and use of risk profiling algorithms are provided in the ADR Research Framework and in the FRAIA.

The weighting factors are said to have been determined by a 'data analysis' of the relationships between different groups of students. This data analysis can no longer be fully reproduced. The development of the weighting factors is discussed in more detail in 4. Results of quantitative analysis.

The ADR Research Framework prescribes that "*relevant stakeholders must be involved when developing the algorithm concerned*" (SV.10). This has partly been done. The enforcement authorities and the ministry are involved in the development. The students who would be subjected to risk profiling are not involved. Students could possibly have pointed out assumptions or biases in the risk profile.

Furthermore, the ADR Research Framework prescribes that "*the objectivity of the algorithm has been specifically elaborated into functional requirements for the algorithm. The extent to which these requirements have been met has been determined*" (DM.1). It was measured at various times before and during the use of the CUB risk profile whether the goal of the risk profile (effectiveness) was achieved [A.11-A.15,A.20-21]. This was not done based on pre-established standards – it was not determined in advance for what effectiveness metrics the risk profile was considered to be successful.

The ADR Research Framework prescribes that "*the choice of the model and the hyperparameters have been motivated and documented*" (DM.2) and that the limits of the applicability of the model have been documented (DM.3). This is partially met [A.40]. The choice of the risk profile and its parameters can be partially reconstructed but are not consistently documented [C.42-1]. There is no motivation for the used weighting factors (also known as risk codification). Why the used selection criteria were chosen is not explained in detail [E.9-4].

This is also at odds with guideline 9 of the Assessment Framework for Discrimination through Risk Profiles of the Netherlands Human Rights Institute: "*Selection decisions must be explainable at all times*".[31]

[31] p.30

It also did not appear that "*the target population has been determined and that it has been checked that the test data is representative of the data of the various subgroups that appear in the production data*" (DM.7). It has not been shown that "*the input and output data are satisfactory in terms of quality, completeness and reliability*" (DM.9). The selection of 934 students in 2010 considered the subgroups of education, age and distance to parent(s) [A.47]. The representativeness of other subgroups – particularly relevant to this study, the subgroup migration background/non-Dutch ethnicity – has not been examined. The research did not show that "*during the development of the model, it was assessed whether there is a difference between the performance of the model between different subgroups*", which is recommended by the ADR Research Framework (DM.20).

The additional requirement is stated in guideline SV.4 of the ADR Research Framework: "*the impact of the algorithm has been inventoried and assessed*". This has been done in terms of the effectiveness and cost of the algorithm. No attention was paid to the emergence of possible (indirect) bias against certain groups. This should be a focus point in the future; not only with respect to the subgroups that were included in the quantitative research and that reached media attention, but also other subgroups.

In more general terms, the FRAIA confirms the importance of making public values explicit when drawing up an algorithm. If those values – including possible bias – are made explicit, they can then be checked:

> *It is not always easy to know what public values are, but they are always manifestations of the general interest. (…) Examples of public values are equality, (…), legal certainty, distributive justice, respect for vulnerable groups, participation and efficient use of resources. At a more concrete level, the protection of fundamental rights (...) can also be regarded as public values.*
>
> *Algorithms can serve to translate certain public values into concrete decision-making procedures. Algorithms can strengthen certain values, but they can also undermine public values – such as fundamental rights. Precisely for this reason it is important to identify which public values may be relevant when using the algorithm.*[32]

All of this – paying attention to subgroups, involving all stakeholders in the development of a risk profile, and defining performance indicators in advance – will have to be done better if risk profiles are used in the future.

Part 2: Processes to counter the risk of bias during deployment

The ADR Research Framework provides guidelines for mitigating bias: "*The definition of the different groups and the desired performance of the model for these groups must be included in the functional requirements*" (DM.16). The subgroups of type of education, age groups and distance to parent(s) have been partly determined. Other subgroups –

[32] Dutch FRAIA p. 14.

including migration background/non-Dutch ethnicity – are not defined. The desired performance has not been determined for any of the groups. Also, "*the degree of accepted bias in the outcome has not been determined and not elaborated into measurable performance criteria*" (DM.17). By not recording all this, it is much more difficult to monitor or check whether the risk profile has resulted in an overrepresentation of certain subgroups in the CUB process while using the risk profile. Processes can be set up to monitor this, such that possible unintended bias can be mitigated.

Such processes are also prescribed by the ADR Research Framework. Guidelines DM.18, DM.19 and DM.21 prescribe: "*The methods to prevent, detect and correct bias are laid down; the outcome bias of production data has been assessed for the different subgroups and meets the performance criteria; the degree of bias in the data, data collection and the model have been mapped out.*" How this can be monitored is discussed in 4. Results of quantitative analysis.

Because bias has not been measured, guideline DM.22 cannot be met: "*the observed bias has been assessed as to whether it indicates discrimination*". On page 29 from the FRAIA the importance of mitigating bias is confirmed. The Netherlands Institute for Human Rights confirms that apparently neutral selection criteria can have discriminatory effects.[33]

To ensure that the above guidelines are met, processes can be set up to minimize the risk of bias in the risk profile. The ADR also prescribes this: "*risk management takes place in a structured manner before and during the use of the algorithm*" (SV.13). More specifically, "*a process for periodically evaluating the quality of the algorithm must be documented and in place. The results should be shared with stakeholders*". (SV.16) Periodic evaluation of the CUB process was common practice but did not relate to assessing the quality of the risk profile and the risk of bias [A.24]. As part of periodic evaluation, the "*responsible person must account for the development, deployment and operation of the algorithm*" (SV.19).

