<table>
<thead>
<tr>
<th>Title</th>
<th>Gender differences in personality traits of software engineers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Russo, Daniel; Stol, Klaas-Jan</td>
</tr>
<tr>
<td>Publication date</td>
<td>2020-06-18</td>
</tr>
<tr>
<td>Type of publication</td>
<td>Article (peer-reviewed)</td>
</tr>
</tbody>
</table>
| Link to publisher's version | https://ieeexplore.ieee.org/document/9120355  
http://dx.doi.org/10.1109/TSE.2020.3003413  
Access to the full text of the published version may require a subscription. |
| Rights | © 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. |
| Item downloaded from | http://hdl.handle.net/10468/10183 |
Gender Differences in Personality Traits of Software Engineers

Daniel Russo, Member, IEEE, and Klaas-Jan Stol

Abstract—There is a growing body of gender studies in software engineering to understand diversity and inclusion issues, as diversity is recognized to be a key issue to healthy teams and communities. A second factor often linked to team performance is personality, which has received far more attention. Very few studies, however, have focused on the intersection of these two fields. Hence, we set out to study gender differences in personality traits of software engineers. Through a survey study we collected personality data, using the HEXACO model, of 483 software engineers. The data were analyzed using a Bayesian independent sample t-test and network analysis. The results suggest that women score significantly higher in Openness to Experience, Honesty-Humility, and Emotionality than men. Further, men show higher psychopathic traits than women. Based on these findings, we develop a number of propositions that can guide future research.

Index Terms—Personality traits, Gender, Empirical software engineering, Bayesian statistics, Network analysis.

1 INTRODUCTION

Several scholars have proposed to close the gender gap that exists in many fields through proactive policies [1], [2], and also in software engineering [3], [4], [5]. Studying how gender differences influence the relations within software organizations has essential theoretical and practical implications. For example, unfair performance evaluations can lead to trust erosion [6], may reduce job satisfaction [7], and increase absenteeism and staff turnover [8]. In short, it is clear from these studies that gender plays a critical role in effective software development, in both corporate development teams and in open source communities.

Most of the gender research in software engineering has considered only two genders: women and men. Women who work in domains dominated by men face more gender bias [9]; software engineering is such a domain. Recent studies suggest that teams with more gender diversity have fewer “community smells” [10], and more gender diversity is linked with a shorter issue fixing time, and increased politeness [11]. Similarly, gender diversity is associated with a higher level of productivity [12]. In open source communities, men engage for more extended periods than women who disengage more quickly [13]. One reason for this is that women ask more questions than men on average, which often remain unanswered, leading to an ‘unhealthy’ community [13]. Gender differences play a significant role in professional environments; for example, men often receive more favorable performance evaluations than women [14], [15], and therefore have a higher probability of getting promoted [16]. Women face specific contribution barriers in online communities [17], which are disproportionate compared to men [18]. Moreover, a 2017 survey of 5,500 GitHub users [19] shows that women encounter more often than men language or content that makes them feel unwelcome, stereotyped, and face unsolicited sexual advances.

Given that software development is inherently teamwork, a second important factor to study is developer personality. There is a considerable body of studies on personality spanning over half a century [20], linking personality to a variety of aspects of teams, such as effective team structures [21]. Personality has also been linked to software engineers’ attitudes [22], which affects how engineers collaborate. Given the important role of both gender and personality, it is surprising that there are very few studies that focus on the intersection of these two topics. Hence, we set out to investigate gender differences in personality traits of professional software engineers. Our aim is to establish generalizable results, advancing stable and long-lasting theoretical contributions through a representative sample study. Furthermore, stable and long-lasting contributions to understanding the studied phenomenon are among the key characteristics of personality trait research [23]. This is not only true for the general population, as Cobb and Schurer pointed out, but also for software engineering professionals [24], and therefore used in software engineering research [25].

In this article, we seek to explore gender differences among software engineers in terms of their personality traits. Because our study relies on statistical techniques, it needs a large sample size of each gender considered; hence, our investigation is able to consider only the genders with large numbers: men and women. Thus, our research question is:

Research Question: How do personality traits differ in men and women software engineers?

To address this question, we conducted a large-scale sample study. Section 3 presents the design of our study. Following a rigorous participant selection process, we selected a representative sample of the software engineering population; 483 valid responses were included in the analysis. We first used...
Bayesian independent t-tests to analyze the degree to which men and women software engineers differ in terms of their personality traits. We then performed a network analysis to study the relations among these personality traits. Network analysis offers a complementary perspective on personality traits [26] as it focuses on the relationships between traits. Section 4 presents the results of these two types of analyses. We discuss the implications of these results in Sec. 5, based on social psychology literature evidence, suggesting specific work-related performances of each specific trait, such as team and individual job performance, generating a set of propositions. Section 6 concludes this article.

2 RELATED WORK

2.1 Personality Traits

Personality is a set of patterns of thinking, feeling, and behaving based on a set of traits that are predictors of an individual's behavior and action [27]. The first personality trait-based theories were proposed in the 1930s, with pioneering work from Karl Gustav Jung [28]. However, it was not until the mid-1970s that traits were operationalized employing measurement instruments [29]. Since then, personality-based theories have flourished, and several models have been proposed, such as the Minnesota Multiphasic Personality Inventory [30]. The software engineering research community has adopted several of those [20]. Two of the most commonly used personality theories in software engineering are the Myers-Briggs Type Indicator (MBTI) derived from Jung’s Personality Type Theory [31] and the Five Factor model based on the Big Five theory [32].

2.1.1 Meyers-Briggs

The Meyers-Briggs type indicators consider an individual's personality composed of four dichotomous dimensions [33, p. 6]:

1) Extraversion (E) vs. Introversion (I), indicating whether an individual directs her attention to the external world of people or the inner world of experiences and ideas.
2) Sensing (S) vs. Intuition (N), indicating whether an individual perceives the world by observing reality or through imagination.
3) Thinking (T) vs. Feeling (F), indicating whether an individual makes decisions based on logical reasoning or beliefs.
4) Judging (J) vs. Perception (P), indicating whether an individual considers the social world as planned or unexpected and spontaneous.

Together these four dimensions define 16 different possible combinations, each representing a distinct personality type, labeled using a four-character string specified above. One study suggested that men software developers with an ENFJ personality (indicating the combination Extraversion, Intuitive, Feeling, and Judging) are more efficient when they work with team leaders exhibiting an ENTJ personality type (Extraversion, Intuitive, Thinking, and Judging) [34].

