Real-life Performance of Fairness Interventions Supplemental Material

Daphne Lenders
University of Antwerp
Antwerp, Belgium
daphne.lenders@uantwerpen.be

Toon Calders
University of Antwerp
Antwerp, Belgium
toon.calders@uantwerpen.be

A Datasheet for Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to provide a new benchmarking tool for fair Machine Learning algorithms. Different than other datasets typically used for evaluating fair ML, our data contains a fair and biased version of its decision label. Using this, we can check the effectiveness of a fair ML intervention, by checking how well it can predict the fair labels after being trained on the biased ones.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Daphne Lenders and Toon Calders were responsible for collecting the biased labels of this dataset. Both are affiliated with the ADReM Data Lab of the University of Antwerp. The rest of the data (including its fair label) was based on an already existing dataset, which is publicly available online[1].

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the dataset was funded by the University of Antwerp - Research Excellence Center DigiTax.

Any other comments? /

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance in this dataset represents a high school student, following either a Portuguese or Maths course. There is a variety of information available for each student, including, e.g., information about

[1] https://www.kaggle.com/datasets/uciml/student-alcohol-consumption (license CC0)

Supplemental Material for: Lenders, D. & Calders, T. (2023). Real-life Performance of Fairness Interventions: Introducing A New Benchmarking Dataset for Fair ML. In *Proceedings of ACM SAC Conference (SAC'23)*

their studytime, their school absences and their free time behaviour. Special variable of interests are students' sex, their performance on an exam (pass vs. fail) and their predicted performance on that exam. The predicted performance was based on an experiment, where students were presented with some information about the students, based on which they had to make grade predictions. Comparing the predicted exam performance with the actual exam performance we observe clear bias against boys.

How many instances are there in total (of each type, if appropriate)?

Our dataset consists of a total of 856 instances. Note, that the two dataset entries might relate to the same student, whereas one entry corresponds to the student's performance in a Maths course and the other to the performance in a Portuguese course.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The instances in our dataset are sampled from a larger dataset that is publicly available. In sampling from this data, we excluded all students whose grade for the last exam of the course was 0. Further, we randomly sampled 428 from the 430 male instances, and 428 from the 560 female instances. These steps were taken so that in the collection of our biased labels, we could present each participant with four male and four female student profiles, which the participants had to make grade predictions for.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of several features, that are either categorical or numerical in nature.

Is there a label or target associated with each instance? If so, please provide a description.

While other features of the data could potentially be used as a target as well, our intended target label is whether each student passed or failed the third exam of the course they were following. Our dataset consists of both a biased and a fair version of this label. The fair version was obtained, by checking whether a student's grade for the last exam was ≥ 10. There are multiple ways in which the biased label can be obtained and all are based on a human experiment where participants ranked eight student profiles according to their expected performance and predicted their grade for the exam. The first way to obtain the biased label is by checking whether the predicted grade for the exam is ≥ 10. The second way is to look at the ranking position each student was assigned to: the passing-labels of most highly ranked instances were always changed to "true", while they were changed to "false" for the two lowest rated instances.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between individual instances are not made explicit.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended data splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Since our dataset was partly based on an already existing dataset, and partly based on our own experiment there are two types of noise in the data. The first relates to the noise that was already present in the original data, which consists of information of high school students and their exams for a course. A lot of information of the high school students was self-recorded, like, e.g., information on their drinking behaviour or their studytime. Thus it is questionable to which extent the students were truthful in reporting this information. Also, in the collection of this data, the sex of the students was treated as a binary variable. Thus, important information on non-binary gender identity may have been lost. Second, our biased labels also contain some noise, due to the fact that they were gathered through a human experiment. In this experiment 107 participants each made grade predictions for 8 student profiles. Because each participant might have different stereotypes and biases when estimating students' exam performance, the labels may not be very consistent. It should be noted that this noise was intentionally introduced, to reflect the complexity of real-life decision making and discriminatory biases.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

While the main part of our dataset can be used as it is, it is possible to extent it further using the original data it was based on. As this data was publicly available online (license CC0), we included a preprocessed version of it on our kaggle and github page. This data contains some information of the students, that we did not present when asking participants to make grade predictions for them. It is possible to link this information to our collected data, using the indices of both datasets. Because we make both the preprocessed original data and our collected data available online, we can guarantee that both will keep existing without any changes (except changes we might make ourselves).

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, this dataset does not contain any of such data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

As mentioned, our dataset consists of both a fair and biased version of its decision label. Comparing the biased label with the fair one, we see that participants had discriminatory biases (mostly targeted against boys) when making grade predictions. These biases may be offensive. Also, "sex" is treated as a binary variable in our dataset, which was a direct consequence of the binary gender categories used in the original "Student Alcohol Consumption" dataset. While we recognize that this may be offensive to some people (especially those who identify with a non-binary gender category), we emphasize that this does not reflect our own beliefs about gender identities.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Yes, our datast contains information about students' sex, and it is also possible to infer their age using the original data it was based on. In our dataset half of the students (i.e. 428 out of 856) are male and the other half are female. The ages of students range between 15 and 22. There are 169 students who are 15 years old, 230 who are 16 years, 230 who are 17 years, 172 who are 18 years, 41 who are 19 years and 15 students who are 20 years or older (9 students that are 20, 3 that are 21 and 2 that are 22).

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, it is not possible to identify individuals from out dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The sex of the students might be considered as sensitive data. Further, in the original data there is some information about the parents of the students including their job and education level. Both might give some indication of the socio-economic status of the students. Finally, the data also contains information about the drinking behaviour of students. This information can especially be considered as sensitive, given that not all of the students are of legal drinking age (in Portugal, the country where the data was collected this age is 18+). We emphasize that we do not want to encourage illegal drinking, by distributing our dataset. Still, we also highlight that none of the students of which the data was collected have to fear unwanted consequences for their actions, as the data is completely anonymous and cannot be traced back to any individuals.

Any other comments? /

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The information about the students as well as their grades were gathered in a previous study by Cortez and Silva. The data was gathered from two high schools in Portugal, for more information on how they collected this data we refer to their paper[2]. We collected the biased labels for this dataset through a survey, where participants were asked to make grade predictions for students, based on short descriptions about them. In the paper corresponding to our dataset, we give a more detailed description of this survey, including the exact task set-up, the materials we used and information about participant recruitment procedures.

