Technical University of Munich

Application Project

Winter semester 2018/19

Data Mining for Hidden Bias

Author:
Patricia Goldberg
Matriculation Number:
03694460

Advisor:
Ana Petrovska
Supervisor:
Prof. Anne Brüggemann-Klein

May 28, 2019

Abstract

The concept of Hidden Bias is widely seen in many different areas and is regarded as one of the main reasons for the gender disparity between students pursuing degrees in Computer Sciences. Since less than 20% of Computer Science students are female, the information about informatics-gender bias is relevant for those female students in this field. As there is a lack of available datasets containing this informatics-gender bias information, the aim of this project was to investigate if and by how much minorities are affected by a likely bias in their academic life.

Since hidden bias is not trivial to measure and analyze, a web-based Implicit Association Test (IAT) was developed during this research, together with other 3 types of self-report questions. The data collected in this research was used to evaluate the automatic association that students have towards a specific gender and the computer science field, leading to an understanding of a different implicit gender association to Computer Science among females and males students, as the male students showed a stronger automatic association of Male and Computer Science compared to female students.

Contents

1 Introduction

Unconscious Bias (also known as Hidden Bias [1]) is a term from the psychology field which describes the bias an individual has, unconsciously, against a person or a situation [2]. Since the access to the unconscious part of the brain is much faster than to the conscious, this ability is therefore useful in dangerous situations, such as an encounter with a lion. In such events, the brain automatically associates "lion" with "dangerous", providing an instantaneous reaction to the situation. However unconscious bias is not always a good thing, it is also related to the labels which burden certain groups of people, based on their gender, color of the skin, age, religion or country of origin. The process of labeling a person is automatic, and people develop these behaviors due to exposure to cultural attitudes against certain groups.

The concept of Hidden Bias is widely seen in many different areas and is regarded as one of the main reasons for the disproportion in the numbers of male versus female students pursuing degrees in Computer and Information Sciences, with around 20% of computer science degrees awarded to women [3]. In previous research, the math–gender stereotype was studied using a mixture of Implicit and Explicit Tests to gather information about the math–gender relationship [4], but the informatics-gender stereotype has not been researched yet.

To lessen the hidden bias consequences, the Informatik Forum Frauen (IFF) conducts Unconscious Bias training within the Technical University of Munich (TUM), primarily focusing on the Department of Informatics. The goal of this training is to promote and increase awareness through real-life experiments and examples. The Unconscious Bias training has been previously conducted in different environments inside the University, including Fit in die Lehre workshop for Ph.D. students and pro-seminars [5].

Inspired by the above mentioned Unconscious Bias training and by the lack of available datasets containing the informatics-gender bias information, the project "Data Mining for Hidden Bias" was designed. This inspiration was triggered by the TUM Department of Informatics gender proportion. Every year, the Department releases a report called "Facts and Figures" [6] which contains the current Department number of female and male students and professors. Having in mind that less than 20% of the department students are female and only 11% of the professors, this project's aim is to investigate if and how much minorities are affected by a likely bias in their academic life.

2 Questionnaire development

The first problem this project faced was in the psychology field, the main issue was which type of questionnaire could possibly analyze a person's hidden bias. In principle, self-report questions are not adequate in this situation, since people tend to lie in this type of reports [1]. The solution for this problem lay in the psychology concept named "Implicit Social Cognition" [7], which was used to develop the so-called Implicit Association Test (IAT). IAT is used as a tool that estimates bias towards a specific stereotype issue and it was first developed in 1994. Since then, it has been researched and improved by several research groups. This includes a non-profit organization called Project Implicit, an on-going project which hosts several IAT types. The project can be found on the Webpage: https://www.projectimplicit.net.

The second problem faced was the lack of available online survey tool which collects the time each respondent takes to complete a section of the survey, which is an essential information used to calculate the IAT results. The solution was to develop an online questionnaire which could gather data from an Implicit Association Test, focusing on the gender stereotype problem [4], together with a set of self-report questions. The self-report questions are particularly useful for cases in which we want to evaluate, not only what a person thinks, but also how one would explicit expose its ideas.