While the Myers-Briggs classification is widely used, it has been criticized for its lack of validity and utility, such as unstable test-retest reliability and inaccurate predictive validity [35]. Also, MBTI does not correlate with other personality scales [36] and is not consistent with research evidence [37]. Furthermore, scholars have observed that MBTI does not measure qualitatively distinct types, generating quasi-random traits assignments [38]. McCrae and Costa concluded that Jung’s theory is either incorrect or not adequately operationalized in the MBTI measurement instrument; thus, it can not provide a sound basis for interpreting personality. MBTI relies on dichotomous preference scores rather than continuous scales; it limits to grasp the degree of every single dimension since the aim is to classify subjects within one personality type [39]. Especially borderline cases are assigned to one or the other dimension, which can easily lead to misclassification of a person’s personality type.

2.1.2 Five Factor (OCEAN) Model

To address the issues with the Meyers-Briggs test, McCrae and John developed the Five Factor Model [40]. While not based on any prior theory, the Five Factor Model has been developed through a series of empirical studies [38], [40]. The Five Factor model defines five factors: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).1 Combining these starting letters, this model is also known as the OCEAN model. McCrae and John suggested that these five are universal characteristics, or traits, which vary among every person; as such, any person can be categorized with this model.

Personality traits can be characterized as either ‘bright’ or ‘dark’ [41]. In this study, we focused on both types of personality traits of software engineers; we describe these below. Within the OCEAN model, which focuses on bright traits, high scores (except for Emotional stability) indicate bright aspects, while lower scores suggest dark ones. There are also specific models such as the “Dark Triad” [42]: narcissism, psychopathy, and machiavellianism, which measure dark traits, where low scores suggest bright personality traits. The combination of both bright and dark personality measurement instruments provides a comprehensive understanding of the personality of people [41]. To the best of our knowledge, this is the first study in the software engineering literature to use both [20].

2.2 Gender Studies in Personality Research

2.2.1 Gender in Software Engineering

There is a growing body of gender-related studies in software engineering on a variety of topics. Gender is often included in studies as a variable of interest but not always explicitly recorded in platforms such as Stack Overflow. Bin and Serebrnik evaluated a number of “gender guessing” approaches to generate gender information [43]. Other studies have addressed topics such as tenure [12], online participation and related barriers [13], [17], [44], [45], [46], gender bias in pull-request acceptance [47], [48], bug fixing [11], team composition [10], and tools [49]. Burnett et al.’s Gender Inclusiveness Magnifier (GenderMag) method seeks to help software developers to create gender-inclusive designs [50]. A common technique to create products targeting specific

1. Neuroticism should be defined more correctly as Emotional stability, since it does not refer to any mental health disorder.
types of users is the development of personas, though this may also lead to simplistic stereotypes. Hill et al. explored how to overcome this tension between personas and stereotypes [51]. While these studies address a variety of aspects of gender, in this study we specifically link gender differences to personality traits of software engineers, and so in the remainder of this section we focus on prior work that addresses the intersection of these two areas.

2.2.2 Gender in Software Engineering Personality Studies

There has been considerable attention for personality research in software engineering, with a dramatic increase in the last 15 years or so. Cruz et al. identified about 90 papers published between 1970 and 2010, with over 70% published since 2002 [20]. As we argued earlier, very little attention has been devoted to understanding the role of gender differences among developer personality, despite a rise in attention in recent years for gender issues in software engineering [12], [52]. There are a number of notable exceptions. Gilal et al. have conducted a series of studies with student teams who completed small projects using agile methods [34], [53], [54], [55], [56], [57]. Several of these studies focused on team performance, linking it to MBTI personality types of team members. One of their studies found that women were uncomfortable in men-dominated teams, especially when the men were extrovert [53]. Women-led teams, on the other hand, were more welcoming towards other women. Another study by Gilal et al. [55] suggests that the MBTI Feeling trait in men suggests a good fit with the Team Leader role. The MBTI Thinking trait was linked to women’s leadership performance. Likewise, extrovert men were found to be more effective, whereas women were found to be more effective when they scored higher on introversion [57]. The series of studies by Gilal et al. all suggest that personality traits may impact a software team’s effectiveness, and considering these traits could be considered when assigning roles within teams [34], [56]. While these studies shed some light on the importance of personality traits in software development teams, they were conducted with student teams, and given the demonstrated shortcomings of MBTI [35], we suggest these findings should be interpreted with care.

Razavian and Lago found that women focused more on relationships, people, flexibility and intuition than men did [52], suggesting that gender-aware team-building practices might improve an architect team to gain from women’s expertise. For example, women might be more sensitive to customers’ actual needs, suggesting their requirement analysis skills might be more thorough. Grams et al. [58] found that levels of extraversion and neuroticism were higher for women while domain-specific self-efficacy in programming and modeling were much lower for women than for men.

The psychology literature informs us that certain personality traits have a direct influence on work and team performance [59], [60], [61], [62], [63]. Therefore, developing an understanding of personality traits, and gender-based differences among those, can greatly help software teams and managers to build diverse and effective teams. Studying the relations between these traits and gender differences through a network analysis adds further detail to these insights.

3 Research Design

To investigate personality trait differences across gender, we conducted a sample study, considering the individual as a unit of analysis. In designing and reporting this study, we adopted Van Doorn et al.’s guidelines for conducting and reporting Bayesian analyses [64]. This section discusses the measurement instruments (Sec. 3.1), data collection procedures (Sec. 3.2), sample description (Sec. 3.3), and data analysis procedures (Sec. 3.4).

3.1 Research Instruments

3.1.1 Bright Personality Traits

To measure the bright traits, we used an updated version of the Five Factor Model: the HEXACO model of personality [65], which has become the new standard in social psychology [66]. The main difference with the OCEAN model is that it adds a sixth trait: Honesty-Humility. Since this trait was previously included in the Agreeableness trait, the latter is defined slightly differently in HEXACO than in the OCEAN model. This new trait emerged from large-scale studies that identified it as a distinct personality trait [67], and was confirmed through replication studies of the Five Factor Model using more advanced computation techniques that were not available when the OCEAN model was initially proposed [68].

Each of the six personality traits defined by the HEXACO model has four facets:

1) Honesty-Humility (H): Sincerity, Fairness, Greed Avoidance, Modesty.
2) Emotionality (E): Fearfulness, Anxiety, Dependence, Sentimentality.
4) Agreeableness (A): Forgivingness, Gentleness, Flexibility, Patience.
5) Conscientiousness (C): Organization, Diligence, Perfectionism, Prudence.
6) Openness to Experience (O): Aesthetic Appreciation, Inquisitiveness, Creativity, Unconventionality.