[2]Cortez, P. & Silva, A. (2008). Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira Eds., *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We used the survey platform Qualtrics[3] to set up our survey. Our survey was based on multiple pilot studies, were we tested its clarity and suitability.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

As previously mentioned the data we collected our biased labels for, was sampled from a slightly bigger dataset that was already publicly available online. We excluded all students with a grade equal to 0 from this dataset, and randomly sampled 428 male and 428 female instances from the rest of the data.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The collection of our data was based on voluntary participation. Further, we also put our survey on the Survey Exchange platforms SurveySwap and SurveyCircle[4]. Here participants who filled out our survey were rewarded with survey-responses for their own survey.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The biased labels were sampled from January until March 2022. This timeframe does not match the creation of the data associated with the instances: the information about each student was recorded in 2005 and 2006.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

To collect our biased labels we obtained ethical approval by the *Ethics Committee for the Social Sciences and Humanities* of the University of Antwerp, under reference number SHW_21_128. To start the review process we gave a detailed description of our experimental setup, possible risks involved in participation and the way in which we would store and process the collected data. We received a positive outcome for this review process.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset relates to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We obtained the data about the students through an already existing dataset (this existing dataset was based on asking individuals directly for their information). The biased labels were also obtained directly, by specifically asking our participants to make grade predictions for the students.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and

[3] `https://www.qualtrics.com`

[4] `https://surveyswap.io/` and `https://www.surveycircle.com/en/surveys/`

provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Again, for details on the data collection of the original data (with the information about the students) we refer to the paper of Cortez and Silva. To collect the biased labels for this dataset, participants first had to fill out a consent form, before they could start the survey. Here it was also explained that the collected data would be made publicly available.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, our participants had to consent to the collection and use of their data. The consent form can be seen when following our survey link: `https://uantwerpen.eu.qualtrics.com/jfe/form/SV_5gOFzeF3xtGSinI`

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

As the participants' data was collected anonymously, and survey responses could not be matched to individuals' identity, we did not provide participants with a mechanism to revoke their consent.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Because we are dealing with anonymous data, no formal analysis has been conducted.

Any other comments?

/

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Before collecting our biased decision labels based on the original dataset, we applied some pre-processing steps on it. All steps are described below:

- **Parent's education** - This was a variable that was not part of the original dataset. Instead the dataset consisted of two variables, namely *Fedu* and *Medu* to respectively denote the father's and the mother's education level. We obtained our variable *Parent's education* by taking the maximum of the two.
- **Studytime** - Originally, this variable consists of four levels (*less than 2 hours* vs. *2-5 hours* vs. *5-10 hours* vs. *more than 10 hours*). Because there were little students with a studytime of longer than 10 hours, we decided to merge the latter two levels
- **Absences** Originally, this variable ranges from 0 to 93. Since high values for this variable were quite uncommon, we decided to bin all absences ≥ 7 into one level called *More than 6*
- **Freetime** - Originally, this variable ranged from 1 (very low) to 5 (very high). We binned this variable into three categories, where 1 & 2 are binned into level, and 4 & 5 are binned as well

- **Gooing out** - Again, this variable originally ranged between 1 and 5. We decided to bin the last two levels (4 & 5)
- **Alcohol Consumption** - The original dataset consisted of two variables denoting the student's alcohol consumption namely *Walc* (alcohol consumption in the weekend) and *Dalc* (alcohol consumption throughout the week). For our experiment we only showed the students' alcohol consumption in the weekend. This variable originally consisted of 5 levels, where we binned the latter two ones

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, a not-preprocessed version of the data is being provided on our kaggle page as well.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the (Python) code that we used to preprocess the data is available on github: `https://github.com/calathea21/settingUpBenchmarkCollection`

Any other comments?

/

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has been used for our own benchmarking experiment. Here we tested the effectiveness of several fairness interventions, by checking how well they can predict the fair labels of the dataset after being trained on the biased ones.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No such repository is available.

What (other) tasks could the dataset be used for?

The dataset was created mostly for benchmarking studies, like the one previously described. However, there are some other interesting use cases for the fair Machine Learning community. It could for instance be interesting to use the data to better understand the dynamics behind discriminatory decision making, by checking how exactly the fair labels relate to the biased ones, and if there are some clear patterns in which discrimination/favouritism occurs. This knowledge could then also be exploited to create better fairness interventions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Though the dataset can be a useful tool for benchmarking fair ML algorithms, users should be careful not to overgeneralize their findings to other decision tasks. A fairness intervention that performs well on our dataset, is not guaranteed to work well on others, and may not be as fair/accurate as intended.

Also, users should be cautious with the fact that we treat “sex" as a binary variable in our dataset. If researchers start developing new fairness algorithms based on our data, they should take into account that our data does not provide information on the types of discrimination non-binary people might face.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Our dataset was made to test the effectiveness of fairness interventions targeting label bias. We say that our collected version of the decision labels (i.e. whether students pass an exam or not) contain label bias as they do not accurately reflect whether the students actually passed or not, and where instead the result of a biased decision process. Label bias is different than other forms of biases, like selection bias or historical bias. Hence, fairness interventions that specifically target this kind of bias, should not be benchmarked on our dataset.

Any other comments? /

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Our dataset is publicly available online (under license CC BY-SA 4.0[5]) meaning that any third party can access and use it, as long as they give appropriate credit, provide a link to the license and indicate if changes were made to the data.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is distributed via kaggle: `https://www.kaggle.com/datasets/daphnelenders/performance-vs-predicted-performance`. It’s DOI is: `10.34740/kaggle/dsv/3689065`

When will the dataset be distributed?

Our dataset is already publicly available on kaggle.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Our dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit `http://creativecommons.org/licenses/by-sa/4.0/` or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No restrictions has been imposed on the data.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[5] `http://creativecommons.org/licenses/by-sa/4.0/`

No export controls or other regulatory restrictions apply to the dataset.

Any other comments?