To make evaluation possible, "*the functional requirements must be developed into adequate and measurable performance criteria*" (DM.4). This is in line with the already discussed recommendation to draw up internal standards. For this purpose, performance metrics relevant to the risk profile as well as evaluation criteria relevant to the CUB process must be drawn up.

Part of the evaluation should also include "the to analyze whether (internal and external) complaints and incidents can result from the use of the algorithm" (SV.17). If there were already methods and measures to mitigate bias in the CUB process – which this audit did not find[34] – then these are not documented. It is worth noting that this audit did not reveal any systematic complaints about the risk profile.

[33] p.18 et seq. *Assessment Framework for Discrimination through Risk Profiling* of the Netherlands Institute for Human Rights.

[34] And what the ADR also concludes, see B.8 and B.42-4.

As discussed, the CUB process of which the risk profile was a sub procedure was indeed monitored and evaluated. M&O studies have consistently established that the risk of unduly use of the college grant is high. It is not known what influence this has on the risk profile and the CUB process. It is possible that a process can also be set up for these periodic M&O investigations to mitigate bias.

The CIO of the Dutch Government may play a role in drawing up a general risk management framework. The Implementation Framework for Responsible Use of Algorithms, which has not yet been published in final form, states:

> *Together with the ministries, CIO Rijk also sets up an internal supervisory structure in which the roles of the various lines of defense are defined. The different lines of defense provide multiple control points in the supervision of algorithms. The supervisory structure will also be described in the next phase of the implementation framework. The way in which this can be structured within municipalities, provinces and water authorities will also be discussed.*

In sum, it is recommended to apply a risk management system for designing and deploying risk profiles – or algorithms in general by Dutch public sector organisations. This reduces the risk of unintentional bias and makes decisions explainable.[35]

Results of research question 7:
The risk profile did not meet several of the standards that are currently set for algorithms used by Dutch public sector organisations. The answer to research question 7 is therefore negative.

5.2 Processes to address the risk of bias

Research question 8:
If the answer to research question 7 is negative, what is needed to use the risk profile responsibly in the future?

Answer research question 8:
Research question 8 was partially answered in the discussion of research question 7. The recommendations will identify concrete measures for the (possible) responsible use of the risk profile in the future.

[35] p.30

6. Disclaimer

This report does not aim to provide an exhaustive, definitive overview of the CUB process. Various documents have been requested from DUO and various documents have been shared by DUO on its own initiative. There may be information available that the researchers have not seen and that (partly) contradicts the analysis.

Algorithm Audit further requested data from data warehousing experts at DUO. Checks have been carried out to verify the correctness of the data. Algorithm Audit cannot guarantee that the data on which the report is based, the associated queries and/or the underlying data structures are complete free of errors or imperfections.

Algorithm Audit is not responsible for any decisions made as a result of this report.

This report – and/or parts of the report – may not be shared with parties outside DUO without prior permission from Algorithm Audit.

The protection of personal data was taken into account when preparing this report.

- Algorithm Audit's internal processes comply with the GDPR;
- DUO shared its documents via its own platform. Algorithm Audit has therefore never had documents originating from DUO in its hands. When access to DUO systems is denied, Algorithm Audit permanently no longer has access to the documents.

7. Conclusion: findings and recommendations

Based on this analysis, points for improvement can be identified to prevent bias in the CUB process in the future. This report makes a start in this regard in seven findings (§7.1) and three recommendations (§7.2).

7.1 Findings

This section summarizes the findings. The findings arise from the quantitative and qualitative analysis of the CUB process and how design and deployment of the process is structured within the organization.

Finding 1 – A rule-based algorithm has been used for years, which assigned a risk score to students who received a college grant based on type of education, age and the distance between the student's address and the address of his/her parent(s). No self-learning algorithm or artificial intelligence systems has been used. The use of risk profiling has proven to be effective.

Between 2012 and 2023, a risk profile was used in the CUB process to automatically assign risk scores to all students who received a college grant from DUO. The risk profile is a linear model that assigns students a risk score based on three criteria. These criteria are type of education, age and the distance between the address where they were registered and the address of their parent(s). Each criterion is divided into categories, for example students who live between 1m-1km from their parent(s), students who live 50-500km from their parent(s), are following vocational training (mbo 1-2) etc. After applying predetermined weighting factors per category, a student's profile leads to a risk score. The risk score can be increased if the age of the student known to DUO differs from the age of the student in the General Registration of Persons (BRP). All students have been assigned a risk score. Applying the CUB process to trace unduly usage of the college grant, of which the risk profile is an important part, has proven to be effective. The effectiveness is evident from the fact that more unduly use of the college grant has been identified with application of the CUB process than with random samples. This concerns 3.6% and 3.8% effectiveness respectively in the random samples of 2014 and 2017 and 38.9% and 35.3% effectiveness respectively when applying the CUB process in 2014 and 2019 in which the risk profile was used.

More information about the risk profile and the CUB process is provided in 2.3 Overview of CUB process. More information about the effectiveness figures can be found in 3.1 Quantitative analysis.

Finding 2 – The random samples from 2014 and 2017 show insufficient statistical relationship between the selection criteria type of education and age, and unduly use of the college grant. A statistical relationship exists between specific categories within the selection criterion 'distance to parent(s)' and unduly use of the college allowances. Insufficient statistical support has been found for the division into six risk categories compared to a binary risk classification.