Different measurement instruments have been developed for the HEXACO model. Common instruments are HEXACO-PI-R, which defines 100 questions [69], and the HEXACO-60, comprising 60 items [70]. A drawback of having such extensive measurement instruments is that it takes considerable time to complete them, which may negatively affect the response rate in a survey. Hence, we adopted a more recently developed instrument, the Brief HEXACO Inventory (BHI), which defines 24 questions, namely one for each of the four facets of each of the six personality traits. As it is much shorter, this questionnaire can be completed more quickly. Studies have suggested that the results of this instrument exhibit a high level of stability, accuracy, and correlation with the much longer HEXACO-PI-R instrument [71]. Table 1 summarizes the meaning of these personality traits.

3.1.2 Dark Personality Traits

Social psychology scholars consider narcissism, psychopathy, and machiavellianism as to the three dark traits of personality, or the Dark Triad due to its malevolent qualities [42]. Table
We collected data through a sample survey, which is a widely used in computer science [77], as well as in other disciplines such as economics [78], psychology [79], and food science [80]. The Prolific platform facilitates an elaborate screening and selection process, discussed below. We used Qualtrics to administer the questionnaire and shared it on the Prolific platform. While collecting data, we were not only concerned about the number of responses but also about collecting high-quality data. We mainly focused on the sample representativeness, sample size, and ethics [81]. We discuss these concerns next.

### 3.2.1 Representativeness

In order to achieve a representative sample, we collected data using a cluster sampling strategy [82] through the data collection platform. The use of a specialized platform to collect data offers several benefits, including reliability, replicability, and data quality [83], particularly when compared to a population sample, such as the pool of university computer science students [84]. Data reliability, replicability, and high-quality data are pivotal for any study. The Prolific platform supports a systematic selection process to collect high-quality data. We implemented several sample selection strategies; the overall process is represented in Figure 1, and comprised the following steps.

**Pre-screening.** We pre-screened the members of the data collection platform according to the following criteria. Members were required to have knowledge of software development techniques, do computer programming for a living, use technology at work, and have an approval rate of 100%. The last criterion refers to the level of reliability of Prolific platform members in Prolific past surveys. From 75,296 Prolific members that had been active during the last three months, we included 2,897 members.

**Competence Screening.** After pre-screening, we conducted competence screening. We run a randomized screening study, advising that selected members would have participated in another study. From the 2,897 potential subjects, the screening survey was randomly sent to a subset of this population, until we reached around 1,000 participants willing to be part of our study. Only those members who self-identified as a software professional were invited to do the study. This screening step comprised a questionnaire with three...
competency-based questions: one about software design and two about programming. The purpose of this step was to include only those professionals who displayed an adequate level of knowledge of software development. Seven hundred sixty subjects agreed to participate in this study, while 276 other participants who had an initial interest in the study withdrew their participation in this screening phase. We excluded those informants who did not correctly answer two out of three questions (n=154), resulting in 606 potential candidates. We also excluded responses that took more than three minutes to be completed since we considered this suspicious behavior (n=92). At this point, 514 candidates were included based on this criterion.

Quality Screening. All screened informants were invited to take the full questionnaire. To improve data quality, three attention checks were randomly allocated in the survey. We received 491 complete responses. If participants failed to recognize attention checks, we assumed that they did not read the questions sufficiently carefully and discarded them.

At this point, we excluded 8 participants. After the selection process was completed, we included 483 valid and complete responses.

Questions were randomized within their blocks to minimize response (or survey) bias, which refers to respondents’ tendency to respond to a questionnaire inaccurately or dishonestly, for example, by over-reporting good behavior and give responses that are socially desirable [82], [85].

### 3.2.2 Sample Size
Achieving a sufficiently large sample of responses is critical to ensure the generalizability of findings. Yamane suggests to define sample size $n$ as follows [86, p. 549]:

$$ n \geq \frac{N}{1 + Ne^2} $$

(1)

where $N$ is the population size, and $e$ is the level of precision, also known as sampling error, typically set to 1% or 5%. Determining the size of the population of software professionals is extremely challenging. While there is no consensus on the number of software developers worldwide, we identified several estimates (see Table 3). A study in 2019 by SlashData suggested there are 18.9 million software developers, and of those, 12.9 million professionals. An International Data Corporation (IDC) report suggested a global number of software developers of 18 million. An study by Evans Data Corporation indicated a population of 23 million in 2018. Wagner et al. suggested considering the population of GitHub users, which is reportedly over 50 million in 2020. However, this number is somewhat problematic in that it does not take into account duplicate accounts, it does not consider inactive users, does not distinguish professionals from non-professional software developers, and does not consider the fact that much activity on GitHub does not represent software development.

Based on these estimates, it is not unreasonable to assume the actual population size lies somewhere between these

![Fig. 1. Participant selection process](attachment:image)
lower and upper boundaries. Using the highest number as a conservative estimate, which at the time of this study was 36.5m, and a precision of 0.05, Equation 1 suggests a minimum sample based on the conservative population estimate of 36.5 million.

We monitored the evidence as the data accumulated in the form of Bayes factors and posterior distribution through a Sequential Analysis (when data are evaluated as they are collected) [89]. Thus, we stopped our data collection when the tendency was clear enough. As an example, we can see from Figure 2 that the initial gender difference of the first 200 subjects was quite negligible with the Bayesian factor close to zero (or even negative). After collecting additional data, the evidence for $H_1$ became very strong. In other words, the findings based on a sample of n=200 are very different from those based on a sample of n=400. The appendix includes plots for all sequential analyses.

### 3.2.3 Research Ethics
We followed the guidelines of the Declaration of Helsinki [90], which states principles such as appropriate consideration of risks and benefits of a research study, identifying potential benefits to the studied population, and that the study should be carried out by trained scholars. Both authors have completed formal training in research ethics for engineering and behavioral sciences.

The interaction with respondents happened only through the Prolific platform. Prolific’s membership policy stipulates that members may not disclose their identity. As we collected data, the first author was available to answer any questions and provide any clarifications as the survey was running. In total, we were contacted 20 times by participants asking questions related to the nature of the survey, clarifications regarding the answers provided, and motivations for rejections. We did not collect any data of a sensitive nature, or which could be traced back to respondents. Respondents could withdraw from participating at any time up to the point of submitting. Several potential respondents (n=306) withdrew during the study. Also, none of the questions were mandatory, which might result in missing data. However, all responses were complete, and so there were no missing data.

Another worthy consideration is the nature of the survey administration. Publishing a questionnaire to mailing lists and professional fora can be a nuisance for recipients, considering the volume of requests that they are receiving. This was avoided by directly addressing a motivated population of respondents available on the Prolific platform.

