This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 110.1109/seaa60479.2023.00042, 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

The Paradox of Analysing Gender-Based Data

Steve Counsell
Brunel University
steve.counsell@brunel.ac.uk

Emily Winter
Lancaster University
e.winter@lancaster.ac.uk

Tracy Hall
Lancaster University
tracy.hall@lancaster.ac.uk

Vesna Nowack
Imperial College, London
v.nowack@imperial.ac.uk

***Abstract*—In this short paper, we analyse "gender" perspectives from a survey of three hundred and seventy-eight industry developers on two aspects of IT industry developer practice: bugs and Automatic Program Repair. We also explore questions of how developers view their job satisfaction. Our key motivation was to show whether there was a difference in the way that males and females viewed these three important concepts. From a total of thirteen survey questions analysed, only two showed any statistical difference between the responses of females compared to males. Those differences were found exclusively in the job satisfaction part of the survey. In terms of the way that male or female developers *think* about technical activities *per se* and diversity and inclusivity more generally, we therefore have the paradoxical issue of whether gender comparisons have any basis. We all think the same way about technical-oriented activities, so perhaps we need to stop trying to find differences and division.**

***Index Terms*—Gender, survey, APR, bugs.**

I. Introduction

The imbalance and lack of diversity between genders in the IT world, whether at industry developer or at first year University level is well-known [1]–[3]; many of the science-based subjects (in the UK at least) have disproportionate numbers of males compared to females. Women remain severely underrepresented in engineering - just 12% of those working in engineering are female, compared with 47% of the overall UK workforce, despite significant effort to address this imbalance[1]. In this short paper, we describe gender perspectives from a survey showing that there are virtually no differences between the way males and females think about technically-oriented aspects of their jobs. The two topics we consider are bugs and Automatic Program Repair (APR). We also explore perceptions of job satisfaction by both males and females. Three hundred and seventy eight participants took part in the survey and results showed that, from a total of thirteen survey questions, only two showed statistical difference between the responses of females compared to males. Moreover, significance was found exclusively for the job satisfaction parts of the survey and not for the more technical questions about bugs and APR. We note that a far wider survey of APR (not gender-based) were first published in a full form by Winter et al., [4].

Two main ways of recruiting participants for the main survey were used. Industry contacts were targeted, as well as social media channels. From this source, 76 responses were obtained. We also used the online platform Prolific[2]. Prolific is designed specifically for academic research and allows participants to be selected and filtered. In this paper, we used 51 female participants and 327 male participants and these numbers according to [4] reflect the proportions of males and females in industry[3].

[1]engineeringuk.com/media/1691/gender-disparity-in-engineering.pdf
[2]https://www.prolific.co

II. Question analysis

In this section, we analyse the three sets of survey data. Firstly, perceptions by male and female participants on bugs; secondly, on the topic of APR and, finally, perceptions on job satisfaction. To facilitate each analysis, we undertook an independent samples Mann-Whitney test to determine if there was a significant difference between how each group perceived questions on each topic. Mann-Whitney is a non-parametric test to determine whether differences exist between two independent groups [5]; it assumes a non-normal distribution. We report, for each analysis, a Z-score and a p-value. A p-value value of $<=0.05$ represent a significant difference between the two groups. If true, we reject the null hypothesis of "no difference" and conclude that there *is* a difference between genders in the way they responded to questions. A p-value >0.05 implies that we cannot reject the null hypothesis, concluding there is no difference between the two groups.

A. Survey questions on Bugs

Question 1 explored views on finding and fixing bugs. Each participant was asked: "Finding and Fixing bugs is:" and the Likert scale options provided were: "Never challenging (5), Always challenging (1) and (4), (3) and (2) values between the two extremes. A Mann-Whitney test was conducted and gave a Z-score of -0.95, p-value value of 0.35 (this and all tests in this paper henceforth are two-tailed). We therefore conclude that there was no significant difference between male and female developer views on the challenge faced when finding and fixing bugs (and we cannot reject the null hypothesis in this case). **Question 2** asked participants about the same topic: "Finding and Fixing bugs is:" and the options were: "Never meaningful (5), Always meaningful (1) and (4), (3) and (2), again representing values between the two extremes. A Mann-Whitney test produced a Z-score of -1.18 and p-value 0.24. Again, we conclude that there is no significant difference between male and female groups on how they perceived the meaningfulness of finding and fixing bugs and, again, we cannot reject the null hypothesis. Next, (**Question 3**) asked: "Finding and Fixing bugs is:" and the options were:

[3]https://www.womenintech.co.uk/8-facts-women-tech-industry

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 110.1109/seaa60479.2023.00042, 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

"Never satisfying (5), Always satisfying (1)" and (4), (3) and (2) options in between. A Mann-Whitney test gave a Z-score of -1.02 (p-value 0.31). A strong similarity therefore existed between the way males and females perceived the satisfaction they derived from finding and fixing bugs (again, we cannot reject the null hypothesis). There is no significant difference between how each perceive satisfaction from fixing bugs.

Question 4 asked: "Finding and Fixing bugs is:" and the options were: "Never frustrating (5), Always frustrating (1)" and (4), (3) and (2) in between. A Mann-Whitney test gave a Z-score of -0.66, with p-value of 0.51. We cannot reject the null hypothesis (this was actually the strongest of the set of results); Fig. I shows this result graphically and we can see a strong tendency by males and females to choose both option 4 and 5 in similar numbers. Option 3, on the other hand, tends to be favoured by male participants.

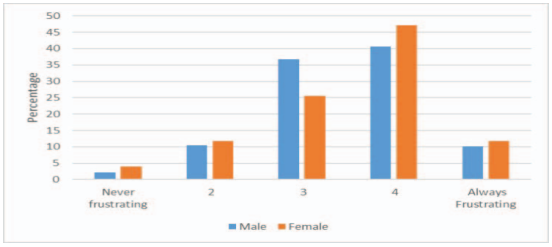

Fig. I. Frustration with finding and fixing bugs

Question 5 asks: "My bug finding and fixing is:" and the options provided were: "Never successful (5), Always successful (1)" and (4), (3) and (2) in between. Yet again, the Mann-Whitney test revealed a Z-score of -0.10 with p-value of 0.32. We cannot once more reject the null hypothesis; no observable difference between females and males on this question could be found. Finally, **Question 6** asked subjects for their views on what makes a bug annoying to fix. They were asked: "What makes a bug particularly annoying to fix?" The text of the question was as follows: "Please choose the three options that most apply". The answers provided were:

- When it's in very old code (1)
- When it's somebody else's [code] (2)
- When it's my own mistake (3)
- When it involves multiple files/parts of the system (4)
- When it relies on an API (5)
- When it's in very complex code (6)
- When it's in poorly documented code (7)

For this question, we simply compared the percentage of choices in each of the 7 categories. Figure II shows the comparison of male and female responses for this question; many of the percentages are similar.

In particular, this is true for options 2, 3, 4 and 7, where there is very little difference in the percentages. The biggest difference was for option 1 (here, 16.51% of males and 25.49% of females chose: "when it's in very old code"). Equally, option 5 shows a difference in opinion (15.60% of males and just 3.92% of males chose this option). The general trend

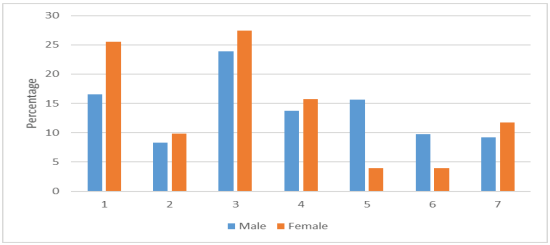

Fig. II. What makes a bug difficult to fix?

in the data therefore tended towards agreement more than disagreement. (We note that a Mann-Whitney test for this question also produced no significant p-values.) *To conclude then, from the six questions posed about bugs, none showed any significant difference between male and female views.*

B. Survey questions on APR

In this section, we discuss the responses from participants to questions related to APR. The topic is a burgeoning area of software engineering research and one that is attracting increasing interest in industry [6]–[9]. The purpose of APR is to replace the onus of bug-fixing on the developer with an automated bug-fixing process; its benefits are reasonably clear. Firstly, if bugs can be patched automatically, then it may free up time for developers to spend on other development activities. Secondly, it removes the "trial and error" approach that characterises the debugging process and the effort of manual regression testing. According to [4], the topic of APR was introduced to participants through a section explaining what it was and how it worked.