Per education, age and distance category, it has been counted how often students that were selected for a control procedure in the random sample in 2014 (n=387) and 2017 (n=293) unlawfully used the college grant. Based on these frequencies, it was tested, using a one-sided Z-test and Fisher's exact test, whether the differences in unlawfulness percentages for the categories used in the risk profile are statistically significant. For example: is there a statistically significant difference between percentages of unduly allocation of college grants between the distance category 2-5km and the category 5-10km? For the profiling criterion age, insufficient evidence was found for statistically significant differences between the used categories. For the criterion type of education, one significant difference was found in the 2014 sample but no significant differences in the 2017 sample. Algorithm Audit considers this to be an insufficiently consistent signal on which risk profiling should be based. A new frequency count on a larger random sample could change this. For distance to parent(s), evidence is found for statistically significant differences between unduly usage percentages in the 1m-1km, 2-5km and 50-500km categories. This provides support for the use of distance as a selection criterion for risk profiling, although the binning thresholds for categorization should be determined more precisely. In addition, there is no quantitative support to divide the assigned risk scores into six risk categories. There is statistical support for a binary risk classification, and this simplification is preferred.

More information about the methodology of the statistical tests can be found in 3.1 Quantitative analysis. The results are discussed in 4.1 Results of random sample 2014 and 2017.

Finding 3 – Students who are registered within a distance of 2 km from their parent(s) are manually selected for a home visit significantly more often than one would expect based on the risk scores assigned by the risk profile. It is likely that specific work instructions, which encourage the manual selection of students who are registered near their parental address, are the cause of this.

In the bias measurement, a strong overrepresentation was observed of students who are registered within 2 km of their parent(s) during manual selection for home visits. This overrepresentation does not stem from the risk profile used.

In 2014, for students who are registered 0km from their parent(s), there is a disproportionate ratio between the probability of being selected for a home visit (8.3x the average of all categories) and the assigned risk score (3.7x the average). This means that students from this category are more than twice as likely to be selected for a home visit than would be expected based on their assigned risk score. For the 1m-1km category, this ratio between the probability of a home visit and risk score (both compared to their averages) is 6.8x : 3.0x and for the 1-2km category 3.5x : 2.7x. This means that the group with a small distance to parent(s) is structurally overrepresented in the manual selection for home visits, compared to the risk scores assigned to them.

This overrepresentation can probably be explained by the presence of specific work instructions that instruct employees to specifically select students with specific living conditions – such as a combination of young age, living close to their parent(s) and/or living with family members – for a home visit. The overrepresentation of students (such as those registered at a small distance from their parent(s)) in the manual selection for home visits could potentially cause an over-representation of other characteristics, such as their migration background. Further research is therefore needed to determine whether there is a connection between the groups that are excessively manually selected for home visits and students with a migration background. See Recommendation 2. In addition, an investigation and improvement trajectory must be set up for the manual selection process, including a review of the work instructions. See Recommendation 3.

The results of the bias measurement, including the figures mentioned, can be found in Figure 8 and in 4.2 Results of bias measurement 2014 and 2019.

Finding 4 – No direct differentiation has occurred on the basis of migration background (nor based on other protected grounds) in the risk profile. Due to insufficient available data, it has not yet been possible to conduct quantitative research into indirect bias towards students with a migration background.

DUO has requested Algorithm Audit to also carry out a bias measurement for possible bias regarding the protected attribute migration background. DUO itself has no data on the migration background of students. The Netherlands' national office of statistics (CBS) has therefore been asked to enrich the DUO data at group level with data on the migration background of students solely for the purpose of carrying out this bias measurement. To date, Algorithm Audit has been unable to obtain this data. In the future, this data may be made available for possible follow-up research. Algorithm Audit has considered alternative methods to measure indirect bias regarding migration background, such as using aggregation statistics per postal code area. However, according to statistical experts affiliated to Algorithm Audit this approach is insufficiently justified from a methodological point of view. At the time of publication of this report, Algorithm Audit was therefore unable to conduct quantitative research into indirect bias of the CUB process regarding students with a migration background. Possible indirect bias in the risk profile or in the further course of the CUB process cannot be ruled out.

How the bias measurement for migration background would have been carried out in the case of sufficient available data is explained in 3.1 Quantitative analysis.

Finding 5 – According to current standards, a well-motivated rationale for used selection criteria and risk categories in the risk profile are lacking, especially regarding possible bias.

The rationale for usage of criteria in the risk profile are largely based on personal experiences and so-called 'common sense' of employees. The suitability of these criteria for the aim pursued is not documented. The origins of the risk profile can be traced back to a series of workshops in 2010. Although subject matter expertise and common sense are important and useful, *self fulfilling prophecies*, *confirmation biases* and the unintentional use of proxy criteria (see Box 1) are a real risk here. The weighting factors used in the risk profile are based on a data study conducted in 2010. Documentation about the origin and methodology used to compose the control group in this data study is lacking. Without further research into the representativeness of this control group, certain groups may be over- or underrepresented in the population, specifically in case the group is manually sampled. Basing weighting factors on that population risks creating *negative feedback loops*. This means that groups that are overrepresented in the control group receive a higher weighting factor, as a result these groups are later systematically assigned an excessive risk score by the risk profile. Substantiating the choice of certain selection criteria and weighting factors and their possible influence on bias is prescribed by the currently applicable frameworks for the responsible use of algorithms such as the Fundamental Rights Algorithm Impact Assessment (FRAIA) and the Algorithms Research Framework of the ADR.