/

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Daphne Lenders will be responsible for supporting, hosting and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Daphne Lenders can be contacted by her institutional email: daphne.lenders@uantwerpen.be

Is there an erratum? If so, please provide a link or other access point.

As for now there is no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset was based on an existing dataset and on a one-time human experiment. Unless new experiments are conducted, or mistakes in the current data are found, it is unlikely that the dataset will be updated.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

There are no limits on the retention of the data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

If new versions of the dataset will be made available, all older versions will still be accessible through the kaggle website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No formal mechanism for contributing to our dataset consists yet, as for now it is unlikely to be expanded. If there are suggestions for extending or augmenting the data, it is possible to send an email to daphne.lenders@uantwerpen.be.

Any other comments?

/

B Proof-of-Concept Study

In this section we are going to give additional information about the study design of our proof-of-concept as well as its results. We set up this study as a precedent for our main experiment, where we test whether people have inherent biases against boys when judging their school performance and whether this bias can be triggered or amplified through stereotype activation. The study was approved by the Ethics Committee of the University of Antwerp, under reference number SHW_21_128.

B.1 Experimental Design

The experimental design of our proof-of-concept study is illustrated in Figure 1. In the main task, we presented 8 student profiles to the participants, containing basic information about each student, for which participants had to make grade predictions. The main manipulation was that part of the participants were presented with a version of these profiles for which the sex was swapped. In other words, a profile that belongs to a female student is, in this condition, said to belong to a male student (and vice versa). By comparing the predicted grades across these two conditions, we could determine whether participants have inherent biases against boys when estimating students' performance. To see whether bias against boys can be triggered, our second manipulation in the experiment was whether participants were exposed to some form of stereotype activation prior to the prediction task. We included three types of stereotype activation, which we will describe in sections. Note, that the complete experimental design, including the type of materials shown to the participants, the task set-up, as well as the task description, were based on a total of three pilot studies. In these pilot studies we let 4-6 participants complete the most up-to-date version of the online survey, and based on their responses and feedback we iteratively improved the overall study design.

Figure 1: The experimental setup to see if participants have inherent bias against boys or whether this bias can be triggered through stereotype activation

B.2 Materials

For the main task of the study, each participant was presented with the same eight student profiles extracted from the "Student Alcohol Consumption" dataset. Half of these profiles belong to male, and the other half to female students. While the original dataset contains more than 30 attributes to describe each student, we chose to only present eight of them per profile, to not overload participants with information. We chose attributes that had high variability between students and that could in legitimate or stereotypical ways be associated with school performance. In Figure 2 we show some examples of the presented profiles, along with the original sex of each student and their obtained grade.

(Org. Girl with Grade = 8/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - between 2 and 5 hours
Absences - 3
Going out - Twice a week
Alcohol consumption - moderate
Freetime - average
In a relationship - no

(a) Student Profile 1

(Org. Girl with Grade = 10/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - less than 2 hours
Absences - 1
Going out - Twice a week
Alcohol consumption - very high
Freetime - high
In a relationship - no

(b) Student Profile 2

(Org. Girl with Grade = 13/20)
Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 4
Going out - Once a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

(c) Student Profile 3

Figure 2: Some of the profiles that participants had to make grade predictions for

As can be seen in this Figure, each student profile was presented in a tabular format. To convey the sex of each student, we randomly assigned each profile to one of four male/female names, depending

on the students' sex in the original dataset and whether the profile was in the 'sex-swapped' condition or not. The four male and female names were chosen to represent common names in English speaking countries. In the experiment all eight profiles were presented on one page, where the order of presentation was randomized. On top of this page, participants were presented with a list of all student names followed by a blank field. They were asked to use a drag-and-drop interface to rank the students according to their expected performance. Additionally, they were prompted to enter specific grade predictions (ranging from 0 to 20) in the blank field next to each students' name.

Before the grading task, participants were exposed to one of three forms of stereotype activation:

1. **None** - Baseline condition in which no extra information is presented.
2. **CaseBased** - Here we presented participants with three student profiles along with the grades of the students. Two profiles belong to male students with low grades (5/20 and 10/20), while one belongs to a female student with a high grade (17/20).
3. **Statistics** - Here we presented a graph showing statistics about how some risk factors affect boys' chance to pass an exam more than they affect girls' passing chances. One presented risk factor was, e.g., having more than 6 school absences, which makes boys $\sim$15% more likely to fail, while girls only $\sim$4% more likely. All risk factors were chosen such that none of the presented profiles contained any of these risk factors.

As can be seen, both the CaseBased and Statistics condition contained stereotypical information against boys. We were interested to see whether presenting this information prior to the prediction task would differently affect the grade predictions for male and female students.

B.3 Participants

The participants were recruited through social media channels and the survey exchange platforms SurveySwap and SurveyCircle. To participate, a consent form needed to be filled out. All responses were completely anonymous, and after a quality controls[6], to filter out short responses and respondents who did not follow the survey instructions correctly, we were left with data of 157 participants.

In table 1 we show how the participants were distributed over the different conditions of our experiment. In table 2 the participant counts per gender and age category are shown. Figure 3 gives an overview of the nationality of the participants.

Table 1: Number of Participants Across the Conditions

Datatype \ Stereotype Act.	None	Case Based	Statistics
Original	N = 28	N = 28	N = 27
Sex-Swapped	N = 29	N = 30	N = 27

Table 2: Number of participants by gender and by age

Gender	Count	% of total
Female	107	63.3%
Male	60	35.5%
Prefer not to say	2	1.2%

Age	Count	% of total
18 - 24	97	57.4%
25 - 34	63	37.3%
35 - 44	7	4.1%
65 - 74	1	0.6%
Prefer not to say	1	0.6%