### 3.3 Sample Description
Using the process described in Sec. 3.2, we collected 483 responses. This sample size is the largest of a personality research study of software professionals; for comparison, the largest sample identified in the review by Cruz et al. contained 128 data points [20].

In this study we consider gender identity, i.e., irrespective of the biological sex of our informant [91]. Because this study relied on statistical techniques requiring large samples, and some studies suggest that other genders are not frequently reported (cf. less than one percent [92]), we could include only two options. Thus, we specifically asked respondents: “Please report your gender” with options ‘man’ and ‘woman.’

Figure 4 describes the gender distribution of our sample; 18.6% of our sample are women and the remaining 81.4% were men. This is considerably higher than other large-scale industry surveys and previous studies, which reported a much lower proportion of women’s presence. It is similar to an earlier study of personality types in software engineering that reported 20% of respondents were women [93].

In a 2010 study, 10% of the respondents were women [46]. The same number was reported in a 2014 study of participants on Stack Overflow and Drupal [13]. A 2015 study of GitHub contributors found that only 6% were women [12]. In a 2016 study of Stack Overflow, women represented 9% of the respondents [17]. The 2019 Stack Overflow Annual Developer Survey included nearly 90,000 developers and reported that 7.5% of the professional population are women. A 2019 study of OpenStack found that participation from women in a variety of activities ranged from 10 to 12% [94].

Most informants (approximately 93%) were born in Western countries. Table 5 lists the Top 15 countries of origin, which accounts for over 90% of respondents.

Table 6 lists the highest degree obtained by respondents. The table shows that more than 70% of the sample has a university degree.

Table 7 lists respondents’ present main roles as a software engineer; more than half of respondents are primarily working in development. However, tasks of software engineers are rarely well defined and might also change with seniority. For example, team leads or C-suite executives can be experienced

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>393</td>
<td>81.4</td>
</tr>
<tr>
<td>Women</td>
<td>90</td>
<td>18.6</td>
</tr>
<tr>
<td>Total</td>
<td>483</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 4
Gender distribution of the sample

<table>
<thead>
<tr>
<th>Source</th>
<th>Year</th>
<th>Estimate</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDC</td>
<td>2018</td>
<td>18m</td>
<td>Estimate 11.65m full-time developers; 6.35m part-time, and 4.3m non-professional developers.</td>
</tr>
<tr>
<td>Evans Data Corp.</td>
<td>2018</td>
<td>23m</td>
<td>Based on secondary research, found 23m developers in 2018 with expected 27.7m by 2023.</td>
</tr>
<tr>
<td>SlashData</td>
<td>2019</td>
<td>18.9m</td>
<td>12.9m professional developers, 6m non-professional. Relies on 5 sources; threshold includes requirement to have involvement in “substantial coding project” which is ambiguous.</td>
</tr>
<tr>
<td>GitHub</td>
<td>2020</td>
<td>50m+</td>
<td>Number of users, but not all users are developers; many users use GitHub for other storing other types of data.</td>
</tr>
</tbody>
</table>
3.4 Analysis Procedures

We conducted two types of analyses. First, we conducted Bayesian analyses to explore the differences between men and women software professionals (Sec. 3.4.1). Second, we conducted network analyses to explore the relationships between personality traits, analyzed by gender (Sec. 3.4.2). The analyses were carried out with IBM SPSS version 26 (to validate the survey, manipulate variables, perform descriptive statistics, and verify the results), and the JASP statistical package (version 0.11.1) [95] for descriptive, Bayesian and network analysis.

3.4.1 Bayesian Analysis

Bayesian statistics has several benefits, as opposed to frequentist null-hypothesis significance testing. Typical reasons claimed by scholars are obtaining evidence in favor of the null hypothesis [97] i.e., understand how likely it is that the null hypothesis is valid providing a better understanding of the phenomenon; discerning between “absence of evidence” and “evidence of absence” [98], which is why it has also been advocated in our research community [99], [100].

One of the main reasons for the increasing level of popularity of Bayesian statistics among statisticians is that it overcomes typical shortcomings of p-values—based findings of frequentist null-hypothesis significance testing [101], [102]. Conceptually, the p-value is the probability of observing something significantly different, assuming that the baseline hypothesis (H0) is true. The p-value does not provide any information about the likelihood that a research hypothesis is correct, which is precisely what a Bayesian approach does. Researchers can assess the probability that H0 will happen over H1. Table 8 presents a set of heuristics for interpreting Bayes factors [96].


| TABLE 5 |
| Respondents’ country of origin |
| --- | --- | --- |
| Country | Frequency | Percent |
| United Kingdom | 141 | 29.2 |
| USA | 135 | 28.0 |
| Portugal | 33 | 6.8 |
| Poland | 22 | 4.6 |
| Italy | 18 | 3.7 |
| Canada | 15 | 3.1 |
| Germany | 12 | 2.5 |
| Spain | 9 | 1.9 |
| Ireland | 9 | 1.9 |
| Greece | 8 | 1.7 |
| Mexico | 8 | 1.7 |
| Australia | 7 | 1.4 |
| France | 6 | 1.2 |
| Hungary | 5 | 1.0 |
| Estonia | 4 | 0.8 |
| Other | 51 | 10.5 |

Generally speaking, Bayesian analysis addresses pervasive questions such as how much evidence do we have from our data against the null hypothesis?

The first step in our analysis was to determine whether we can use parametric tests or whether we should use non-parametric alternatives. To do so, we assessed whether the data followed a normal distribution. We divided our data by gender, resulting in a total of 18 different distributions, i.e., the result for nine traits per two genders. We used distribution and density plots, boxplots, and Q-Q plots to visualize the data for a prima facie assessment of distributional normality; all plots suggested normal distributions. To ensure that parametric tests were warranted, we performed a Shapiro-Wilk test; the results of this test fell between 0.942 and 0.985, (p < 0.05). Kurtosis and skewness values also were between ±1, supporting our assertion that the data followed a normal distribution. Both plots and tests are available in the appendix. Finally, to ensure not to perform an underpowered study, which may lead to biased results, we performed a
power analysis to compute the minimum sample size with G*Power [103]. The result for an error probability of 5% and a power of 95% is 88. Hence, since all preliminary tests supported the reliability and significance of our sample, we conducted a Bayesian independent samples t-test. Section 4.1 presents the results of this analysis.