The first question in this section (**Question 1**) asked participants: "How would you feel about using an automatic software repair tool that found and fixed bugs?". The six possible responses were: "Very positive, Somewhat positive, Neither positive nor negative, Somewhat negative, Very negative and Don't know". Figure III shows the profile of responses on a male/female basis.

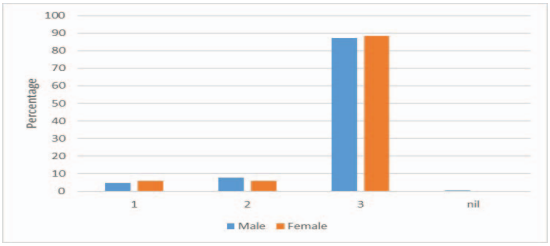

Fig. III. Feelings on the use of an APR for bugs

The first row of Table I shows the results of the Mann-Whitney test for this set of responses and, as *per* the results for bugs, no significant difference was found between male and female participants. **Question 2** asked developers to: "rank the following options according to your preference:

1) An automatic software repair tool that automatically applies fixes (1).

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 110.1109/seaa60479.2023.00042, 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

2) An automatic software repair tool that provides developers with fixes to approve (2).
3) An automatic software repair tool that provides developers with different fixes to choose from (3)".

Results are shown in rows 2, 3 and 4 of Table I which we have labelled as **Questions 2a, 2b** and **2c**. These show the Z-scores for comparisons of choices in each of the three items in the list. For example, Question 2a is the comparison between females and males for choice of option (1) above, Question 2b for option (2) above and Questions 2c for option (3) above. A clear lack of any agreement between subjects for any of the questions is evident by p-values (although the p-value for 2b is quite close to the 0.05 threshold); 88.24% of female developers chose option 3 and the corresponding value for males for that option was 87.16%. **Question 3** asked: "Please rank the following according to when an automatic software repair tool would be useful to you? (From 'most useful' at the top to 'least useful' at the bottom)". The options were: "Bugs found during development (1) Bugs found during testing (2)" and "Bugs found post-release (3)" We found no significant differences between the set of three responses for this question, as can be seen in the Table entries labelled Question 3a, 3b and 3c. For the third option (bugs found post-release) there was almost unanimous agreement between males and females, with Z-score of -0.04 and p-value 0.97. This illustrates the overarching point of our analysis:- that there seems little point in trying to find differences between male and female participants when discussing technical subjects, simply because they will generally tend to agree.

TABLE I
SURVEY QUESTIONS ON APR

Question No.	Z-score	p-value	Sig. (Yes/No?)
Question 1	-1.33	0.18	No
Question 2a	-0.07	0.94	No
Question 2b	-1.80	0.07	No
Question 2c	-1.50	0.12	No
Question 3a	-0.79	0.43	No
Question 3b	-0.79	0.43	No
Question 3c	-0.04	0.97	No
Question 4a	-0.20	0.84	No
Question 4b	-0.11	0.91	No
Question 4c	-0.29	0.77	No
Question 4d	-1.10	0.27	No
Question 4e	-0.10	0.32	No
Question 4f	-1.03	0.30	No
Question 4g	-0.53	0.60	No

Question 4 asked: "When you think about automatic software repair and automatically generated patches, how far do you agree with the following statements?" The possible rankings were: "Strongly disagree (1) Somewhat disagree (2) Neither agree nor disagree (3) Somewhat agree (4) Strongly agree (5) Don't know (6)". The statements with which they had to agree or disagree were:

- "Automatically generated patches would help save me time (1).
- Automatic software repair would not be able to fix complex bugs (2).
- I would be worried about the accuracy of automatically generated patches (3).
- I would find an automatic software repair tool useful (4).
- An automatic software repair tool would make my job easier (5).
- Human-written patches are more reliable than automatically generated patches (6).
- Automatic software repair tools might make software developers complacent (7)."

