Developers' estimations about their users' behaviour

Lina Witzel[*] Sophy Chhong[†]

Department of Informatics
University of Zurich

ABSTRACT

As a developer for end-user software an important concern is whether the software is usable and fulfils the needs of the user. In addition to user research, developers usually make assumptions about their users and try to estimate their behaviour. We wanted to explore how accurate those estimations are and whether it might be influenced by age and gender differences. By quantitatively analysing the accuracy and qualitatively exploring the developers' assumptions about their users, we discover that age and gender differences seem not to have a major influence on the accuracy, but they do seem to be vital factors that developers take into account when estimating their users demographics, characteristics and behaviour.

1 INTRODUCTION

When designing and developing software, one of the common measures for success is the amount and satisfaction of users. Therefore, it is not surprising that one of the urging questions of developers seems to be how users use their software ("How do users typically use my application?" [1]). The developer-user-interaction and its effects on the success of the project has been addressed by Leonard-Barton and Sinha [8] which argue that a higher user involvement leads to higher user satisfaction. Nonetheless, Gallivan and Keil [4] showed with their case study that the extent of user participation in the development process does not generally predict the success but instead the quality of involvement of users influences it. Assuming that the quality of developer-user-interaction influences how well developers understand their users, the developers' ability to assess the users behavior may influence the project's success. Therefore, we propose to compare the developers' assessment of their users' behavior and the actual behavior of the users. Of the several factors that may affect the accuracy, we measure whether the gender and the age difference have an influence on the accuracy. With this approach, it can be tested whether it is harder for developers to assess the behavior of users with very different demographics than themselves.

We therefore formulated our research questions as follows:

- RQ1: How accurate can developers estimate how users will use their software?
- RQ2: Does the combination of the developer's gender and the user's gender and the age difference between the developers and users influence the accuracy?
- RQ3: What assumption do developers' have about their users?

Our results showed that the average accuracy of the developers estimations is around 50% while it seems to be higher for projects where developers have diverse assumptions. Although, we could not find an influence of gender differences on the accuracy of estimations, age seemed to have a slight effect on the accuracy. The accuracy generally seems to be lower for higher age differences. From our interviews we found that developers seemed to have clear images of their users, although variation can be found between the images. Apart from that the developers generally expressed awareness of the variety of users and as such their behaviour.

In the following, we discuss our approach for the recruiting and the study procedure to address our research questions. Afterwards, we explain the data analysis, summarize the results of our experiment and finally, discuss them. We end this paper with addressing the limitations of our study, drawing a conclusion and suggesting future research.

2 APPROACH & EXPERIMENT DESIGN

2.1 Selection & Recruiting

To conduct our experiment, two different types of inclusion criteria were needed. Firstly, participants with the role developers working on the same piece of software were chosen. Secondly, depending on the project, we needed participants with the role user who either were novice or experienced users of the software. For both roles, the range of age should be as high as possible and the amount of each gender as balanced as possible. This contrasts with the assumption of Ko, Latoza and Burnett [7], as this experiment is not about different skill levels but rather different cognitive strategies and behaviour so that age and gender are worthwhile to consider in the recruitment of participants. A nuisance factor could be the company setup of the developers' workplace (e.g., how projects are managed or if they collaborate with user experience designers). All participants needed to be fluent in English. For this pilot study, we wanted to recruit roughly 5 participants for every role.

It became apparent that recruiting a sufficient amount of developers for one project was not feasible. We therefore recruited developers of two distinct projects. The goal of the application of project A is to allow users to focus on their work without being distracted. As such the software is meant to disable notifications of different communication channels, i.e. WhatsApp or Gmail, for a specific period of time. For this project, we chose participants with role user which were not familiar with the application. Furthermore, we recruited developers of an online shop which is mainly selling electronics (project B). The users of project B were recruited to be familiar with the online shop.

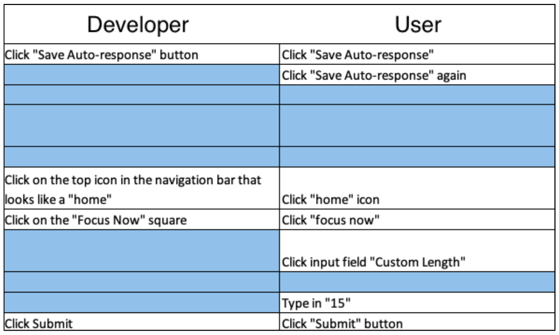
In sum, we were able to recruit 2 developers (D1, D2) and 3 users (U3, U4, U5) for project A, as well as 3 developers (D6, D7, D10) and 5 users (U8, U9, U11, U12, U13) for project B. All of the data was used for the analysis, except for one developer of project B. The data of D6 for the quantitative analysis was lost due to some technical difficulties. Additionally, the CW and the interview session with D7 could not be done on the same day due to connection problems. The interview was held a day after the CW has been completed by the participant. Although the interview covers the mental model of a user while doing the CW, we think that the time period (one day) is short enough for the participant to have sufficient recollection.