[6]see: `https://github.com/calathea21/analyzing_proof_of_concept`

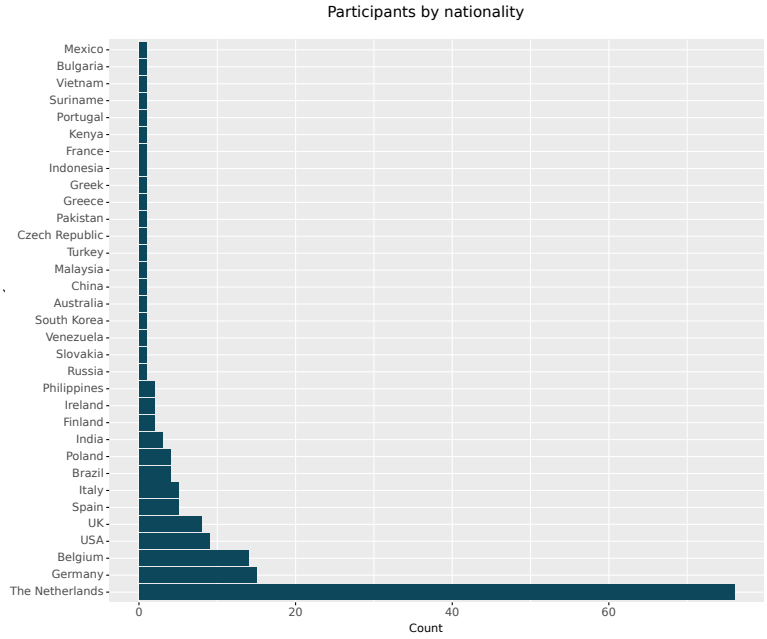


Figure 3: Participants by Nationality

B.4 Data Analysis

On the grade predictions of all eight student profiles we applied the Align Rank Transform procedure with an ANOVA, a nonparametric factorial analysis technique introduced by Wobbrock et al[7]. As factors we used “Stereotype Activation" and “Sex" (i.e. whether a profile was said to belong to a male or female student), while the “Predicted Grade" for each student profile was used as the dependent variable. For post-hoc tests for pairwise comparison, we used Tukey correction. Note, that we chose to not apply multiple comparison correction for the different ANOVAs. Using a significance cut-off value of 0.05 and conducting eight ANOVAs each testing three hypothesis, this results into 1.2 expected false positives (24*0.05). Given that our proof-of-concept study is more of preliminary nature and we also did not want to compromise on the statistical power of our analysis, we deemed this as acceptable.

B.5 Illustration of Most Important Results

In this section we are going to demonstrate the different types of effect we found on three student profiles. We refer to Appendix B.6 for the statistical test results on all eight profiles. The results for this section are visualized in Figure 4, where Profile 1, 2 and 3 correspond to the profile descriptions in Figure 2.

[7]Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143-146).

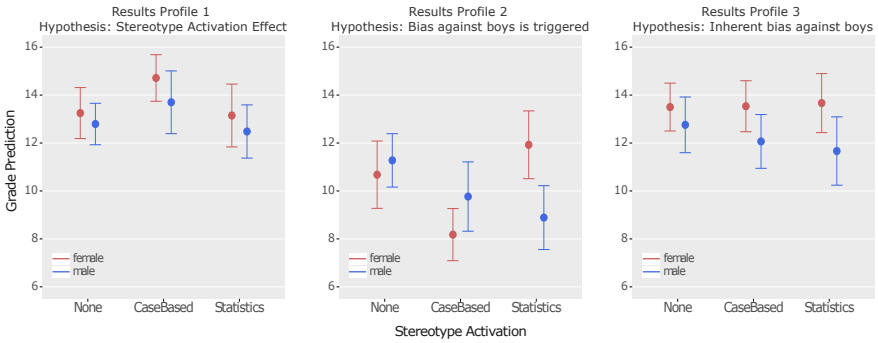


Figure 4: Participants' grade predictions can depend on the stereotype activation they were exposed to and on the presented sex of the student profile (as well as on the interaction between both)

For student profile 3 (right graph of Figure 4), we found a significant effect of "sex" on grade prediction. Averaged across all stereotype activation conditions, higher grades were predicted for this profile if it was tied to a girl's rather than a boy's name ($F(1, 163) = 5.255$, $p = 0.023$), with a mean difference of 1.40. This finding confirms our hypothesis that in some cases humans have inherent stereotypes against male students and expect them to perform less well in high school than girls.

On the grade predictions on student profile 1 (left graph of Figure 4) we found a significant effect of "stereotype activation" on grade predictions ($F(2, 163) = 4.031$, $p = 0.020$). Averaging over the grades assigned to the male and female version of the profile, participants who were exposed to the "CaseBased" condition gave significantly higher grades than participants exposed the "Statistics" stereotype activation (with a mean difference of 1.392). A similar effect was found in student profile 2 (middle graph of Figure 4, $F(2, 163) = 5.157$, $p = 0.007$); here participants presented with the CaseBased condition predicted significantly lower grades than participants in the other conditions. Even though we did not make any hypotheses about the effects of stereotype activation alone, these observations were interesting to see. We hypothesise that an anchoring effect occurred, where information about the grades of other students (as presented in the CaseBased condition) influenced participants' subsequent grade predictions [8]. Differently than expected, this effect does not influence the grade predictions for male and female students differently. This might be the case because in the CaseBased condition only three student profiles (2 male, 1 female) were shown. The presented sex-differences in their grades might have been too subtle, for stereotypes to be activated.

For student profile 2, we observed, next to the significant effect of "stereotype activation" alone, a significant interaction effect of "sex" and "stereotype activation" on the grade predictions. In other words, participants assign different grades to the male and female version of this profile, depending on which stereotype activation was presented ($F(2, 163) = 8.402$ $p = 0.001$). In this case, the difference in female and male grade for participants under the "Statistics" condition ($diff_{female - male} = 3.037$) is higher than under the "CaseBased" ($diff_{female - male} = -1.588$) and "None" conditions ($diff_{female - male} = -0.597$). This confirms our hypothesis that biases against boys are not always inherent but can be triggered, in particular through the "Statistics" stereotype activation.

One more general finding of our study is that significant effects of "stereotype activation" or "sex" (or their interaction), were only found on certain profiles. In particular, we observed that no effects occurred on "stereotypically good" profiles, of students with e.g. high amount of study time or low alcohol consumption (see Appendix, profile 4 and 8). Even though we did not go into a deeper analysis of this, it confirms our hypothesis that the occurrence of bias does not only depend on the sex of the students, but also on their other, more complex characteristics. While a more elaborate study is needed to generalize our findings to real-life human behaviour, our results show that the experimental

[8] Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, 40(1), 35-42.

setup of our study is appropriate to elicit interesting biases in human decision makers. As some of these biases are discriminatory, we deemed the setup as useful for our main study.