2.1 Types of questions

The first type of question developed was the IAT. It collected the time and the number of errors a person did while associating a target to a category. The target is the focus group and its opposite. Thus this research's targets were Female and Male. On the other hand, the category was Computer Science and Arts, since the goal of the project was to measure the association between Computer Science and Female (the target group). The IAT developed was divided into 7 tasks, from which 3 were practice tasks and 4 were critical tasks. Critical tasks are the only ones used to measure hidden bias.

The second type of question developed was the Explicit Test, also known as self-report question. It had 3 sub-types (Matrix, Drag and Drop and Single Choice) and each sub-type could be used more than once in the questionnaire. The matrix had columns containing categories, such as from Strongly disagree to Strongly agree, and rows with statements to associate with the columns. The Drag and Drop allowed the respondent to sort words according to their preferences by dragging a box of the word and dropping it where preferred. The Single Choice consisted of a question and multiple possible answers, while the respondent could only choose one answer.

2.2 Implementation

This research used several different technologies in its development. React was used in the Front-End, connecting with Python in the Back-End. MongoDB was the NoSql chosen to store the respondents' answers due to its flexibility Docker was used as the container, which was executed in a single Linux server hosted by DigitalOcean. In the MongoDB, besides all the answers for the Explicit Test, the time of each IAT task per person, the error of each IAT task per person and the IAT Result (stored in a variable named winner) were stored.

2.3 Questionnaire Order

The ordering between explicit and implicit tests normally has little influence in the final IAT measurements [8]. But, there have been studies which show that this order can influence a "Gender-Stereotype" type of IAT, which is the type of IAT implemented in this project. Therefore, the questionnaire was designed in a specific order.

The first page of the questionnaire was a description of the project according to the GDPR [9], and only after consenting could the respondent start answering the questionnaire. The first section of questions contained Explicit Tests about gender identity, used to specify the context of the questionnaire to the respondent. The second section was the IAT, collecting the measurements from each of the 7 tasks. The last and third section contained Explicit Test about the Department of Informatics at TUM and some other relevant questions to the questionnaire. After all the questions were answered, the IAT result (the winner), further explained in section 3.2, was shown to the respondent, together with a "thank you" page and IFF contact details.

2.4 Data Collection

Together with the permission of the TUM Data Protection Officer and professors, the questionnaire was filled during lectures and sent by e-mail to online Practical Courses. In total, 457 students completed the questionnaire, being 184 females (41%), 267 males (58%) and 6 of gender rather not say (1%). The questionnaire was run in the following courses:

- Functional Programming and Verification
- Foundations of database system
- Betriebssysteme
- Distributed Systems
- Grundlagen der Programm- und Systementwicklung
- Software Engineering Essentials
- iPraktikum
- Einführung in die Rechnerarchitektur

A special appreciation to everyone involved in the data collection: Prof. Bernd Brügge, Dr. Lukas Alperowitz, Prof. Seidl, Prof. Pretschner, Prof. Anne Brüggemann-Klein, Prof. J. Ott, Maiken, Prof. M. Schulz, Prof. H.-A. Jacobsen, Thomas Hutzelmann, Nasirifard Pezhman, Maximilian Schüle, Nico Hartmann, Nane-Maiken Zarges, Raphaela Palenta, Dora Dzvonyar, and Ana Petrovska.

3 Data Exploration

This section is divided in accordance with the subdivision of the questionnaire. Section 3.1 describes all questions and answers related to gender identity using explicit tests. Section 3.2 measures IAT results. Section 3.3 provides details on the answers to all other explicit tests.

3.1 Gender Identity

In order to have a better insight into whom is completing the online questionnaire, some basic personal questions are essential, such as a Single Choice question about the respondent's gender and one Single Choice about their age. Besides classifying the respondent using personal questions, this section purpose was also to extract their gender identity information. Using this type of information, we can better understand the person replying to the questionnaire, what are their personal views about their own gender, and how well can they identify themselves with their gender. This section then contained other two questions: one Matrix and one Drag and Drop. In the matrix, the respondent could associate a sentence with 4 categories: Strongly Agree, Agree, Disagree, and Strongly Disagree. In total, there were 8 gender identity sentences [10] in the same matrix based on the gender of the respondent (Female or Male). If the respondent was "Rather not say", this question was not shown. Although these sentences were all in the same matrix while the data was collected, as they had different tones, we split them into "Negative" and "Positive" tone to evaluate their answers. Such a list of sentences split into different tones and genders are below, provided using the format "Sentence Number-Gender of the respondent: Sentence":