2.2 Procedure & Task

Each developer participant received an introduction and an empty cognitive walkthrough (CW). The walkthrough included a description of the scenario and the use case. The scenarios for the projects were created to result in at least 10 actions. We tried to choose tasks that might not have an obvious way to complete and suited

[*] lina.witzel@uzh.ch

[†] sophy.chhong@uzh.ch



Developer	User
Click "Save Auto-response" button	Click "Save Auto-response"
🟦	Click "Save Auto-response" again
🟦	🟦
🟦	🟦
🟦	🟦
Click on the top icon in the navigation bar that looks like a "home"	Click "home" icon
Click on the "Focus Now" square	Click "focus now"
🟦	Click input field "Custom Length"
🟦	🟦
🟦	Type in "15"
Click Submit	Click "Submit" button

Figure 1: Snippet of a sequence of tasks for a developer (left) and user (right) pair for project A (the same color means the step is identical, different colors account for different steps).

the experience of the users (novice or experienced). After conducting the experiments with the developers from project A, we realised that we did not explicitly tell them that the user has no experience with the software. Since the software is quite new, we assumed that it was clear. We considered this for project B and told the developers that the user is already familiar with the software. The participants needed to fill in the steps they think the average user will take to fulfill the requested use case. Participants should only define and not evaluate the steps, like it is usually done in the CW. After completing the CW, a qualitative interview was conducted to gather information about the user whom the participants have imagined, including demographics like age and gender. Apart from that they received guiding questions (e.g. about their character and cognitive strategy) to describe the imagined user. At the end, they needed to tell us their age, gender, educational level, job title, job role and years of professional development experience.

As for the user, an initial introduction guided them on the experiment. They received the same scenario as the developers, but instead of doing a CW, the user participant interacted with the software and tried to complete the task. The user's screen was captured for analysis purposes. After finishing the task, they needed to tell us their age, gender, educational level, job title, job role and professional development experience. Neither participants with role developer nor role user received any training except for the introduction on what they were asked to do.

3 Data Analysis

3.1 Quantitative

For every developer-user, developer-developer and user-user pair, the accuracy was calculated. The first step to do so, was the qualitative comparison and classification of the steps, retaining the sequence of steps (Figure 1). Based on that, the sum of identical and different tasks for each of the pairs could be calculated.

The following formula was used to calculate the accuracy:

$$\text{Accuracy} = \text{Identical tasks / Sum of tasks}$$

To test the robustness of our approach, both researchers ran the analysis for project A separately to calculate the inter-rater agreement as mentioned by Goodwin [5]. With an agreement of 96.74% we decided to rely on the calculations of one researcher, while having the second researcher check the classification in order to ensure a congruence between the results for both projects while minimizing any bias.

The age difference was calculated by subtracting the lower from the higher age for each pair. Furthermore, the correspondence of the gender was indicated by 1 (same gender) and 2 (different gender). To compare the influence of age together with gender difference, in this pilot study, both numbers were multiplied.

3.2 Qualitative

The interviews were fully transcribed to allow open coding. Open coding can be used for thematic content analysis [3] which helped us understand the existing patterns when developers envision their users' behaviour. The codes were used to compare the similarities and differences across the different developers. We did, for example, classify ”male” and ”man” in the category ”male” whereas ”risky” and ”man” ended up into distinct categories. With the codes, we also wanted to highlight peculiarities that we found. Although fourteen stages of analysis are described [3], we ignored stage 10, since there was no need to cut out sections and paste them to paper, and stage 11 as we saw no need to involve participants in the categorisation. Open coding was done by one researcher with the other complementing and checking the category list in an adapted stage 6.

4 Results

4.1 Quantitative

On average, we collected 24.2 steps per project and per participant, with a median of 26 steps for project A and 23 steps for project B. The average accuracy of the developer-user pairs for project A is 47.6% while it lies at 55.1% for project B (Table 1). For project A, the highest average accuracy occurs for the developer-developer pair (64.0%) while the average accuracy for the user-user pairs is much lower (44.3%). This may point to a generally higher variance of the steps the users take, while the developers' assumptions seem to be more similar. For project B, on the other hand, this does not seem to be the case.

Pairs	Average accuracy [%]
Project A	
All pairs	48.3
Developer-user pairs	47.6
Developer-developer pairs	64.0*
User-user pairs	44.3
Project B	
All pairs	51.2
Developer-user pairs	55.1
Developer-developer pairs	45.8*
User-user pairs	48.5

*due to sample size only one pair considered

Table 1: Average accuracy per type of pairs and project

When comparing the accuracy by age difference, it can be seen that for project A the accuracy of all pairs generally is higher for the pairs with less age difference (Figure 2). For project B this trend can be seen as well, although, it is less pronounced.