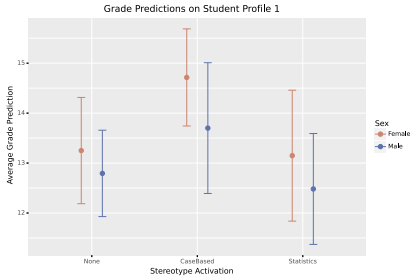
B.6 All Statistical Results

Now that we have illustrated the type of main- and interaction effects we found in our proof-of-concept study, we will show the statistical results on all the student profiles.

B.6.1 Profile 1

(Org. Girl with Grade = 8/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - between 2 and 5 hours
Absences - 3
Going out - Twice a week
Alcohol consumption - moderate
Freetime - average
In a relationship - no

(a) Student Profile 1



(b) Average grade predictions

Figure 5: Results on Student Profile 1

Table 3: Results Statistical Test Profile 1

	Sum of Squares	df	Mean Square	F	p
Datatype	6217	1	6217	2.570	0.111
Stereotype Activation	18868	2	9434	4.0309	0.020
Datatype * Stereotype Activation	14.9	2	7.47	0.0030	0.997

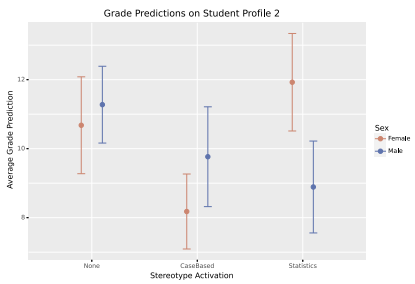
Table 4: Results Post Hoc Test Profile 1

Comparison			Mean Difference	SE	df	t	p_{tukey}
Stereotype Activation		Stereotype Activation					
CaseBased	-	None	21.30	9.03	163	2.360	0.051
CaseBased	-	Statistics	23.17	9.15	163	2.531	0.033
None	-	Statistics	1.87	9.19	163	0.203	0.978

B.6.2 Profile 2

(Org. Girl with Grade = 10/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - less than 2 hours
Absences - 1
Going out - Twice a week
Alcohol consumption - very high
Freetime - high
In a relationship - no

(a) Student Profile 2



(b) Average grade predictions

Figure 6: Results on Student Profile 2

Table 5: Results Statistical Test Profile 2

	Sum of Squares	df	Mean Square	F	p
Datatype	637	1	637	0.260	0.611
Stereotype Activation	23819.80	2	11909.90	5.157	0.007
Datatype * Stereotype Activation	37551.4	2	18775.7	8.4024	0.001

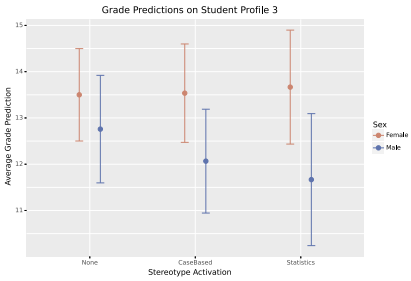
Table 6: Results Post Hoc Test Profile 2

Comparison			Mean Difference	SE	df	t	p_{tukey}
Stereotype Activation		Stereotype Activation					
CaseBased	-	None	-26.95	8.97	163	-3.006	0.009
CaseBased	-	Statistics	-22.33	9.09	163	-2.456	0.040
None	-	Statistics	4.63	9.13	163	0.507	0.868

B.6.3 Profile 3

(Org. Girl with Grade = 13/20)
Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 4
Going out - Once a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

(a) Student Profile 3



(b) Average grade predictions

Figure 7: Results on Student Profile 3

Table 7: Results Statistical Test Profile 3

	Sum of Squares	df	Mean Square	F	p
Datatype	12536	1	12535.5	5.255	0.023
Stereotype Activation	1527.5	2	763.7	0.311	0.733
Datatype * Stereotype Activation	3242.391	2	1621.196	0.664	0.516

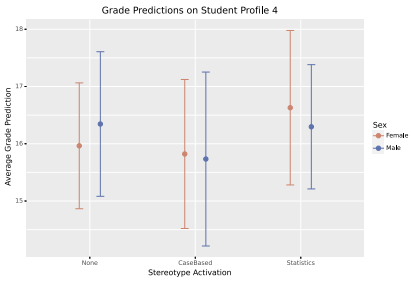
Table 8: Results Post Hoc Test

Comparison			Mean Difference	SE	df	t	p_{tukey}
Datatype		Datatype					
Sex-Swapped (Male)	-	Original (Female)	-17.2	7.52	163	-2.29	0.023

B.6.4 Profile 4

(Org. Girl with Grade = 15/20)
Reason School Choice - Reputation
Parents' education - University
Studytime - more than 5 hours
Absences - 0
Going out - Twice a week
Alcohol consumption - high
Freetime - high
In a relationship - yes

(a) Student Profile 4



(b) Average grade predictions

Figure 8: Results on Student Profile 4

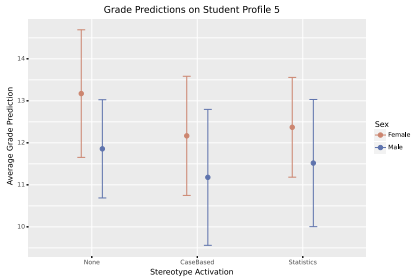
Table 9: Results Statistical Test Profile 4

	Sum of Squares	df	Mean Square	F	p
Datatype	128	1	128	0.0523	0.819
Stereotype Activation	2658.2	2	1329.1	0.5438	0.582
Datatype * Stereotype Activation	4338	2	2169	0.894	0.411

B.6.5 Profile 5

(Org. Boy with Grade = 8/20)
Reason School Choice - Unknown
Parents' education - Middle School
Studytime - between 2 and 5 hours
Absences - 2
Going out - Once a week
Alcohol consumption - very high
Freetime - low
In a relationship - yes