An analysis of the substantiation of the selection criteria and categories used in the risk profile is provided in 5.1 Qualitative analysis of risk profile.

Finding 6 – There has been no internal research into possible biases in the risk profile, neither during development and deployment of the risk profile.

Risk profiling is a means permitted by the Netherlands General Court of Appeals for increasing effectiveness, efficiency, and cost savings in combating unduly use of public allowances.[36] However, conditions apply to usage of risk profiling, including that made differentiation must be suitable, necessary, and proportionate. This must be guaranteed during the design and deployment of the risk profile. There has been no evidence of awareness within DUO (and/or the bodies involved such as the Ministry of Education, Culture and Science, the ADR, the House of Representatives after the introduction of CUB in 2012, etc.) that the use of the risk profile in the CUB process entails a risk of (indirect) unequal treatment. There has also been no evidence that checks and balances have been put in place to monitor or prevent such unequal treatment. No bias was measured, and no investigation or consideration was given to whether the criteria and weighting factors used may have been proxy characteristics for certain demographics.

An explanation of the lack of internal investigation into bias can be found in 5.2 Processes to address the risk of bias.

Finding 7 – There are no internal standards for responsible use of algorithms.

This study found no internal guidelines used by DUO to prevent bias or mitigate other risks when using algorithmic methods. It does not appear to have been tracked how DUO implements general legal frameworks and guidelines from other authorities for the composition of risk profiles. If these guidelines do exist, not the entire organization is aware of them. Guidelines are lacking in at least two areas that are relevant to prevent (indirect) discrimination. Firstly, guidelines for which degree of correlation between risk profiling criteria and protected grounds indirect discrimination occurs. Secondly, to determine under what circumstances an objective justification exists for usage of profiling criteria, because the legal requirements of suitability, necessity and proportionality are met. In the period 2012-2022, annual monitoring

[36] Including CRvB 8 September 2015, ECLI:NL:CRVB:2015:3249, paragraph 4.5.; see also Assessment Framework for Discrimination through Risk Profiles of the Netherlands Institute for Human Rights (2021) Guideline 1.

reports, privacy audits and abuse and misuse checks (misbruik-en-oneigenlijk gebruik – M&O) took place. Random sampling took place in 2010, 2014 and 2017, primarily with the aim of determining the effectiveness of the CUB process and to estimate the financial 'residual risk' in the entire CUB population. The lack of guidelines for dealing with (algorithmic) risk profiling has contributed to the fact that the checks did not pay attention to possible bias.

An explanation of the lack of internal standards for the use of algorithms is given in 5.2 Processes to address the risk of bias.

7.2 Recommendations

The following recommendations follow from the findings mentioned.

Recommendation 1 – Provide a well-motivated rationale for usage of risk profiling and specific criteria before profiling is potentially used again, among others with help of a normative framework.

Inherent in the use of risk profiling is that differentiation is made between groups of students. This is partly the intention: not a *bug*, but a *feature*. However, treating groups of people differently can also exceed legal and social norms. This is, for example, the case when differentiation is not suitable, necessary, and proportionate. At the same time, differentiation that is lawful can still be considered socially and ethically undesirable. This could include differentiation that is not focused on a legally protected ground, but on different characteristics such as education level. It is recommended to devise an internal framework against which can be assessed which forms and to what extent differentiation of groups of people are considered (un)desirable. Such a framework specifically relates to testing criteria that precede usage of the risk profile, such as statistical hypothesis testing. The proposed Phase 2 study of Algorithm Audit provides starting points for making normative considerations to interpret quantitative testing results. Assessing a risk profile against a framework must also contain quantitative support of the selection criteria used, for which the analyzes in this report provides a preliminary step. To motivate the selection profile quantitatively, and not just measure the effectiveness of the CUB process, it is recommended to draw a larger random sample than the random samples from 2014 and 2017.

Recommendation 2 – Further research should be conducted to determine whether there is a link between the groups that are excessively manually selected for home visits and students with a migration background.

Whether groups that are selected excessively during manual selection also have a migration background is not clear and must be further investigated. The work instructions used, for example the guideline to further investigate details of the living situation (such as the combination of young people and living with family), can contribute to possible over-representation of demographic groups, including students with a migration background. In addition, the process of manual selection is susceptible to discrimination based on latent characteristics. At the lists that employees are presented with during manual selection, in addition to the criteria from the risk profile and the risk score, other student characteristics were also visible, such as name, address and date of birth. There is a risk of (unconscious) bias regarding migration background or other characteristics in manual selection. The influence that these and other characteristics play during the manual selection for home visits needs to be investigated. In addition, an improvement trajectory must be set up to improve work instructions for manual selection. See Recommendation 3.

Recommendation 3 – Avoid using the same selection criterion in different steps of the CUB process. Set up an investigation and improvement trajectory for the manual selection process, including redesign of the work instructions.