Positive tone sentences:

1-Female: I feel I fit in with other Females in my department
1-Male: I feel I fit in with other Males in my department
3-Female: I feel comfortable being a Female in my department
3-Male: I feel comfortable being a Male in my department
5-Female: I feel that my personality is similar to most Females personalities in my department
5-Male: I feel that my personality is similar to most Males personalities in my department
6-Female: I feel that the things I like to do in my spare time are similar to what most Females in my department like to do in their spare time
6-Male: I feel that the things I like to do in my spare time are similar to what most Males in my department like to do in their spare time
8-Female: I think I am a good example of being a Female
8-Male: I think I am a good example of being a Male

Negative tone sentences:

2-Female: I feel annoyed that I am supposed to do some things just because I am a Female
2-Male: I feel annoyed that I am supposed to do some things just because I am a Male
4-Female: I feel people interpret my behavior based on my gender
4-Male: I feel people interpret my behavior based on my gender
7-Female: I sometimes think it might be more fun to be of opposite gender
7-Male: I sometimes think it might be more fun to be of opposite gender

Figure 1 represents the percentage of replies for each sentence presented above. The graph is sub-divided into Positive and Negative tone sentences and it is colored by the categories, varying from Strongly Disagree to Strongly

Percentage of responses by category

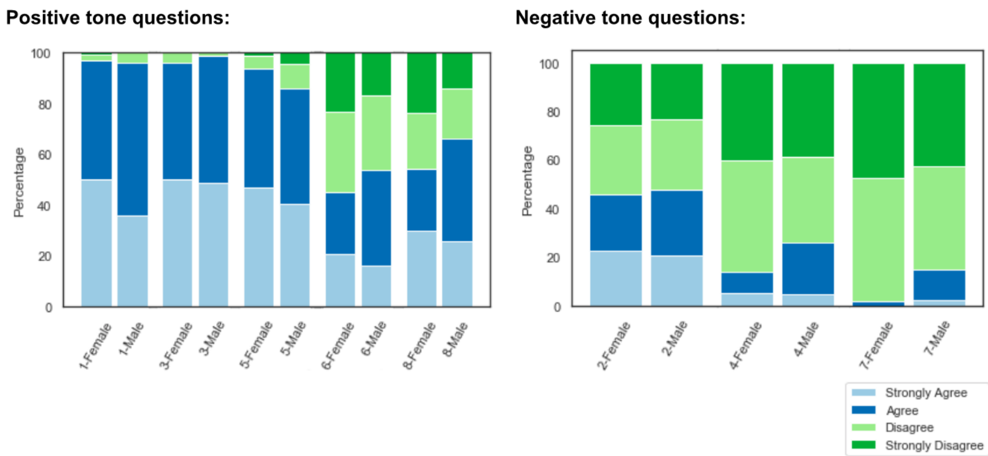

Figure 1: Percentage of replies for each sentence in the Gender Identity matrix section of the questionnaire, sub-divided into Positive and Negative tone sentences and colored by the categories, varying from Strongly Disagree to Strongly Agree.

Agree.

The Negative tone sentences had the most accumulative percentage of "Strongly Disagree" and "Disagree" answers. Sentences 3, 5 and 2 did not show a significant change in the answers among females and males respondents (less than 5% difference in each agreement level). Sentence 1 had a higher number of females strongly agreeing, while males had a higher percentage of respondents agreeing. Sentences 6 and 8 had almost the same amount of females in each level of agreement, while it had a lower number of males strongly agreeing and strongly disagreeing. Sentences 4 and 7 had the least number of respondent agreeing or strongly agreeing, having a lesser percentage of females agreeing than males.

In this question, the data collected did not represent a significant difference in responses among female and male respondents. The most significant difference is between Positive and Negative tone questions, in which a higher number of respondents tended to disagree with Negative tone questions.