3.4.2 Network Analysis

An alternative way of thinking about personality is to consider the traits as an “ecosystem” in which personality characteristics and behaviors interact and affect one another [26], [104]. This network perspective provides complementary insights to the analysis based on latent, variables explained above [26], whereby traits are treated as distinct constructs. A network perspective of personality accepts that there may be feedback between traits [104]. This means that, while personality traits tend to be quite stable over a person’s lifetime, specific events or contexts may alter behavioral patterns that is typically associated with a specific trait. Paraphrasing Cramer et al.’s example, an extrovert who keeps getting ignored while trying to make small talk with strangers might, ultimately, become disillusioned and give up trying [104]. Thus, individual behavior can be highly idiosyncratic due to this ‘organism-environment’ feedback loop [104]. Further, a network perspective also acknowledges that personality traits are not isolated but correlated. A person scoring high on Machiavellianism is unlikely to score high on Honest-Humility (see Tables 1 and 2), as these two traits are highly incompatible.

With this analysis, we aim to investigate the relations among the personality traits of software engineers. The interest for network analysis has been increasing during the last 20 years both in social science as in technical disciplines due to the rich insights that this approach can provide, such as the investigation of the proprieties of a particular network [105]. The concept of centrality is of particular interest, which is the role of a given node within a network. To gather a sharpened understanding of gender differences, we ran two separate networks, which resulted in two different graphs; one for men and one for women. Section 4.2 presents the results of the network analysis.

4 RESULTS

4.1 Bayesian Independent Samples T-Test

Figure 2 shows the degree of evidence of the two hypotheses. The top of the figure shows a “probability wheel” [64], which visualizes the ratio of the evidence for \( H_0 \) (white, not visible in this figure) to the evidence for \( H_1 \) (red). In this case, the wheel is fully red, indicating very strong evidence for \( H_1 \). The figure also demonstrates that a sample of only 200 samples would have provided only weak evidence for \( H_0 \). In the case of Extraversion (see appendix), there is moderately strong support for \( H_0 \), suggesting a difference between men and women on this trait to be unlikely.

4.1.1 Bayesian Analysis

Given the data followed normal distributions, we adopted parametric variants of the Bayesian analysis. To run this analysis, we used the Bayesian t-test framework proposed by Jeffreys [106]. This analysis comprises two major steps.

First, we assess the results for hypothesis testing. Second, we discuss the effects of parameter estimation. Table 9 characterizes how the scores vary by gender.

The two rival hypotheses are \( H_0; \delta = 0 \) and \( H_1; \delta \sim \text{Cauchy} \). The Cauchy distribution is a \( t \) distribution for which the mean and variance are undefined [107]. We use \( \delta \) as the standardized effect size, also known as Cohen’s \( d \) [108]. In other words, with \( H_1; \delta \sim \text{Cauchy} \), one group will be above the mean of the other group, and there is a significant chance that a person picked at random from one group (i.e., women) will have a higher score than a person picked at random from the other group (i.e., men). In our case, \( H_0; \delta = 0 \) means that there is no difference in personality traits between genders. Given the paucity of research on gender differences in personality traits, we do not have any useful prior knowledge on this topic, and so we adopted a default value for the prior for an independent samples t-test, which is a Cauchy distribution with spread \( r = \frac{1}{\sqrt{3}} (0.707) \) [107].

We first establish which personality traits differ significantly across men and women. Table 10 presents the BF values, which indicate the likelihood of the observations under \( H_1 \) versus \( H_0 \). Four traits have Bayes factors larger than 10, which is a minimum recommended value [109], [110]. Psychopathy has a Bayes factor of almost 80, suggesting that the likelihood that men score higher on this trait is 80 times more likely under \( H_1 \) than under \( H_0 \). Similarly, the table suggests strong evidence for Honesty-Humility, Emotionality, and Openness to Experience. The error percentage of the different tests are also very low, namely less than 0.001%, which suggests high stability of the algorithm used to compute the prediction. The appendix shows that the robustness of the Bayes factors concerning our prior specification in these four cases is also quite high. In particular, the Bayes factor Robustness Check shows the Bayes factors under different prior specifications. The rationale behind this check is that if the conclusion does not change through a range of different prior distributions, it is a strong indication of the robustness of the analysis.

In the second step, we consider the effects of parameter estimation. Here we make some considerations about the
posterior distribution δ which is the probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from the collected data (i.e., the standardized mean difference between gender groups). All values are within their 95% credible interval and fairly close to Cauchy. The four identified traits have smaller CI ranges, which means that if the effect is assumed to exist, its uncertainty is low. We can, therefore, assume that the likelihood for H1, as computed by the Bayesian Factors, is accurate.

After assessing the likelihood of a difference between the two gender groups, we are now interested to assess how they differ. Therefore, we look at the descriptive statistics in Table 9. For the four identified traits, we look for the higher mean value and their credible interval. Based on that evidence, we can assess which trait is significantly higher or lower. Accordingly, we conclude that women score significantly:

- lower in Psychopathy;
- higher in Honesty-Humility;
- higher in Emotionality;
- higher in Openness to Experience.

### Table 9: Descriptive Statistics

<table>
<thead>
<tr>
<th>Personality Trait</th>
<th>Gender</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machiavellianism</td>
<td>M</td>
<td>2.31</td>
<td>0.817</td>
<td>0.041</td>
<td>(2.230, 2.392)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>2.247</td>
<td>0.857</td>
<td>0.090</td>
<td>(2.068, 2.427)</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>M</td>
<td>2.471</td>
<td>0.751</td>
<td>0.038</td>
<td>(2.396, 2.545)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>2.150</td>
<td>0.719</td>
<td>0.076</td>
<td>(1.999, 2.301)</td>
</tr>
<tr>
<td>Narcissism</td>
<td>M</td>
<td>2.698</td>
<td>0.790</td>
<td>0.040</td>
<td>(2.620, 2.777)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>2.636</td>
<td>0.871</td>
<td>0.092</td>
<td>(2.454, 2.819)</td>
</tr>
<tr>
<td>Honesty–Humility</td>
<td>M</td>
<td>3.653</td>
<td>0.691</td>
<td>0.035</td>
<td>(3.584, 3.721)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>3.958</td>
<td>0.686</td>
<td>0.072</td>
<td>(3.815, 4.102)</td>
</tr>
<tr>
<td>Emotionality</td>
<td>M</td>
<td>2.722</td>
<td>0.657</td>
<td>0.033</td>
<td>(2.657, 2.787)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>3.156</td>
<td>0.693</td>
<td>0.073</td>
<td>(3.010, 3.301)</td>
</tr>
<tr>
<td>Extraversion</td>
<td>M</td>
<td>3.353</td>
<td>0.681</td>
<td>0.034</td>
<td>(3.286, 3.421)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>3.408</td>
<td>0.744</td>
<td>0.078</td>
<td>(3.252, 3.564)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>M</td>
<td>2.869</td>
<td>0.582</td>
<td>0.029</td>
<td>(2.811, 2.927)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>2.772</td>
<td>0.590</td>
<td>0.062</td>
<td>(2.649, 2.896)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>M</td>
<td>3.508</td>
<td>0.629</td>
<td>0.032</td>
<td>(3.446, 3.571)</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>3.567</td>
<td>0.669</td>
<td>0.071</td>
<td>(3.427, 3.707)</td>
</tr>
<tr>
<td>Openness to</td>
<td>M</td>
<td>3.761</td>
<td>0.592</td>
<td>0.030</td>
<td>(3.702, 3.820)</td>
</tr>
<tr>
<td>Experiences</td>
<td>W</td>
<td>3.969</td>
<td>0.532</td>
<td>0.056</td>
<td>(3.858, 4.081)</td>
</tr>
</tbody>
</table>