It has been established that students who are registered within a distance of 2 km from their parent(s) have been manually selected for a home visit significantly more often than would be expected based on the risk scores assigned by the risk profile. It is obvious that specific work instructions that encourage the manual selection of students who are registered near their parental address are the cause of this. These findings provide reason to examine the work instructions for employees in the selection process. If the risk score assigned by the risk profile already considers characteristics such as living nearby parent(s), further instruction to manually select the same characteristics more often may cause an overreaction. As a result, students with a short distance to their parent(s), for example, are wrongly selected for home visits much more often. The overall manual selection process needs further investigation to identify areas for improvement. The current process is opaque, because the procedure to manually include and exclude students from home visits based on their risk scores and relevant characteristics is not a clear protocol. It is difficult to find out exactly how employees work, and the process is therefore difficult to control. In addition, latent personal characteristics that are visible to employees (such as name, address, date of birth) can pose a risk of (unconscious) bias. Anonymizing students and exclusively showing only the relevant characteristics and the risk score would be a possibility. Further research should reveal what improvements can be made in the manual selection process.

Recommendation 4 – Establish an organization-wide algorithm management policy to reduce the risks associated with the use of algorithms.

Draw up a policy for consistent and responsible handling of algorithms and risk profiling. Such policies can be drawn up in different dimensions. At this point one could think of:

> Governance: centralization of key decision-making in committees, alignment of roles and responsibilities within the existing organizational structure, implementation of 3 lines-of-defense (3LoD) risk management framework.
> Documentation: drawing up standardized documentation requirements and work instructions.
> Algorithm inventory: central overview of all algorithms[37] within the organization for informa-

[37] For a definition of an algorithm see: https://algoritmes.overheid.nl/nl/footer/over-algoritmes.

tion requests, monitoring and evaluation.

- Processes: Establish evaluation and validation processes for each algorithm.
- Monitoring and reporting: Monitoring of reporting on algorithmic risks promotes risk-oriented practices.

Inspiration for a different division can also be drawn from existing frameworks for responsible handling of algorithms in the public sector.[38]

[38] See, for example, the Implementation Framework 'Responsible Use of Algorithms' of the National Government https://www.rijksoverheid.nl/documenten/rapporten/2023/06/30/implementatiekader-verantwoorde-inzet-van-algoritmen, Algorithm Research Framework of the Dutch Government Audit Agency (ADR) https://www.rijksoverheid.nl/documenten/rapporten/2023/07/11/onderzoekskader-algoritmes-adr-2023 and the Algorithms Assessment Framework of the Court of Audit https://www.rekenkamer.nl/onderwerpen/algoritmes-digitaal-toetsingskader.

Appendix A – Further references

Algorithm register

> Documentation 'Slimme bijstandscheck levensonderhoud' [Smart check sustenance] ' https://algoritmeregister.amsterdam.nl/ai-system/onderzoekswaardigheid-slimme-check-levensonderhoud/1086/.

Societal attention for algorithms

Furthermore, several (research) journalistic articles have been published about bias in algorithms. For example:

> https://www.vpro.nl/argos/media/luister/argos-radio/onderwerpen/2021/In-het-vizier-van-het-algoritme-.html
> https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
> https://www.lighthousereports.com/methodology/suspicion-machine/.

Academic data science

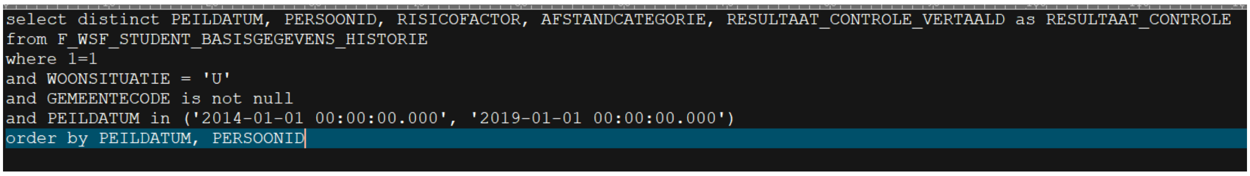
> Qualifying types of proxy discrimination: Michael Carl Tschantz. 2022. What is Proxy Discrimination? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1993–2003. https://doi.org/10.1145/3531146.3533242.

> Tension between AI and European non-discrimination law: Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." Computer Law & Security Review 41 (2021): 105567.

Appendix B – Datawarehouse query

An example of a SQL query to request the following data of students living away from the data warehouse for the reference date 01-01-2014:

- Personal ID (pseudo anonymized)
- Risk factor
- Distance to ouder(s)
- Result of the check (if available).



```
select distinct PEILDATUM, PERSOONID, RISICOFACTOR, AFSTANDCATEGORIE, RESULTAAT_CONTROLE_VERTAALD as RESULTAAT_CONTROLE
from F_WSF_STUDENT_BASISGEGEVENS_HISTORIE
where 1=1
and WOONSITUATIE = 'U'
and GEMEENTECODE is not null
and PEILDATUM in ('2014-01-01 00:00:00.000', '2019-01-01 00:00:00.000')
order by PEILDATUM, PERSOONID
```

Figure 10 – Example of a query from DUO data warehouse to retrieve personal number (pseudo anonymized), risk factor, distance to parent(s) and result of the check (if available).

Appendix C – Additional information statistical analysis

About the Z-test and Fisher's exact test as used in 3.1 Quantitative analysis.

First, a so-called Z-test is used to compare two percentages. Let p_A and p_B be the unduly use percentage for respectively group A and group B, with size N_A and N_B. We define the test statistic Z as follows:

$$Z = \frac{p_A - p_B}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{N_A}+\frac{1}{N_b})}}, \text{where } \hat{p} = \frac{N_{A,unduly}+N_{B,unduly}}{N_A+N_B}.$$

When the sample size is large enough, and H_0:p_A=p_B applies, then Z follows a standard normal distribution – this is a normal distribution with a mean of 0 and variation of 1.