(a) Student Profile 5



(b) Average grade predictions

Figure 9: Results on Student Profile 5

Table 10: Results Statistical Test Profile 5

	Sum of Squares	df	Mean Square	F	p
Datatype	9711.7	1	9711.71	4.045	0.046
Stereotype Activation	3880.80	2	1940.40	0.797	0.453
Datatype * Stereotype Activation	482.29	2	241.15	0.098	0.907

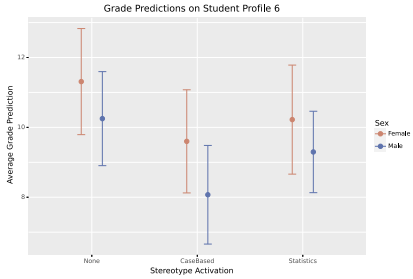
Table 11: Results Post Hoc Test

Comparison			Mean Difference	SE	df	t	p_{tukey}
Datatype		Datatype					
Sex-Swapped (Male)	-	Original (Female)	15.2	7.54	163	2.01	0.046

B.6.6 Profile 6

(Org. Boy with Grade = 10/20)
Reason School Choice - Curriculum
Parents' education - Middle School
Studytime - less than 2 hours
Absences - 4
Going out - Twice a week
Alcohol consumption - high
Freetime - high
In a relationship - no

(a) Student Profile 6



(b) Average grade predictions

Figure 10: Results on Student Profile 6

Table 12: Results Statistical Test Profile 6

	Sum of Squares	df	Mean Square	F	p
Datatype	9710	1	9710	4.041	0.046
Stereotype Activation	22613	2	11307	4.869	0.009
Datatype * Stereotype Activation	1317	2	658	0.270	0.764

Table 13: Results Post Hoc Test Datatype (Profile 6)

Comparison			Mean Difference	SE	df	t	p_{tukey}
Datatype		Datatype					
Sex-Swapped (Female)	-	Original (Male)	15.22	7.55	163	2.01	0.046

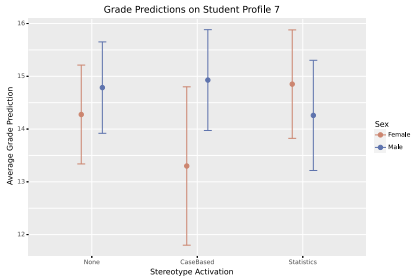
Table 14: Results Post Hoc Test Stereotype Activation (Profile 6)

Comparison			Mean Difference	SE	df	t	p_{tukey}
Stereotype Activation		Stereotype Activation					
CaseBased	-	None	-28.0	8.99	163	-3.11	0.006
CaseBased	-	Statistics	-12.4	9.12	163	-1.36	0.365
None	-	Statistics	15.6	9.15	163	1.71	0.206

B.6.7 Profile 7

(Org. Boy with Grade = 13/20)
Reason School Choice - Close to home
Parents' education - High School
Studytime - between 2 and 5 hours
Absences - 0
Going out - Twice a week
Alcohol consumption - moderate
Freetime - low
In a relationship - yes

(a) Student Profile 7



(b) Average grade predictions

Figure 11: Results on Student Profile 7

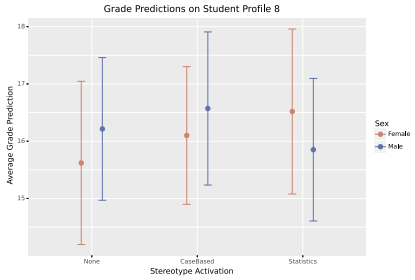
Table 15: Results Statistical Test Profile 7

	Sum of Squares	df	Mean Square	F	p
Datatype	1626	1	1626	0.665	0.416
Stereotype Activation	281	2	141	0.057	0.944
Datatype * Stereotype Activation	7474	2	3737	1.557	0.214

B.6.8 Profile 8

(Org. Boy with Grade = 15/20)
Reason School Choice - Reputation
Parents' education - University
Studytime - more than 5 hours
Absences - 2
Going out - Once a week
Alcohol consumption - moderate
Freetime - high
In a relationship - no

(a) Student Profile 8



(b) Average grade predictions

Figure 12: Results on Student Profile 8

Table 16: Results Statistical Test Profile 8

	Sum of Squares	df	Mean Square	F	p
Datatype	49.5	1	49.5	0.0201	0.887
Stereotype Activation	1303.0	2	651.5	0.265	0.767
Datatype * Stereotype Activation	8253	2	4127	1.724	0.182

C Survey Proof of Concept Study

Collecting a Benchmarking Dataset for fair ML – Proof of Concept

Start of Block: Introduction

Dear participant,

You are invited to participate voluntarily in this research study. Before you consent to participate it is important to read the text below carefully. Here we will give you information about the study itself, as well as your rights in this study.

Any questions you may have because of this information, you can ask by sending an email to: daphne.lenders@uantwerpen.be

Goal and description of the study
This survey is part of a scientific study for my doctoral degree. The goal of this study is to see how well people can predict the study performance of students, given information about their personal background as well as study behaviour.

Duration of the study and task description
Completing this survey should take you 10 – 15 minutes. If you agree to participate, you will first be presented with the task instructions. Reading this information should take about 2-4 minutes. Afterwards, you can start with the main task, which should take about 5-10 minutes. Finally, we will ask you some questions about your demographics, this part of the survey should take around 1 minute.

Voluntary participation
Your participation in this study is strictly voluntary and you have the right to refuse participation. When you accept to take part in this study, you can download this information for safekeeping. Further, you will be asked to digitally give your consent to participate. You can do this by answering the question at the bottom of this page.
You have the right to discontinue your participation at any given time, even after having signed the consent form. You do not have to motivate discontinuing your participation. If you stop participation before the survey is completed, your responses will be deleted and not used in our data analysis.

Benefits
We cannot guarantee you that, if you take part in this study there will be a direct benefit for you.

Re-use of Data
Any data that is collected in this study may be re-used for future scientific studies. This means that we might share your survey responses with other researcher outside the university of Antwerp. These researchers will have access to your anonymous responses, no personal information will be shared with them.

Privacy Policy
If you click on this link we will redirect you to a second survey, where we ask you to give your email address. We will use this address to send you a mail about the study results, once the study has been completed. Your email address will not be linked to the rest of your survey responses. Further, it will be deleted once the debriefing mail has been sent.