### Table 10: Bayesian Independent Samples t-Test

<table>
<thead>
<tr>
<th></th>
<th>BF10</th>
<th>Error %</th>
<th>Median(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machiavellianism</td>
<td>0.158</td>
<td>1.442 × 10⁻⁵</td>
<td>0.07</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>78.590</td>
<td>8.513 × 10⁻⁸</td>
<td>0.04</td>
</tr>
<tr>
<td>Narcissism</td>
<td>0.158</td>
<td>1.442 × 10⁻⁵</td>
<td>0.07</td>
</tr>
<tr>
<td>Honesty–Humility</td>
<td>113.897</td>
<td>6.051 × 10⁻⁸</td>
<td>0.43</td>
</tr>
<tr>
<td>Emotionality</td>
<td>275,953.415</td>
<td>3.241 × 10⁻¹¹</td>
<td>0.63</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0.160</td>
<td>1.435 × 10⁻⁵</td>
<td>0.07</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.335</td>
<td>8.351 × 10⁻⁶</td>
<td>0.14</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.172</td>
<td>1.358 × 10⁻⁵</td>
<td>0.09</td>
</tr>
<tr>
<td>Openness to</td>
<td>11.260</td>
<td>4.837 × 10⁻⁷</td>
<td>0.34</td>
</tr>
</tbody>
</table>

### 4.2 Network Analysis

We used network analysis to discover relations among personality traits and to analyze the structures of the relations of such traits, using graph theory and relational algebra. From an operational analysis perspective, data are organized in a relational matrix, personality traits are represented as nodes, and their relations as edges between pairs of nodes.

#### 4.2.1 Network Structure

To estimate the network structure, we used the `EBICglasso` function, which is a combination of the Extended Bayesian Information Criterion [111] and the Least Absolute Shrinkage and Selection Operator [112] with the automatic correlation method and normalized centrality measures. This approach supports better visualization since small edge weights are neglected from the model, allowing to only focus on those relationships that are significant.

The network analysis reveals two polarized sub-networks. The Dark Triad, together with Honestly-Humility, and Agreeableness are positioned on one side, while Conscientiousness, Openness to Experience, Extraversion, and Emotionality are on the other side. Both networks show tight and positive interactions among dark traits (Nar, Mac, Psy). Dark traits are negatively linked with bright ones, in particular with Honestly-Humility and Agreeableness. This first finding is not surprising since the two aforementioned bright traits are related to sincerity, greed avoidance, forgiveness, and gentleness. Both men and women software engineers show consistent trait relationships; for example, the relationship between Machiavellism and Honestly-Humility is a negative one for both men and women. Thus, we can exclude a personality disorder, which is the typical case where we see a group of personality traits isolated from other traits [113].

Looking closer at the network differences, Figure 3a shows that Conscientiousness is on the right-hand side of the graph (i.e., the distance of such node with the others on the same side is rather close), although it is very weakly linked to the other three traits (Openness to Experience, Extraversion, and Emotionality). From a practical perspective, these traits would appear to be similar, despite the fact that they do not correlate with each other. On the other hand, Figure 3b shows a more densely connected network. Women have more relations between the traits (see also the appendix) (14 non-zero edges), with respect to men (12 non-zero edges), suggesting higher personality complexity [114]. Women’s traits are more correlated than men’s, meaning that a single trait cannot be interpreted in isolation, but must be explained in relation to other traits.

Network stability was assessed by computing a non-parametric bootstrap analysis (re-sampling subsets of the data with replacement and using the same sample size) [115] with 1,000 iterations. We can see that the network edges remain stable across these 1,000 sub-samples. Fig. 20 in the appendix shows that the sample fits well with the bootstrapped mean and that each of the estimated edge weights is within the bootstrapped confidence intervals.

#### 4.2.2 Centrality Analysis

To explore which personality traits are the most influential, we performed a centrality analysis. Centrality is the degree to
Fig. 3. Network analysis of personality traits of men (a) and women (b) (software engineers). Nodes relations can be positive (blue) or negative (red). Nar: Narcissism, Mac: Machiavellism, Psy: Psychopathy, HonH: Honesty-Humility, Emo: Emotionality, Ext: Extraversion, Agr: Agreeableness, Con: Conscientiousness, OtE: Openness to Experience

which a node occupies a central position in the network. Central nodes are better positioned to spread information across a network. We analyzed centrality using three measures: Betweenness centrality, Closeness centrality, and Degree centrality [116], which are visualized in Fig. 4. The first two are related to the shortest paths of the network, i.e., the least number of steps to reach one node from another.

Betweenness centrality is the number of shortest paths of a given node. In our case, the betweenness of Psychopathy and Extraversion is high for both genders. So, the paths to pass through those traits are shorter than others, and it is also easier to pass from the other nodes to Psychopathy and Extraversion with respect to the others. Those two traits are the positive (Extraversion) and negative (Psychopathy) gatekeepers of personality. In practical terms, this suggests that if an organization wants to plan a change within the organization, e.g., the introduction of a new tool, extrovert people may be more suitable to do this successfully. On the other hand, psychopaths are more likely to oppose any change [117]. While such findings might seem obvious, they have to the best of our knowledge never been scientifically demonstrated so far in the context of software development.

In terms of gender differences, based on the centrality analysis, we suggest that men scoring high on Honest-Humility, and women with high Agreeableness scores are also more likely to advocate organizational changes within an organization. This hypothesis is also confirmed by general population studies, showing how the Honest-Humility trait is a predictor of positive attitudes to workplace diversity [118], and support organizational change to improve performances [62]. Similarly, Agreeableness is also considered a relevant predictor for working performance, also by supporting organizational changes [119]. Not surprisingly, Honest-Humility and Agreeableness are typically highly correlated traits [70].