In the Drag and Drop question, the respondent was asked, given 10 words [11], to rank them by arranging qualities in the order they thought it was relevant to how to be successful in their career or studies (most important at

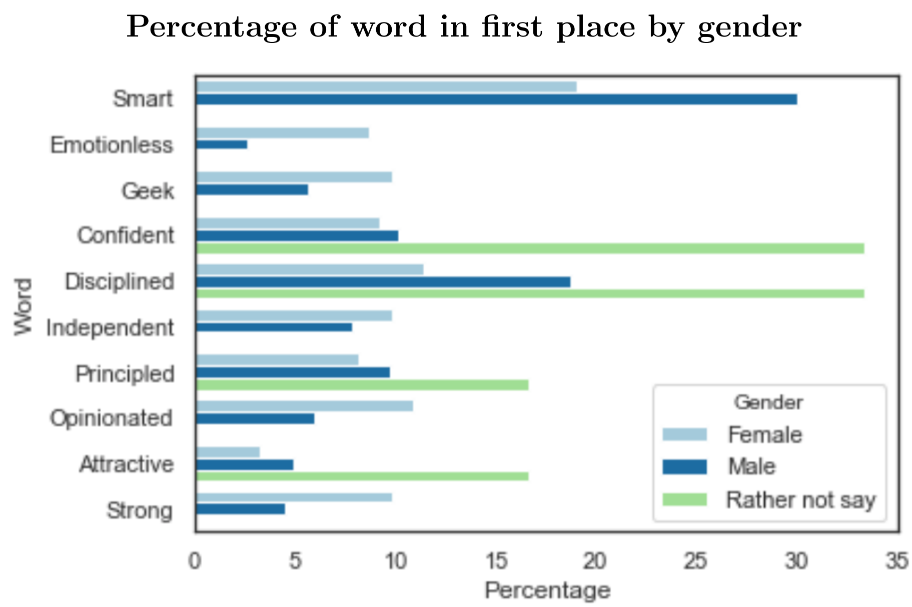

Figure 2: Percentage of the words ranked in first place of the gender identity Drag and Drop question divided by the respondent gender.

the top). Figure 2 shows the percentage of the words ranked in first place by all of the respondents, divided by gender. Since the number of respondents who replied to "Rather not say" as their gender was low, the focus of this section will be only on Females and Males answers. The word with the highest percentage, for both Female and Male, was "Smart", 20% Females and 30% of Males ranked that as first place. The word which appeared in the first place the lowest was "Attractive". It is worth noting that the words "Emotionless", "Geek", "Opinionated" and "Strong" had twice the number of Female respondents ranking it first place than Male respondents.

Although the data collected in the gender identity matrix do not show a significant difference between male and female respondents' answers, the data collected in the Drag and Drop question do. The Drag and Drop data show a significant difference in adjectives which normally are used to describe the opposite gender. In this case, words such as "Emotionless" and "Strong" normally associated as a male characteristic, were better positioned by female respondents as an important quality to have in order to succeed in their career or studies rather than by male respondents. This result shows a weaker gender identity of female students compared to male students, raising a theory that informatics female professionals may have to adapt their behavior in order to succeed in a male-dominated environment.

3.2 Implicit Association Test

Following the "Gender Identity" section is the Implicit Association Test. This is the section in which the hidden bias will be calculated. As the IAT procedures are not trivial, instruction on how to proceed were first shown to the user, together with a table listing the 5 words which represented each category (Arts and Computer Science) and target (Female and Male).

Arts	Sculpture	Music	Theater	Painting	Melody
Computer Science	Programming	Technology	Code	Mathematics	Electronics
Female	Sister	Mother	Aunt	Grandmother	Daughter
Male	Brother	Father	Uncle	Grandfather	Son

Table 1: Table of words which represent the categories and the targets of the Implicit Association Test.

3.2.1 Method

Using the words listed in Table 1, the respondents were asked to associate each word to the corresponding category or target. This was done by 5 steps, subdivided into 7 association tasks. Steps 1, 2 and 4 were practice sessions, thus, the respondent answers in these steps did not count into the final Hidden Bias calculation. Steps 3 and 5 were the critical sessions, therefore the time and the number of mistakes the respondent made in each task were collected.