Both developers of project A achieve a higher average accuracy for users with a different gender. For project B, one developer achieves the same accuracy for both categories while the second developer generally achieves a better accuracy for users with the same gender (Table 2).

When combining the age and gender difference, the trend for project A (Figure 3) looks similar to the one in the age difference (Figure 2), but slightly more pronounced. On the other hand, as seen in Figure 3, the trend for project B generally stays stable independent of the age and gender differences.

4.2 Qualitative

All developers estimated the user to be of similar age (mostly around 30 or 35 years). D1 argued that ”... *older people might not even know that such apps exist so they will not search* [for it]”. D2

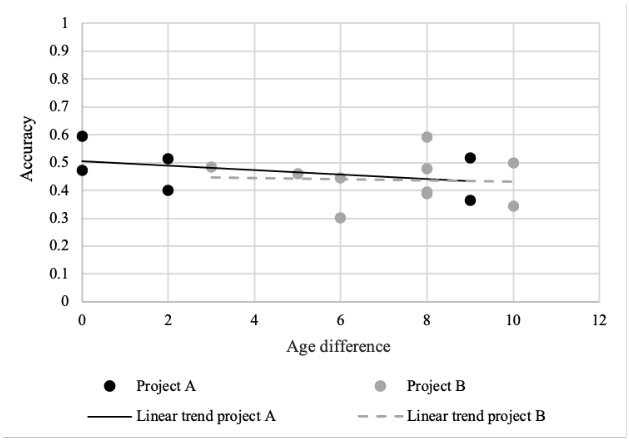


Figure 2: The difference of age and accuracy for each developer-user pair of both projects, including the linear trend for each project

Gender difference	Average accuracy per developer [%]	
Project A	D1	D2
Same gender	51.3*	41.7
Different gender	55.5	52.0*
Project B	D7	D10
Same gender	43.6	48.8
Different gender	43.6	37.4

*due to sample size only one pair considered

Table 2: Average accuracy per project and developer for users with the same or a different gender

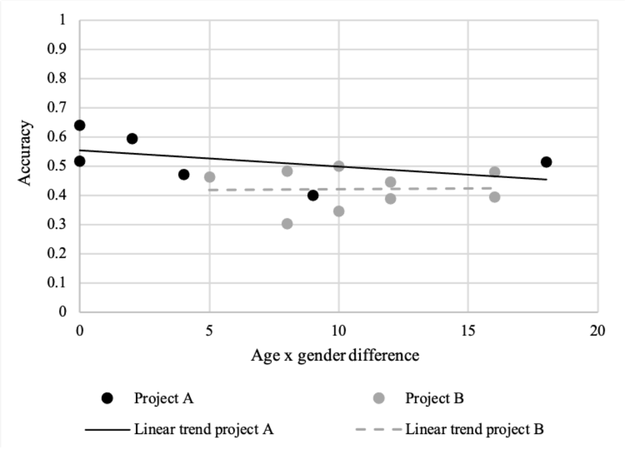


Figure 3: The difference of age x gender difference and accuracy for each developer-user pair of both projects, including the linear trend line for each project

on the other hand mentioned software developers and information workers as the target audience: ”*To where they have had enough time in the company to be promoted and well needed, so they receive a lot of messages and need time to focus.*”. D7 and D10 stated that their online shop mainly sells electronic consumer goods and therefore mostly targets a younger audience. When it comes to gender, D2 was the only developer who assumed the user to be male. The explanation was that the application is targeted towards software engineers where D2 thinks of as a mostly male domain. Furthermore, part of the task was to add a pre-defined automatic response, for which D2 interpreted it as a male signature message. Although the rest of the developers thought of no gender, both D1 and D6 used the pronoun ”he” to refer the user in some parts of the interview.

In terms of the risk attitude of the user, D1 stated that the imagined user is not risk oriented but might explore the application first before completing any task. D2 said the opposite, the user would be risk oriented in the sense to test new applications to figure out if it might be useful. D6 and D10 imagined the user to be risky, D10 stated: *”I think uh creative people do work a lot with computers therefore are also keen to learn new things with computers”*. D7 did not imagine the user to be risky and that the user would simply leave the website if the desired product could not be found.

All developers imagined the user to have medium to high computer self-efficacy and to use technology a lot with the exception of D7 and D10. D7 thought that the specific scenario is more task or goal oriented and as such does not encourage exploration. D10 did not have an answer to the question about the user’s motivation but later stated that the imagined user is creative and therefore would use the computer a lot.