Table I shows the Mann-Whitney Z-scores and p-value for each of the seven options in the list above; these are labelled Questions 4a-4g. The most extreme value in terms of agreement was for option 2 (Automatic software repair would not be able to fix complex bugs). This was closely followed by option 1 (Automatically generated patches would help save me time). Surprisingly, the largest disagreement was for option 4 (I would find an automated software repair tool useful). Inspection of the survey data from the replication package for this option showed that males tended to "Somewhat agree" more than females, who, in turn, tended to opt for "Neither agree or disagree" to a larger extent than males; how this result could be interpreted needs further research.

C. Survey questions on job factors

The first question in this section **Question 1:** asked about the meaningfulness of the work of participants. The options were: "My work is... Never meaningful (1), Always meaningful (5)" and options 2, 3 and 4 for values in between. In this case, we found a significant difference between male and female views in their responses. A Z-score of -2.70 and p-value 0.01 showed that males and females thought significantly differently. So, we reject the null hypothesis and conclude that there is a difference between views on how male and females perceive the meaningfulness of their work. Figure IV shows the distribution in percentages for this question. We see a clear difference between the two groups. In some cases, male responses are significantly higher than female responses and in other cases the opposite is true. Generally speaking however, females tended to under-estimate the meaningfulness of their work. Responses of 2 and 3 were dominated by females, whereas male responses tended to be higher in the 4 and 5 category options. This could suggest that females felt that their work was less worthwhile compared with males and were consequently less confident with the scores they gave. Further research will explore this facet of the study in more depth.

We then conducted the same test for the second of three questions, **Question 2** in this category. This question was related to the frustration felt by developers in their jobs. The question posed was as follows: "My work is..." and the options were "Never frustrating (1), Always frustrating (5)" and values of 2, 3 and 4 in between. We again ran a Mann-Whitney test and found that the Z-score for this analysis was -2.75 and with p-value of 0.01 (two tailed). Once again we reject the null hypothesis; there is a significant difference between male

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 110.1109/seaa60479.2023.00042, 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

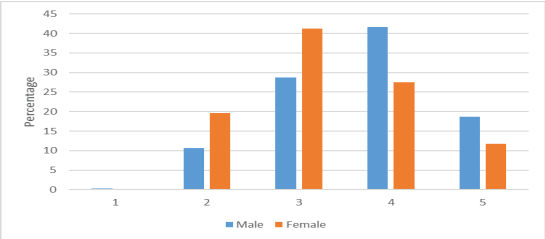

Fig. IV. Meaningfulness of subjects' work

and female in terms of the frustration they felt in their role. Exactly why and how that arose of course is unknown in this context. The numbers actually showed that 41% of females found their work to be in the 4 category, compared with just 22.02% of males. Put another way, female participants were far more likely to express high frustration with their work.

Finally, **Question 3** asked whether participants thought they were successful at work where "Never successful = 1" and "Always successful = 5" with values of 2, 3 and 4 in between. This produced a Z-score of -1.06 and p-value of 0.29. In this case, we cannot reject the null hypothesis and we therefore conclude that there is a similarity between male and female subjects in terms of how successful they felt they were at work.

III. Study Context

A short paper sadly precludes a full, page-long description of literature in the area; we instead provide a "research" context. Our work is supported by Russo et al., [10]. Their research explored personality data from 483 software engineers with a view to studying differences between genders. The work was motivated by the need to understand differences in team performance across genders. Results suggested that in terms of personality, females scored significantly higher in Openness to Experience, Honesty-Humility and Emotionality than males. Males show higher psychopathic traits than females! Baltes et al., [11] stresses the need to reach out to professional developers as an essential part of empirical software engineering. The work reported on experiences with different sampling strategies used and motivated the need to record demographics about software developers (for external validity purposes). We agree entirely with their paper's sentiment, but disagree that gender data should be collected, unless it is used for a dedicated purpose such as personality analysis. Work by Kaiser et al., [12] reported personality differences between males and females; in summary these could be stated as differences between: sensitive, aesthetic, sentimental, intuitive, and tender-minded (for females) versus utilitarian, objective, unsentimental and tough-minded (for males). May et al., explore the role of gender on stack overflow [13]. One of their conclusions was that: "the average woman has roughly half of the reputation points, the primary measure of success on the site, of the average man." A hypothetical redesign of the site's scoring system was one of their key recommendations.