These association tasks are made dividing the screen into right and left. With the target Male on the left, the respondent should associate all the words listed as Male to the left part of the screen. The unsuccessful association of the word to the correct side of the screen is considered as an error, and a red cross appears in the middle of the screen. The association process is different across platforms. If the user is using a laptop, this is done by representing the left side of the screen with the keyboard letter E, and the right side with the letter I. If the user is using a mobile device, they could just click on the right or left side of the screen in this association task.

Step 1 had only 1 association task, which was to associate the target Male to the left of the screen and the target Female to the right. Step 2 had the category association task, associating Computer Science and Arts to the left and right respectively. Step 3 had two consecutive critical association tasks, which the user had to associate words related to Male or Computer Science to the left side of the screen, and words related to Female or Arts to the right. Step 4 had one practice association task, the aim of this task was to change the side of the screen of the targets, thus Female was located in the left side, while Male in the right. The last step, such as step 3, had two critical association task. The target Female was in step 5 on the left side though, while the categories remained on the same side as step 3.

The results are calculated by counting the errors and response time of step 5 minus step 3. If the result is negative, the winner is female, meaning that the respondent has an automatic association between Female and Computer Science. If the result is positive, the winner is male. The automatic association term refers to the hidden bias a person has of the target and the category. This association is subdivided into 4 classifications groups: "little to no", "slight", "moderate" and "strong". The group "little to no" means that the respondent demonstrates hardly any automatic association to that gender and Computer Science, while "strong" means a strong automatic association. If the absolute value of the result is between 0 and 6, the association is "little to no". If this value is between 6 and 12, the association is "slight". If higher than 12 and lower than 18, "moderate". Otherwise, the automatic association is considered "strong" [1].

3.2.2 Limitations

It is important to understand that this IAT platform had two main limitations. The first would be that one participant can make the test multiple times, endangering the data independence assumption. This is minimized by a large number of samples. The second was the order of Step 3 and Step 5 which, in this IAT, was fixed, so the "Compatible Category" (meaning: Men and Computer Science) is shown first. Therefore steps 3 and 5 had two critical association tasks each since this reduces the typical effect of the pairing order [12].

3.2.3 Results

Figure 3 represents the distribution of the IAT results' variables. The x-axis represents the classification categories, while the y-axis the number of

Number of responses per classification and winner

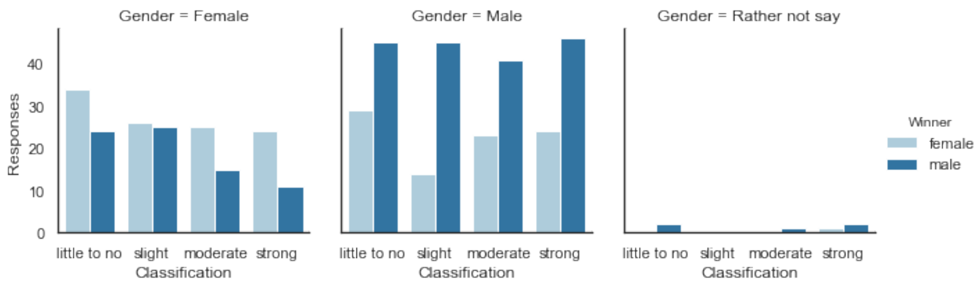

Figure 3: Bar plot representing the total number of responses per classification category of the Implicit Association Test, sub-divided by the gender of the respondent and the Implicit Association Test winner of that respondent.

responses in that category. Each bar plot represents the gender of the respondent, divided by "Female" (41% of the respondents), "Male" (58%) and "Rather not say" (1%). And the variable "Winner" represents the winner of the Implicit Association Test, explained in Section 3.2.1. It worth noting that 60% of female respondents showed an automatic association with Female and Computer Science, and this association is spread into the classification categories, varying from "little to no" until "strong". Meanwhile, most of the male respondents (67%) showed a stronger automatic association with Male and Computer Science, having roughly the same intensity of classification categories.