Closeness centrality is the inverse sum of all shortest paths from the node of interest to all other nodes in the network. It describes how much one node is responsible for spreading information to others. Here, we confirm what we have observed in Figures 3a and 3b regarding the higher degree of polarization since men’s traits (5 nodes) are higher than women’s (2 nodes). This can be considered as a consequence of a higher personality complexity for women since their traits are more balanced and less extreme.

Degree centrality is the sum of the absolute input weights of one specific node. It predicts the direct impact of one node on the entire network. Also, it characterizes the centrality of a node within the network. Practically, in a scenario of very limited resources, software engineers who exhibit high Honest-Humility levels will commit to organizational change. This evidence also resonates well with the fact that this trait is associated with pro-social behavior, treating people fairly, and being unconcerned with self-promotion. Also, people high in Honest-Humility tend to have good job performances (i.e., this predictor outperforms the other five factors for job performance [61]). On the opposite side of the Degree graph we have Conscientiousness, which is not surprisingly, typical for career-driven and risk-averse people [120]). Hence, due to their risk-aversion, conscientious professionals will most likely be late adopters of any change within the organization. Finally, since the Degree values for men and women mostly overlap (see the graph on the right-hand side of Fig. 4) we cannot draw distinct conclusions regarding gender.

5 Discussion
In this article we seek to shed light on patterns of personality traits among software developers, and to interpret these patterns in light of implications for software development organizations. We wish to emphasize that this study does not seek to point out specific, individual detrimental behaviors of employees. Similarly, we do not claim that our findings apply to each individual software engineer. Instead, the findings of this study represent average effects; for example, we do not assert that all men in our sample of software professionals are psychopaths while women are not. We
do claim, based on the findings of this study, that men score on average higher than women on this particular psychological trait. We highlight the likelihood that specific traits are prevalent to a specific gender, and their meaning, in aggregate terms, within software development environments. Our findings can help software organizations to predict work and team performance of new employees based on their gender, and also understand, from a personality perspective, which employees might be best suited to evangelize new projects.

5.1 Implications for Research and Practice
The results of our study have several important implications; Table 11 provides a summary. First, within the context of software development, there are clear personality differences between men and women. Such differences likely have an impact on the way of working, how tasks are performed, or how interactions unfold with a team. Second, we suggest that differences in personality could, or perhaps should, affect how software organizations pursue initiatives for process improvement. We elaborate on both points below.

Gender personality differences are also relevant in the general population, showing a higher score in Emotionality and Agreeableness for women [121]; therefore, it is unsurprising that we identified differences between men and women software professionals. Similarly, men score higher on the so-called Dark Triad traits, especially psychopathy, than women [122]. Thus, our results are in line with and confirm previous psychology research, suggesting new and specific insights for software professionals.

In particular, we found that women score higher in Honesty-Humility. This personality trait is highly related to work performance [61], [62], [63]. Johnson et al. found that it is the best predictor of job performance [61]. A large-scale study by Owens et al. extended these results [62]. Honesty-Humility is a predictor of both individual performance and contextual performance, i.e., quality of team member contribution. Also, this trait compensates for lower general mental ability and increases, in case of a managerial role, employee retention, engagement, and job satisfaction. Finally, employees scoring high on the Honesty-Humility trait show a significantly lower degree of workplace delinquency and also serve as a great moral example among peers [63]. People who score high on this trait tend to be aware of their limits and exhibit a willingness to compensate for their weaknesses, have a good understanding of their role, have a strong work ethic, and exhibit a high level of commitment. Therefore, Honesty-Humility professionals are precious in software teams, as they can serve as role models to other team members.

The results for team performance are more complex. Women software professionals score high both in Emotionality (which decreases team viability through social cohesion [123]), and Openness to Experience (which supports team performance [124]). Other scholars have found similar results; for example, Bradley et al. concluded that openness to experience and emotional stability are essential moderators of the relationship between task conflict and team performance [125]. An individual factor is creativity; especially when working in teams, Openness to Experience is significantly related to team creativity [126]. On the one hand, it seems that the high level of emotionality of women might affect a team’s cohesion and performance negatively. On the other hand, women’s higher levels of Openness to Experience may mitigate such adverse effects, leading teams to be more creative and receptive, strengthening the team spirit.

Adding another significant difference, such as Psychopathy, might offer a better understanding of team performance. Among the Dark Triad, Psychopathy is the most detrimental trait for team performance [122]. Teams with psychopathic members decrease their level of innovativeness, creativity, commitment, leading to a revenue decrease [127]. Psychopaths tend to exhibit anti-social behavior, such as bullying towards colleagues, which may temper their motivation and increase the odds of members leaving the team or even the organization [128]. Although men have substantially lower Emotionality levels than women, their team performance is mitigated by higher psychopathy levels.

To summarize, women software engineers significantly differ from men in terms of personality traits, which are related to higher job performance, ethics, and creativity. Men, despite having lower scores on Emotionality, exhibit higher scores on the Psychopathy trait, which may lead to a reduced level of team performance. Taken together, we offer the first two propositions:

Proposition 1. Including women in software teams increases team performance and decreases workplace delinquency such as absenteeism and alcohol abuse.

Proposition 2. As both men and women exhibit negative and positive traits linked to teaming, mixed-gender teams will perform better than non-mixed teams.

Earlier we offered an explanation of the likelihood that specific personality traits are significantly different in men than in women. Such information is descriptive. Looking into the relations of personality traits provides both researchers and practitioners with a better understanding of leveraging specific traits to drive organizational changes.
and transformations. As new practices, processes, and tools become popular, organizations will seek to exploit these. It is common that organizations’ top decision makers impose such changes from the top. It is also common, however, that process improvement initiatives do not achieve the desired outcome. For example, many organizations have sought to introduce large-scale agile transformations, however, the degree of success varies considerably [129]. For organizations to achieve success in such initiatives, it is important to identify champions who help to convince and entice others [130]. A key question is, then, how such champions can be identified; what behavior characteristics and personality traits might such evangelists possess? The network analysis can offer answers to this question.