If you want to download this information as a PDF click here: Information sheet

P.S. This survey contains a completion code for SurveySwap.io and Survey Circle

---

I have provided my email address in the separate survey. I understand that it will not be linked to the rest of my survey responses, and that it will be deleted, after a debriefing mail has been sent.

- ○ Yes
- ○ No

I have read the information presented above, I understand it, and I consent completely voluntarily to participate in this study

- ○ Yes
- ○ No

End of Block: Introduction

Start of Block: Task Description

In this study we are going to look at real-life data of high school students, who are between 16 and 18 years old. All of these students followed an English course for which they took one exam.
We want to see how well you can predict the students' performance on the exam, given some facts about each student:

- the highest education level of the student's parents
  (lower education vs. middle school vs. high school vs. university)
- the reason the student attends this high school
  (*reputation vs. school's curriculum vs. close to home vs. other)*
- the time the student studied for their exam
  (less than 2 hours vs. 2-5 hours vs. more than 5 hours)
- the number of classes the student missed
- whether the student is in a romantic relationship *(yes vs. no)*
- the amount of free time the student has (*low vs. average vs. high)*
- the number of times the student goes out in a week
  (never vs. once a week vs. twice a week vs. thrice or more)
- the student's alcohol consumption
  (*low vs. moderate vs. high vs. very high)*

For each profile we are going to ask you what grade you expect the student to get for the English exam. Grades range between 0 and 20 (10 is the minimum passing grade) and the expected passing rate for the exam is 70-80%.

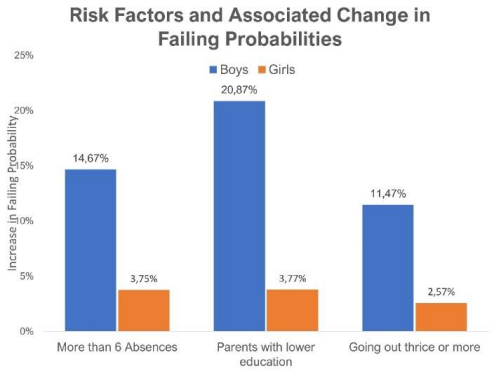
It is important that you completely follow your intuition when predicting the students' grades. Do not overthink too much and do not worry about giving a right or wrong answer. Remember that all of your responses are processed anonymously.

End of Block: Task Description

Start of Block: Statistics Stereotype Activation

We already have some information about the associations between a student's characteristics and their chance to pass the English exam. More specifically, we identified some *risk factors*. Risk factors are characteristics of a student that are associated with a higher chance of failing the exam. One risk factor is, for instance, if a student goes out thrice or more per week. This student has a lower chance of passing the exam, than other students who go out less often.

Interestingly we found that risk factors affect boys more than girls. Take a look at the graph below where some risk factors are outlined:



To explain this graph look for instance at the risk factor "more than 6 absences": girls who fall into this category have a 3,75% higher chance of failing the course than other students. The effect for boys is bigger: if they miss more than 6 classes their chance of failing increases by 14,67%.

Please look at the graph for a bit and make sure you understand it. We are going to ask some questions to make sure you understand the effects of different risk factors.

*

How much more likely is a boy that goes out thrice or more to fail the exam compared to other students of the English course?

- 2,57%
- 14,67%
- 11,47%

How much more likely is a girl that goes out thrice or more to fail the exam compared to other students?

_____

*

Girls whose parents did lower education are 20,87% more likely to fail the class than other students

- True
- False

*

In this graph, what is the second biggest risk factor for boys?

- Having more than 6 absences
- Having parents with lower education
- Going out thrice or more

End of Block: Statistics

Start of Block: Case Based Stereotype Activation

To guide you in your task, we will first show you three profiles of students following the English course. Their names are Tom, Lucas and Emma. As you see, Tom is a low performing student, Lucas a medium and Emma a high performing student. Please take some time to look at the profiles and answer the questions below, before proceeding with the survey.

Tom (5/20)

Parents' education - *Middle School*
Reason School Choice - *Close to home*
Studytime - *less than 2 hours*
Absences - *4*
In a relationship - *no*
Freetime - *low*
Going out - *Thrice a week*
Alcohol consumption -*high*

Lucas (10/20)

Parents' education - *University*
Reason School Choice - *Reputation*
Studytime - *between 2 and 5 hours*
Absences - *1*
In a relationship - *no*
Freetime - *high*
Going out - *Twice a week*
Alcohol consumption -*moderate*

Emma (17/20)

Parents' education - *High School*
Reason School Choice - *Curriculum*
Studytime - *between 2 and 5 hours*
Absences - *2*
In a relationship - *yes*
Freetime - *average*
Going out - *Twice a week*
Alcohol consumption -*moderate*

Which of Tom's characteristics do you think affected his grade the most? (select up to three options)

- [ ] His parents' education
- [ ] His reason for choosing the school
- [ ] His amount of studytime
- [ ] His number of absences
- [ ] His relationship status
- [ ] His amount of freetime
- [ ] The number of times he goes out
- [ ] His Alcohol Consumption

Which of Lucas' characteristics do you think affected his grade the most? (select up to three options)

same options as in previous question are provided

Which of Emma's characteristics do you think affected her grade the most? (select up to three options)

same options as in previous question are provided

End of Block: Case Based Stereotype Activation

Start of Block: Ranking Task

We arrived at the main part of this survey! Once you've completed this part, we will only ask you some more personal questions before you're completely done.

Below we present you with all student profiles. Please take some time to look at them and afterwards rank them according to how well you expect the students to perform in the English exam. Start with the student you think will get the highest grade.
You can use your mouse/touchscreen to drag and drop the student names.

For each student also specify in the blank field which grade you think they'll get. The grades range between 1 and 20 (10 being the minimum passing grade), and the expected passing rate for this exam is 70-80%.
Remember to completely follow your intuition and to not overthink your predictions.