The data result, shown in figure 3, show that the majority of male respondents demonstrated a strong automatic association between Male and Computer Science. This result leads to the conclusion that men in the department of informatics at TUM have a stronger implicit association between Male and Computer Science. Meanwhile, women have a less strong association between Male and Computer Science, tending to demonstrate actually a strong association between Female and Computer Science.

3.3 Department Related Questions

The last section of the questionnaire contained explicit questions about the Department of Informatics at TUM in a matrix form. This section was mainly designed to comprehend the student's perception of the department gender equality issues. Here, 8 sentences were presented in rows, regardless of the gender of the respondent. The categories of responses (presented in the columns of the matrix) were “Strongly Agree”, “Agree”, “Strongly Disagree”, “Disagree” and “Not Applicable”. The category “Not Applicable” was added since some students may have neither an advisor nor a supervisor, thus some sentences were not applicable to those students. The list of sentences can be found below:

1: Students in my department are treated equally by the staff regardless of their gender
2: I am confident that the staff of my department would address sexism
3: I have witnessed gender discrimination from the staff of my department
4: I feel that my supervisor/advisor treats men and women equally at the lab
5: I feel that my professor treats men and women equally during the lectures
6: I feel that my supervisor/advisor aligns her/his research with mine
7: I feel that my supervisor/advisor helps me identify my training/development needs
8: I feel that my supervisor/advisor shows interest in my progress/success

Figure 4 represents the percentage of replies for each sentence related to the Department of Informatics, colored by the category of the answer and sub-divided by the gender of the respondent. Sentences 3, 6, 7 and 8 had almost 50% of the respondents answering as “Not Applicable”, this was indeed expected for the phrases 6, 7 and 8 since there were students that do not have supervisor or advisor. Sentence 3 cannot be explained in the same way, as it was a question about gender discrimination. Furthermore, this sentence had the highest percentage of “Disagree” answers, leading, initially, to no conclusion. In the matrix as a whole, hardly any difference can be spotted while comparing the gender of the respondent to their answer.

4 Data Analysis

Data analysis of this project was focused on statistical evaluations of the IAT measurements. This was mainly done using boxplot [13], which uses as a basis the data quantiles in order to be constructed. Following the IAT best practises [14], the measurements' outliers were deleted using the z-score given by $z = \frac{x-\mu}{\sigma}$, in which μ represents the mean and σ the variance [15].

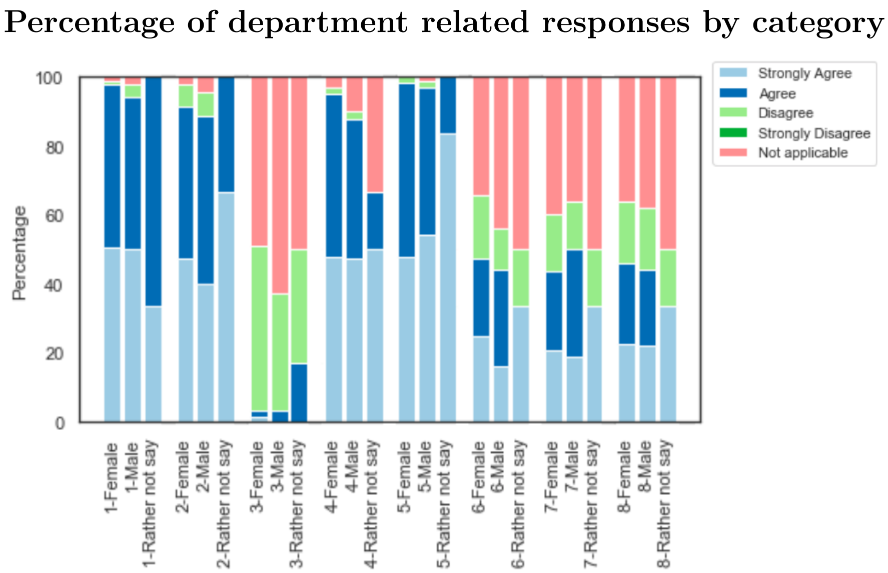

Figure 4: Percentage of department related replies for each sentence, subdivided by gender and colored by the categories, which vary from Strongly Disagree until Strongly Agree.