A p-value can then be derived from the distribution of a standard normal distribution. This p-value informs us what the probability is that we will happen to observe p_A and p_B, given that H_0 is true. When calculating this p-value, the group size of A and B is considered. The significance level of the hypothesis test is set at 5%, which means that if the p-value is less than 0.05, the null hypothesis is rejected. This means that $p_A>p_B$, in other words: the unduly use percentage for the different groups are not the same. If the p-value is greater than 0.05, there is no reason to suspect that the unduly use percentages differ for the groups.

In some cases, when the sample size is very small, using the Z test can lead to incorrect conclusions. To prevent this, Fisher's exact test is applied as a check. The test is also a statistical test for comparing two percentages and is often recommended when working with small samples. This may be relevant to some of the comparisons – for example, there are often relatively few observations of certain types of education in the samples. The results of Fisher's exact test are available when contacting Algorithm Audit.

Because the subcategories are compared against each other for each variable of education, age and distance, there is a risk of incorrect detection of statistical (in)significance. The Bejamini-Yekutieli procedure can offer a solution in this case. The results of this procedure are available when contact Algorithm Audit.

Appendix D – Document list

A. L: …/Algorithm Audit/2023-10-10

Reference	Work name	Document name
[A.1]	Woo-besluit 2023	1 - Getekend Woo-besluit 28-6-23 Woo-verzoek Investico
[A.2]	Bijlagen Woo-verzoek 2023	2 - Bijlage 2 Inventarisatielijst Woo-verzoek Investico
[A.3]	Beleid Controle Uitwonendenbeurs	3 - Beleid misbruik uitwonende beurs passages m.b.t. fraude selectie, onderzoek, vaststellen fraude, huis- buurtbezoek
[A.4]	Procesbeschrijving CUB	4 - Procesbeschrijving Uitvoeren controle Misbruik Uitwonendenbeurs
[A.5]	Richtlijnen voor controleurs 2022	5 - Richtlijnen voor controleurs 2022
[A.6]	Werkinstructies afhandelen & verwerken CUB	6 - Werkinstructie WI MUB afhandelen & verwerken
[A.7]	Onderzoeksopzet CUB 2023	7 - Onderzoeksopzet CUB versie 02 28062023 CONCEPT
[A.8]	Opdrachtbevestiging onderzoek privacyaspecten CUB aan de ADR 2023	8 - Opdrachtbevestiging onderzoek Beheersing privacyaspecten proces uitwonendencontrole 21122022
[A.9]	missing	missing
[A.10]	Organogram 2023	10 - Organogram OVG 2023 (versie 1.2)
[A.11]	Rapportage CUB 2013	11 - Rapportage MUB 2013 v0.2
[A.12]	Rapportage CUB 2014	12 - Rapportage MUB 2014 v1.0
[A.13]	Rapportage CUB 2015	13 - Rapportage MUB 2015 v1.0
[A.14]	Rapportage CUB 2016	14 - Rapportage MUB 2016 v1.0
[A.15]	Rapportage CUB 2017	15 - Rapportage MUB 2017 v1.0
[A.16]	Kamervragen Spekman 2009	16 - 2009 Kamervragen Spekman cs ah-tk-20082009-3686
[A.17]	Actieplan CUB 2009	17 - 200911 Actieplan misbruik uitwonendenbeurs
[A.18]	Voortgangsrapportage 2018	18 - 20100701 Voortgangsrapportage inzake de uitvoering van het Actieplan misbruik uitwonen-

Reference	Work name	Document name
[A.19]	Kamerbrief 2012	19 - kamerbrief-over-eindrapportage-actie-plan-misbruik-uitwonendenbeurs
[A.20]	Rapportage CUB 2012	20 - Rapportage MUB 2012 v1.1
[A.21]	Eindrapportage Intensiveringen IHS en CUB 2018	21 - Eindrapportage IHS en Mub 2018 1.0
[A.22]	Avg verwerkingsgrond uitwonende-ontrole	22 - ZZZ GGL Henl - uitwonendencontorle 000
[A.23]	Avg verwerkingsgrond procedure uitwonendecontrole	23 - ZZZ GGL Henl – uitwonend prodecure 000
[A.24]	Informatieverkenning CUB 2020	24 - 20200629 Informatieverkenning H&I proces MUB
[A.25]	Ordeningsplan CUB 2022	25 - Kopie van 20221118 Ordeningsplan HI versie 0.4 netto lijst
[A.26]	Datawarehouse overzicht	26 - Datawarehouse SFS
[A.27]	Kopie risicocodering 2023	27 - Kopie van 20230704 Risicocodering
[A.28]	Mogelijkheden invoer resultaat controle	28 - Mogelijkheden invoer resultaat controle
[A.29]	Resultaten controle mogelijkheden	29 - resultaten controle mogelijkheden
[A.30]	Resultaten Den Haag 2 2013	30 - resultaten Den Haag 2
[A.31]	Resultaten Den Haag 3 2013	31 - resultaten Den Haag 3
[A.32]	Resultaten Den Haag 2013	32 - resultaten Den Haag
[A.33]	Mailwisseling uitbreiding risicoprofiel 2023	33 - Risicoprofiel MUB
[A.34]	Overzicht Den Haag	34 ----
[A.35]	Risicocodering 2021	35 - 20210630 Risicocodering
[A.36]	Snippet datalevering #1	36 - acess aangeboden ter controle
[A.37]	Basisbestand Utrecht	37 – basisbestand utrecht
[A.38]	Snippet datalevering #2	38 - basisbestand
[A.39]	Informatie en bron analyse rapport CUB 2015	39 - DSF080 Selectie uitwonende fraude v0.2