_____ Anna
_____ Jenny
_____ Brian
_____ Oliver
_____ Sarah
_____ Michael
_____ Lisa
_____ David

Here all the student profiles are shown, to give two examples of a student profile

Michael

Absences - *4*
In a relationship - *no*
Going out - *Twice a week*
Alcohol consumption - *high*
Studytime - *less than 2 hours*
Freetime - *high*
Reason School Choice - *Curriculum*
Parents' education - *Middle School*

Anna

Freetime - *average*
Absences - *3*
Studytime - *between 2 and 5 hours*
Alcohol consumption - *moderate*
Parents' education - *Middle School*
Reason School Choice - *Curriculum*
In a relationship - *no*
Going out - *Twice a week*

(note in the other version of this survey the sex of the students will be swapped)

End of Block: Rank Table Format

Start of Block: Demographics

You've nearly made it to the end of the survey. Before you're done, please answer these last questions:

What is your gender?

- ○ Male
- ○ Female
- ○ Other, please specify _____
- ○ Prefer not to say

Please select your age

- ○ 16-24
- ○ 25-34
- ○ …*options are continued*

What is your nationality?

_____

How would you describe your English proficiency?

- ○ Basic
- ○ Intermediate
- ○ Advanced
- ○ Native/Bilingual

Please fill in your background (current or most recent field of work/study)

_____

End of Block: Demographics

D Survey Main Study

Because the survey we used for our Main Study was very close to that of our proof-of-concept study, we do not provide a copy of it (the survey used for the proof-of-concept study can be seen in Appendix C). As mentioned previously, the main change in this study lay in the presentation of the student profiles. Whereas in the proof-of-concept study each participant had to rank and grade the same eight student profiles, different profiles were presented per participant in our main study. Further, because the data from our main study is made publicly available, we added the following remark to our consent form:

> **Re-use of Data** The data that is collected in this study will be made available to public, so that it can be re-used for future scientific studies. This means, that researchers outside the university of Antwerp might access your anonymous responses. No personal information will be shared, however.

E Results Main Study

E.1 Participants

In table we present the number of participants in our main study, distributed over the different "Stereotype Activation" conditions. In Table 18 we show the number of participants by gender and age, while in Figure 13 we show the number of participants by their nationality.

Table 17: Number of Participants Across the Conditions

Stereotype Act.	Count	% of Total
None	32	30.2%
CaseBased	34	32.1%
Statistics	40	37.7%

Table 18: Number of participants by gender and by age

Gender	**Count**	**% of total**
Female	74	69.2%
Male	32	29.9%
Prefer not to say	1	0.9%

Age	**Count**	**% of total**
18 - 24	68	63.3%
25 - 34	33	30.8%
35 - 44	5	4.7%
45 - 54	1	0.9%

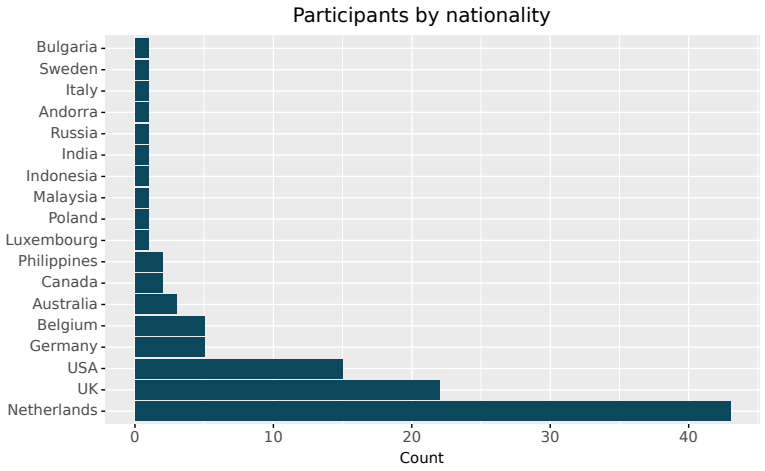


Figure 13: Participants by Nationality

E.2 Understanding the Effect of Stereotype Activation

In our paper we have shown how our collected benchmark data contains biases based on students' sex, that are mostly targeted against boys. However, this is only one of many introduced biases, and another interesting type of bias may stem from the type of stereotype activation that different participants were exposed to before the grading task. In particular, some of stereotype activation conditions may strengthen or weaken existing biases. To estimate these effects we used the binary version of the biased decision labels and compared how these relate to the actual labels per stereotype activation condition. In Figure 14 we visualize the results.

One interesting observation from these plots is that the difference in discrimination rates between boys and girls is lowest if no stereotype activation is presented (difference of 5.31%), and highest for the "Statistics" condition (difference of 17.5%). The difference between these rates in the "CaseBased" condition lies with 11.76% between both. It is difficult to say whether these differences can completely be attributed to the various stereotype activation conditions, since participants in each condition got presented with different profiles and direct comparison for predictions on the same profiles is impossible. Still these numbers follow the patterns observed in the proof-of-concept study, and thus suggest that the different types of stereotype activation have some effect on how participants graded student profiles of different sexes. In regards to favouritism, it is hard to draw strong conclusions from the data. Relatively speaking, the difference in favouritism rates between boys and girls is highest when no stereotype activation is presented. However, given that not much favouritism occurs in the data overall, it is harder to observe clear patterns from this small fraction of the data.

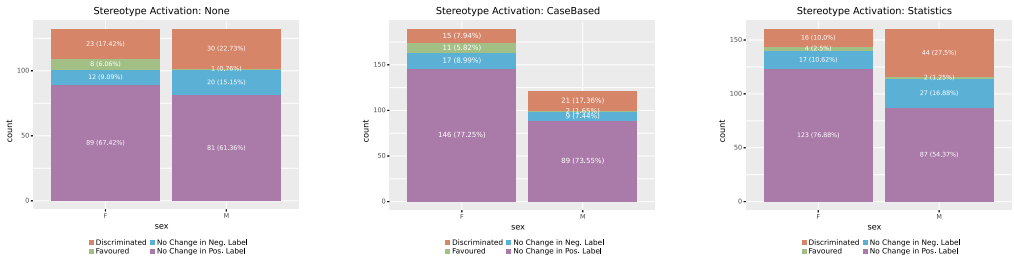


Figure 14: Visualization of how different stereotype activation conditions affected the type/amount of bias introduced by our experiment.