Figure 5 is a boxplot which represents the duration of the respondents (in seconds), divided by gender, per task. Having in mind that step 3 associates Male and Computer Science in two association tasks, and step 5 associates Female and Computer Science in two tasks also, the boxplot shows a significant wider interquartile range (IQR) of the male respondents in step 5 comparing to female respondents. Therefore, the time range of male respondents when associating Female and Computer Science was bigger than the time range of female respondents. The median of the male respondents in step 5 is significantly higher, comparing to their median in step 3, explaining the IAT results found in Section 3.2.3.

Figure 6 represents the mistakes per step, per respondent gender. While the boxplot for Females remains almost constant across tasks, the boxplot for the Males shows a higher IQR for step 5, explaining the reasoning behind the IAT results obtained in section 3.2.

There are significant differences in the median and in the interquartile range of the duration and the number of mistakes when grouping the respondents per gender. This result leads to the conclusion that men take

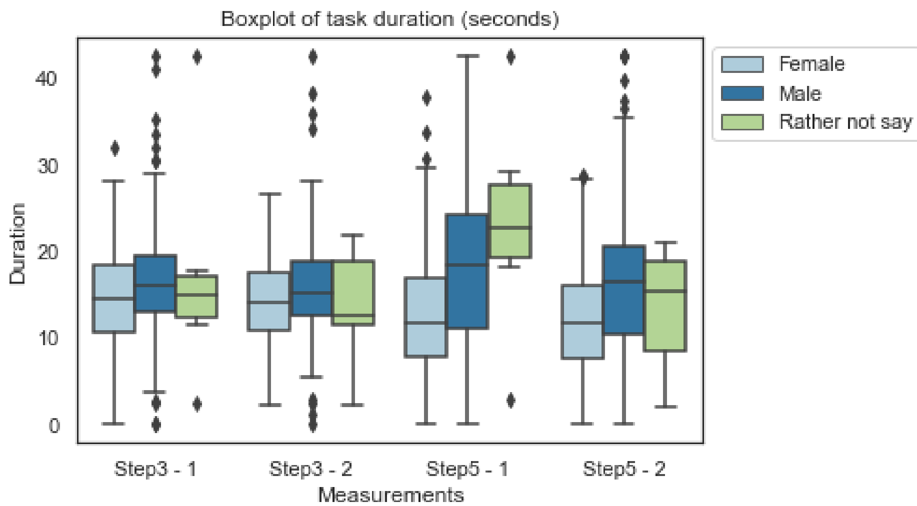

Figure 5: Boxplot displaying the duration of the respondents (in seconds), divided by gender, per task.

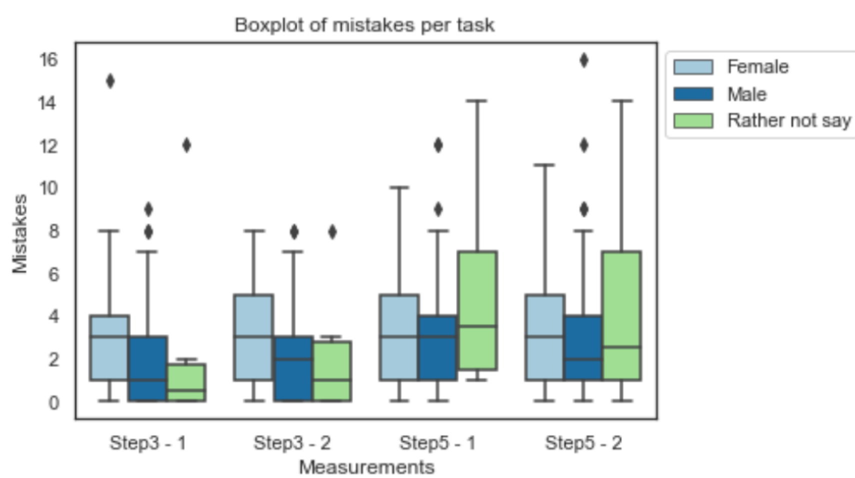

Figure 6: Boxplot displaying the mistakes of the respondents, divided by gender, per task.

longer and make more mistakes than women while associating Female and Computer Science.