Reference	Work name	Document name
[A.40]	Rapportage aselecte steekproef 2014	40 - Rapportage a-selecte steekproef MUP 0.1
[A.41]	MUB-AP	41 – MUP-AP
[A.42]	AP 20230705	42 Notulen 8 juni 2023 en 25 mei 2023
	Notulen 2021 en 2022	43 - Notulen 2021 2022, Notulen 20211111
[A.44]	Kalender 2022	44 - MO kalender 2023 concept (geclusterd)
[A.45]	Viermaandse rapportage MUB 2023	45 - 20230612 1ste viermaandsrapportage MUB regulier en MUB HBB tijdvak jan-apr 2023 v05TK
[A.46]	Uitvoeren Controle MUB	46 - Uitvoeren controle misbruik Uitwonendenbeurs (2.0)
[A.47]	Risicocodering 2010	47 - Risicocoderingen misbruik uitwonendenbeurs
[A.48]	Kwalitatieve onderbouwing art.11	48 - Kwalitatieve onderbouwing MO-inventarisatie art 11 HOS 2022
[A.49]	Kwalitatieve onderbouwing art.12	49 - Kwalitatieve onderbouwing MO-inventarisatie art 12 HOS 2022
[A.50]	Kwalitatieve onderbouwing art.13	50 - Kwalitatieve onderbouwing MO-inventarisatie art 13 HOS 2022
[A.51]	Aselecte steekproef 2010 verdeling	51 - A-selecte steekproef def
[A.52]	Aselecte steekproef 2010 verdeling fout	52 - A-selecte steekproef fout def
[A.53]	Sjabloon Beheersdocument	53 - IAR Misbruik Uitwonendenbeurs (MUB) v.1.1

B. L: .../Algorithm Audit/2023-10-10/41

Reference	Work name	Document name
[B.41-1]	Properties risicofactoren 2018	----
[B.41-2]	Informatie Analyse Rapport - Misbruik Uitwonendenbeurs	IAR - Misbruik Uitwonendenbeurs (MUB)
[B.41-3]	Inhoud H SWF	----
[B.41-4]	Jira HENI fiPF3&	----
[B.41-5]	Query documentatie	----

C. L: .../Algorithm Audit/2023-10-10/42 - AP 20230705/

Reference	Work name	Document name
[C.42-1]	Notulen 05-01-23 MUB overleg	20230105 Verslag MUB overleg
[C.42-2]	Notulen 02-02-23 MUB overleg	20230202 Verslag MUB overleg
[C.42-3]	Notulen 15-02-23 MUB overleg	20230215 Verslag MUB overleg
[C.42-4]	Notulen 16-03-23 MUB overleg	20230302 Verslag MUB overleg
[C.42-5]	Notulen 30-03-23 MUB overleg	20230316 Verslag MUB overleg
[C.42-5]	Notulen 30-03-23 MUB overleg	20230323 Verslag MUB overleg
[C.42-6]	Notulen 17-04-23 MUB overleg	20230417 Verslag MUB overleg
[C.42-7]	Notulen 11-05-23 MUB overleg	20230511 Verslag MUB overleg
[C.42-8]	Notulen 25-05-23 MUB overleg	20230525 Verslag MUB overleg
[C.42-9]	Notulen 08-06-23 MUB overleg	20230623 Verslag MUB overleg
[C.42-10]	2021 Leidraad selecteren	2021 leidraad selecteren
[C.42-11]	2021 Werkinstructies	2021 Werkinstructie en actielijst MUB
[C.42-12]	Access zoeken	Access zoeken
[C.42-13]	Woonoppervlakte tool	BAG Viewer woonoppervlakte
[C.42-14]	BRP documenten	BRP ...
[C.42-15]	Aanlevering formulier	FORMAT AANLEVERING NIEUWE STIJL 2023
[C.42-16]	Voorbeeld inschrijving	Studie inschrijving
[C.42-17]	Overzicht bestand	WI MUB selecteren versie 0.1

D. L: …/Algorithm Audit/2023-10-10/43

Reference	Work name	Document name
[D.43-1]	Notulen 23-04-20 MUB overleg	Notulen 23-04-20 MUB overleg
[D.43-2]	Notulen 16-09-21 MUB overleg	Notulen 16-09-21 MUB overleg
[D.43-3]	Notulen 30-09-21 MUB overleg	Notulen 30-09-21 MUB overleg
[D.43-4]	Notulen 14-10-21 MUB overleg	Notulen 14-10-21 MUB overleg
[D.43-5]	Notulen 28-10-21 MUB overleg	Notulen 28-10-21 MUB overleg
[D.43-6]	Notulen 11-11-21 MUB overleg	Notulen 11-11-21 MUB overleg
[D.43-7]	Notulen 25-11-21 MUB overleg	Notulen 25-11-21 MUB overleg
[D.43-8]	Notulen 09-12-21 MUB overleg	Notulen 09-12-21 MUB overleg