IV. Conclusions and Future Work

In this paper, we described the results of an analysis of gender-based responses from a wider survey carried out on the topic of APR by [4]. We used data from 378 industrial developers as a basis of our analysis and showed that while there were differences in terms of job satisfaction of males and females, when it came to bugs and APR (i.e., technically-oriented aspects of their jobs), no real differences were observed. We therefore question whether gender comparisons have any basis whatsoever in this context. Maybe this should be more the domain of social science studies, rather than computer science. For the sake of inclusivity and diversity in the workplace, we feel strongly that issues such as those analysed are relevant and important [14]. To our knowledge, this is the first analysis of gender-based questions in surveys. Materials from this study are provided at: https://github.com/winterem/APRsurvey.

Acknowledgment

This work was funded by the Engineering and Physical Sciences Research Council, UK (grant number EP/S005730/1).

References

[1] L. J. Sax, K. J. Lehman, J. A. Jacobs, M. A. Kanny, G. Lim, L. Monje-Paulson, and H. B. Zimmerman, "Anatomy of an Enduring Gender Gap: The Evolution of Women's Participation in Computer Science," *The Journal of Higher Education*, vol. 88, no. 2, pp. 258–293, 2017.

[2] E. Winter, L. Thomas, and L. Blair, "'it's a bit weird, but it's ok'? how female computer science students navigate being a minority," in *Innovation and Tech. in Comp. Sci. Educ.* ACM, 2021, pp. 436–442.

[3] A. Durndell, "The persistence of the gender gap in computing," *Computers Education*, vol. 16, no. 4, pp. 283–287, 1991.

[4] E. Winter, D. Bowes, S. Counsell, T. Hall, S. Haraldsson, V. Nowack, and J. Woodward, "How do developers ¡italic¿really¡/italic¿ feel about bug fixing? directions for automatic program repair," *IEEE Transactions on Software Engineering*, pp. 1–20, 2022.

[5] A. Field, *Discovering Statistics using SPSS*. Sage Publication, 2006.

[6] A. Marginean, J. Bader, S. Chandra, M. Harman, Y. Jia, K. Mao, A. Mols, and A. Scott, "Sapfix: automated end-to-end repair at scale," in *International Conference on Software Engineering, Montreal, Canada, 2019*. IEEE / ACM, 2019, pp. 269–278.

[7] J. Bader, A. Scott, M. Pradel, and S. Chandra, "Getafix: Learning to fix bugs automatically," *Proc. ACM Program. Lang.*, vol. 3, no. OOPSLA, Oct. 2019. [Online]. Available: https://doi.org/10.1145/3360585

[8] S. Kirbas, E. Windels, O. McBello, K. Kells, M. Pagano, R. Szalanski, V. Nowack, E. R. Winter, S. Counsell, D. Bowes, T. Hall, S. Haraldsson, and J. Woodward, "On the introduction of automatic program repair in bloomberg," *IEEE Software*, vol. 38, no. 4, pp. 43–51, 2021.

[9] C. Le Goues, M. Dewey-Vogt, S. Forrest, and W. Weimer, "A systematic study of automated program repair: Fixing 55 out of 105 bugs for $8 each," in *Intl. Conference on Software Engineering*, 2012, pp. 3–13.

[10] D. Russo and K. Stol, "Gender differences in personality traits of software engineers," *IEEE Trans. on Software Eng.*, pp. 1–1, 2020.

[11] S. Baltes and S. Diehl, "Worse than spam: Issues in sampling software developers," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 2016.

[12] T. Kaiser, M. Del Giudice, and T. Booth, "Global sex differences in personality: Replication with an open online dataset," *Journal of Personality*, vol. 88, no. 3, pp. 415–429, 2020.

[13] A. May, J. Wachs, and A. Hannák, "Gender differences in participation and reward on stack overflow," *Empirical Software Engineering*, vol. 24, no. 4, pp. 1997–2019, feb 2019.

[14] B. Trinkenreich, R. Britto, M. A. Gerosa, and I. Steinmacher, "An empirical investigation on the challenges faced by women in the software industry: A case study," in *International Conference on Software Engineering, Pittsburgh*. IEEE/ACM, 2022, pp. 24–35.