5 Conclusion

Since hidden bias is not trivial to calculate and analyze, one existing study called Implicit Association Test (IAT) was used as the base to our web-based questionnaire. The data collected in this research was used to evaluate the automatic association that students have towards a specific gender and the computer science field. Moreover, explicit questions helped in understanding the students' ideas towards their gender identity and their current perception of how the department deals with gender equality issues.

In the gender identity section of the questionnaire, the ordering of the words in which the respondents felt was more important in order to succeed in their careers or studies, differs when grouping females and male students. Females tend to think it is more important to succeed in the technology area having well-known male characteristics, such as “Geek” or “Strong”, showing a probable weaker gender identity of female students compared to male students.

For the Implicit Association Test, the majority of men had, as a result, an automatic association of Computer Science to Men. Meanwhile, there were a majority of results from women associating Computer Science to Women in all four levels of association. Furthermore, there are significant differences in the median and the interquartile range of answers for the IAT's duration and the number of mistakes when grouping the respondents per gender. Both conclusions lead to an understanding of a different implicit gender association to Computer Science among females and males students, as the male students showed a stronger automatic association of Male and Computer Science compared to female students.

In the Department Related questions, around 50% of all respondents clicked in “Not Applicable” for the gender discrimination question. Which leads to the conclusion that this question was either not well formulated or did not contain enough information to be answered. In all of the other questions in this section, no great difference among female and male respondents was spotted, therefore the perception of the Informatics Department at TUM is similar to all students.

6 Future Work

As for future works, one recommendation is to group the department questions into two groups: Gender equality and Supervision. While Gender equality section could include self-perception sentences of the equality topic in the University, the Supervision section could include specifics sentences about supervisors and advisors at the Department.

As the gender discrimination question had a very high number of respondents who did not reply to it as expected, one idea is to address the sexism in a different question, or in a different project if necessary.

Group results from this work into the "Unconscious Bias Training" is a must have future work, since the idea of this research was indeed to broader the Training with TUM data. Another possible contribution to this research would be to use this data to train machine learning model, this would give a better insight for future "Unconscious Bias Training".

The formulas to calculate the IAT results in section 3.2.3 were based on the formulas described in the book "Blindspot" [1]. Since this book was written in 2013, more accurate formulas are being developed by researchers. Therefore, more recent formulas can be used in future works.

References

[1] Mahzarin R. Banaji. *Blindspot : hidden biases of good people*. New York : Delacorte Press, [2013] ©2013, [2013].

[2] Claude M Steele. *Whistling Vivaldi: And other clues to how stereotypes affect us (issues of our time)*. WW Norton & Company, 2011.

[3] National Center for Science and Engineering Statistics. Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019. Technical report, National Science Foundation, 2019.

[4] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. Math= male, me= female, therefore math≠ me. *Journal of personality and social psychology*, 83(1):44, 2002.

[5] IFF. Unconscious Bias Training, 2018.

[6] TUM Department of Informatics. THE DEPARTMENT OF INFORMATICS IN FACTS AND FIGURES 2018, 2018.

[7] Anthony G Greenwald and Mahzarin R Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4, 1995.

[8] Brian A Nosek, Anthony G Greenwald, and Mahzarin R Banaji. Understanding and using the implicit association test: Ii. method variables and construct validity. *Personality and Social Psychology Bulletin*, 31(2):166–180, 2005.

[9] Council of the European Union European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[10] Susan K Egan and David G Perry. Gender identity: a multidimensional analysis with implications for psychosocial adjustment. *Developmental psychology*, 37(4):451, 2001.

[11] Michelle Levi Brittany Welby. The Shriver Report Snapshot: An Insight Into the 21st Century Man. Technical report, A Woman's Nation, 2015.

[12] Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197, 2003.

[13] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[14] Anthony G Greenwald and Brian A Nosek. Health of the implicit association test at age 3. 2001.

[15] Dhiren Ghosh and Andrew Vogt. Outliers: An evaluation of methodologies. In *Joint statistical meetings*, pages 3455–3460. American Statistical Association San Diego, CA, 2012.