Based on the Betweenness outcome of the Network analysis, we previously concluded that software professionals who score high in extraversion are the best candidates to become organizational change agents, or champions. This insight is also substantiated by previous management scholars [131]. However, if there is a specific need to appoint a man or a woman to the role of evangelist, the choice should rely on men scoring high on honest-humility or women scoring high on agreeableness (see Fig. 4). Psychopaths, on the other hand, will generate the opposite effect, opposing any action towards change. If top-management wants to address a long-term and organization-wide transformation psychopaths will likely oppose it. Hence, we offer Proposition 3:

**Proposition 3.** Extrovert employees are best suited to drive long-term, organization-wide transformation processes.

If, on the other hand, leadership wishes to evaluate minor changes to see how they affect, for example, development teams, our findings suggest that people who score high on Honest-Humility are best suited. An organization can identify a few teams and invite those members who score high in the Honest-Humility personality trait to start using a new coding practice or tool, and asking them to explain it to the others. Only the honest-humility developer would likely attend the training, leaving others with their development tasks. Once such developers go back to the team, they can teach their team members the new practice or tool through a peer-learning approach [132]. Hence, we introduce Proposition 4:

**Proposition 4.** Software professionals who score high on the Honest-Humility personality trait are best suited to conduct pilot evaluations of new practices and tools.

This study suggests that women display a higher personality complexity. As Razavian and Lago suggested, women software professionals can deal more effectively than men with complex social tasks such as developing and imagining mental models of customers, bridging such understandings effectively with development teams [52]. One study found that women have a different approach to program comprehension as they showed a tendency to begin at a low level of abstraction and move towards a higher level, i.e., bottom-up of code comprehension [133]. However, the limited number of studies and subjects involved in investigating specific software engineering tasks using cognitive neuroscience research tools, such as functional magnetic resonance imaging (fMRI), did not provide any significant gender-related differences as of yet and are typically addressed as a limitation [134].

**Proposition 5.** Women software engineers can deal better with complex social tasks, especially in relation to people.

Finally, all those considerations are very likely to be stable in time and consistent with future studies. The reason is that personality traits are stable over time, i.e., they do not change along the maturation process of people [23]. This stability has also been recently confirmed for software developers [24].

### 5.2 Threats to Validity

We adopt Gren’s five-facets framework to discuss the threats to validity of this study [135].

**Reliability.** The use of Bayesian statistics provides here a great degree of control over the collected data. Since we were able to assess the stability and consistency of our results through sequential analysis combined with robustness analysis, we conclude that our data are reliable.

This is a sample study with self-reported values, which might limit the study’s validity. To address this, we followed a rigorous data collection process, leading to 483 validated questionnaires considering the 760 initial subjects who started the competence screening phase. To our knowledge, this is the first sample study in the software engineering literature that followed such a rigorous selection.

**Construct validity.** The two measurement instruments we used, reflect the purposes for which it was developed, namely to measure bright and dark traits on individuals. Indeed, they have been developed by personality scholars, grounded in well-established theories. Similarly, the wording of the test and its perception of the participant has been substantiated by using widely used and validated measurement instruments. Also, both work well through different cultural groups, leading to generalizable findings. Moreover, suspicious, unreliable, or unlikely answers were discarded along our data collection process. However, we do recognize that we used short versions of the original inventories, which were slightly less accurate. Nevertheless, we made such a trade-off since both inventories estimate with high accuracy the original long inventories, giving us the possibility to engage our informants with two tests: the HEXACO and the Dark Triad, with a manageable drop-out ratio.

**Conclusion validity.** Using Bayesian statistics and overcoming the shortcomings of p-values based findings of frequentist null-hypothesis significance testing provided us with a fairer understanding of the investigated phenomenon. Our results show the likelihood of a personality trait difference among the two groups and the degree of it, rather than a positive or a negative answer. Two of the observed differences are also common to the general population (psychopathy and emotionality), confirming the soundness of our conclusions.

**External validity.** The sample study strategy adopted in this research has a high level of potential to achieve generalizability [76]. To that end, we made a considerable effort to establish a representative sample of the software engineering population. Following Yamane’s formula [86] and the highest estimate of the population at the time of this study, we

9. These instruments have been used in nearly 1,000 articles, primarily in social psychology: www.hexaco.org/references.
established a minimum sample size of 400; our sample size of 483 is well above that. Until now, no other personality study of professional software engineering reached such a threshold. Moreover, the demographics which we collected were comparable to other large-scale surveys, e.g., The State of Developer Ecosystem by Jet Brains. Another limitation is that we collected gender as a binary characteristic. We did so because this study relied on statistical techniques that require large samples, and so this study only represents software developers who identified as men or women.

6 Conclusion

Several scholars have expressed concern about a lack of diversity in software engineering teams [47], [136], [137], and indeed, there is increasing attention for diversity within the SE literature. Several recent studies have discussed gender-related issues, focusing on differences between (primarily) men and women. Another stream of studies within the SE literature has focused on personality traits [138], and the role that personality plays in building teams. Very few studies have, however, combined these two perspectives, which led us to pose the question: what are gender differences in terms of personality traits among software engineers? The literature on the role of gender differences is still in its nascent phase, though it has grown considerably in recent years.

This study seeks to understand how personality traits differ in men and women in the software industry, and to develop an understanding of the potential implications of such differences for software development workplaces. We collected data through a sample study of almost 500 software engineers, and have analyzed these data using Bayesian statistics—as such, this study also contributes to the literature as a showcase of Bayesian analysis.

We identified four main differences in personality between men and women. Women score lower in Psychopathy, higher in Honesty-Humility, higher in Emotionality, and higher in Openness to Experience than men. The relations between traits is also different for men and women. Drawing on our findings and linking these to social psychology literature as well as previous gender studies in the software engineering literature, we presented a number of propositions (Table 11) that could guide future research.

To conclude, this article contributes to the nascent literature on gender-related personality studies. While considerable attention has been dedicated to both gender studies and personality traits research in software engineering, we observe that few studies have attempted to investigate both aspects, and no previous study has sought generalizability to the larger population of software developers utilizing a sample study. We hope that this study provides a useful starting point for future work on gender-related studies and personality traits from a range of perspectives such as team dynamics, considering groups as a unit of analysis. Further, whereas our study is limited in that it employed a binary gender identity, future work could enrich this field further.

Acknowledgments

We thank the three anonymous reviewers for constructive feedback. We are particularly grateful for the reviewers’ guidance in relation to our discussion of gender. This work was partially funded by the Science Foundation Ireland grant 15/SIRG/3293 and 13/RC/2094.

References


Daniel Russo is an Assistant Professor in the Department of Computer Science, Aalborg University. He holds a PhD in Computer Science & Engineering from the University of Bologna and was a Postdoctoral Researcher at Lero—the Irish Software Research Centre. He is a member of the ACM Future of Computing Academy.

Klaas-Jan Stol is a lecturer in the School of Computer Science and Information Technology, University College Cork (UCC), and a Science Foundation Ireland Principal Investigator. He is a Funded Investigator with Lero—the Irish Software Research Centre, and leads the UCC Software Engineering Research (USER) Group.