E. L: …/Algorithm Audit/2023-10-18/

Reference	Work name	Document name
[E.1]	Mail agenda stuurgroep 22-02-2010	Agenda Stuurgroep uitwonendenbeurs 22022010.doc
[E.1-1]	Notitie Risicoprofielen 22-02-2010	DUO-notitie Risicoprofielen
[E.1-2]	Agenda stuurgroep 22-02-2010	Agenda Stuurgroep uitwonendenbeurs 2202010.doc
[E.1-3]	Besluitenlijst Stuurgroep Misbruik Uitwonendenbeurs 25-01-2010	Concept Besluitenlijst Stuurgroep Misbruik Uitwonendenbeurs 25 jan 2010.doc
[E.1-4]	DUO Opsporing & Handhaving	DUO-opsporing&handhaving.doc
[E.2]	Nota Misbruik controlevolume 2015 en jaarrapportage 2013	----
[E.3]	Opdrachtbrief uitvoering wetsvoorstel Studievoorschot	----
[E.4]	Verslag uitvoeringsoverleg 08-07-2016	----
[E.5]	Mail stuurgroep vergadering MUB 15-03-2010	FW Stuurgroep vergadering ‘misbruik uitwonendenbeurs’ 15 maart 2010

Reference	Work name	Document name
[E.6]	Mail onderbouwing uitvoeringskosten Studievoorschot	Onderbouwing uitvoeringskosten Studievoorschot
[E.7]	Agenda, actie- en besluitenlijst uitvoeringsoverleg 08-07-2016	----
[E.8]	Mail stuurgroep MUB 21-01-2010	Stuurgroep Misbruik Uitwonendenbeurs
[E.8-1]	Agenda Stuurgroep misbruik uitwonendenbeurs 25-01-2010	Agenda Stuurgroep uitwonendenbeurs 25012010.doc
[E.8-2]	Besluitenlijst Stuurgroep 14-12-2009	Besluitenlijst Stuurgroep dd 14122009.doc
[E.8-3]	Communicatieplan Misbruik Uitwonenden Beurs 2010	Communicatieplan fraude berus uitwonenden_0.2.doc
[E.8-4]	Notitie afspraken over convenanten voor pilots uitwonendenbeurs	Notitie Stuurgroep 25012010 convenanten.doc
[E.8-5]	Raming kosten Misbruik Uitwonenden Beurs	Raming Project Misbruik Uitwonendenbeurs-21012010.doc
[E.8-6]	Notitie ter voorbereiding voor overleg met BZK (toezichthouder en adresadministratie)	Notitie ter voorbereiding voor overleg met BZK
[E.8-7]	Planning Stuurgroepagenda	Agenda- planning Stuurgroep.doc
[E.9]	Mail stuurgroep MUB 24-06-2010	Stuurgroepvergadering Project Misbruik Uitwonendenbeurs 28 juni
[E.9-1]	Agenda Stuurgroep MUB 28-06-2010	Agenda Stuurgroep 28.06.10.doc
[E.9-2]	Besluitenlijst Stuurgroep 19-04-2010	Besluitenlijst Stuurgroep Misbruik uitwonendenbeurs 19 april 2010
[E.9-3]	Concept artikelen wetsvoorstel aanpak MUB 24-06-2010	1759-artikelen-24-6-10.doc
[E.9-4]	Overdrachtsdocument 22-06-2010	Overdrachtsdocument 22.06.10.doc
[E.9-5]	Nota voortgangsrapportage Actieplan uitwonendenbeurs 2010	Ter ondertekening 15.06.10.doc
[E.9-6]	Voortgangsrapportage Actieplan uitwonendenbeurs 2010	Voortgangsrapportage TK 24 06 10 (incl WJZ+DUO+HB) (laatste versie) (2) (2).doc

F. L: .../Algorithm Audit/2023-11-1/

Reference	Work name	Document name
[F.1]	Formulier Uitwonendentoelage 2006	6 formulier uitwonendentoelage 2006
[F.2]	Agenda Workshop Misbruik Uitwonendenbeurs	Agenda Workshop misbruik uitwonendenbeurs
[F.3]	Brief Keupink	Brief melding fraude Keupink
[F.4]	Controlebeleid IB-Groep	Controlebeleid uitwonendheid IB-Groep
[F.5]	PowerPoint Workshop maart 2010	DUO Workshop uitwonendheid 25032010
[F.6]	Gegevens voor ontwikkeling risicoprofiel	Gegevens voor ontwikkeling risicoprofiel studerenden met een uitwonende beurs
[F.7]	Nota gesprek 13 november 2007	Nota gesprek over fraude 13 november
[F.8]	Notitie	Notitie 06112007 uitwonendencontrole
[F.9]	Proces uitwisseling gemeenten	Proces uitwisseling gemeenten
[F.10]	Verslag workshop 25 maart 2010	Verslag workshop misbruik uitwonendenbeurs 25 maart 2010
[F.11]	Voorgangsrapportage 28 juni 2010	Voortgangsrapportage 28.06.10 (eindversie)

G. L: .../Algorithm Audit/Gegevenslevering/

Reference	Work name	Document name
[G.1]	Steekproef 2014	Tabel 1 – Aselect 2014
[G.2]	Steekproef 2017	Tabel 3 – Aselecte 2017
[G.3]	CUB-2014	Tabel 1 – CUB 2014
[G.4]	CUB-2019	Tabel 2 – CUB 2019

H. Additional documents received after December 14, 2023

Reference	Work name	Document name
[H.1]	Inventarisatie CUB werkwijze	Inventarisatie CUB werkwijze versie 1.0

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