Interestingly, D7 and D10 of project B mentioned Personas and patterns that they use in their company. As such they were considering the variety of users and the different demographic groups when doing the CW. D7 concluded that the scenario would fit a technical person whereas D10 tried to follow the most common user pattern to define the actions. D1 and D2 mentioned no formal method to differentiate between user groups but were nonetheless aware that users might take different paths to complete the tasks.

5 Discussion

With regard to our small sample for this pilot study, we were still able to generate insightful results. We discovered that the overall accuracy of the developers’ estimated steps and the actual steps the users take can differ greatly. The similarity in the developers’ estimations and the variance of the users’ steps for project A could therefore result in less coverage of user patterns and behaviour and hence lower accuracy. This seemed to be more accurately addressed in project B, showing more variance for the developers estimations but higher average accuracy for the developer-user pairs.

For project A, a slight influence of the age difference on the accuracy could be seen while for project B, the difference in gender seemed to have an influence on the accuracy. Both factors - age and gender - seemed to be something that the developers intuitively took into account when estimating their users’ behaviour. Based on these factors, they even concluded on other characteristics of the user (e.g. familiarity with technology). The developers of project A seemed to have a clear picture of their intended user while still being aware of the variance in their users’ behaviour. Similarly, for project B the developers mentioned Personas and common user behaviour patterns they use in the company and as such considered different user groups and demographics.

We discovered a higher variance in estimations for the developers of project B than for project A, quantitatively as well as qualitatively. The target audience of the applications could have played a role. Whereas project A seems to focus on information workers, project B targets a wide audience with their online shop. As mentioned before, Personas may have helped to develop a better understanding for different user demographics, characteristics and cognitive strategies. Nonetheless, the average accuracy for the developer-user pairs was very similar for both projects.

We assumed that developers will most likely imagine the user to be male since people usually link males to the genderless word "user" [2]. From our interviews, most developers did not associate a specific gender with the user. Although, two of those participants later on used the pronoun "he" to refer the user.

6 Limitations

Even though our study revealed interesting results, these results are limited by our small sample size. As the range of age is small and the genders not appropriately balanced, our findings could change, especially when including participants from very different age groups. Furthermore, our findings are limited to two projects. Already for those projects, one can see that the findings vary depending on the company or software. To be able to generalize the results, more projects and bigger samples would need to be included.

Although we tried to minimize the bias in the qualitative analysis, it is worth to mention that the results might slightly vary depending on the involved researchers.

7 Conclusion & Future Research

Generally, our research provided us with interesting insights regarding the accuracy and assumptions that the developers used to estimate their users' behaviour. We showed that there seems to be a gap between the developers' estimations about their users' behaviour and their actual behaviour. Furthermore, we discovered a variance of accuracy which may even be influenced by age and gender differences. While developers are aware of the variance, they seem to have clear and similar assumptions about their users. This conflicts with the divergence of the users behaviour. Measures like working with Personas may enhance the awareness for different user demographics, characteristics and cognitive strategies. As such, they can increase the overall accuracy of developers' estimations.

For our study, we focused on gender and age differences and did not take into account whether the assumptions the developers have about their users (e.g. computer self-efficacy) result in higher accuracy if they are correct. In further experiments, we would want to explore whether the user experience improves for users whom the assumptions fit for. This could be achieved by adding an interview or survey to the users' experiment in order to assess the behaviour and characteristics of the users.

Furthermore, it would be interesting to explore which measures (e.g. using Personas) developers take into account in order to achieve a better understanding of their users behaviour and thus result in higher accuracy when estimating the user's behaviour.

Future research could focus on whether a high accuracy predicts success of a project, similar to Tesch, Sobol, Klein and Jiang [10], and if so, find tools that enhance the accuracy, e.g. by improving the communication between users, designers and developers. A similar approach was used by Guzman and Maalej [6] or Oh, Kim, U. Lee, J. G. Lee and Song [9] studying the effect of app reviews on application development.

8 Acknowledgements

We want to especially thank all the participants for their time and interest to participate in our study. Furthermore, we are greatly thankful for the continuous and insightful feedback from Prof. Thomas Fritz.

References

[1] A. Begel and T. Zimmermann. Analyze this! 145 questions for data scientists in software engineering. In *Proceedings of the 36th International Conference on Software Engineering*, pages 12–23, 2014.

[2] A. Bradley, C. MacArthur, M. Hancock, and S. Carpendale. Gendered or neutral? considering the language of hci. In *Proceedings of the 41st graphics interface conference*, pages 163–170, 2015.

[3] P. Burnard. A method of analysing interview transcripts in qualitative research. *Nurse education today*, 11(6):461–466, 1991.

[4] M. J. Gallivan and M. Keil. The user–developer communication process: a critical case study. *Information Systems Journal*, 13(1):37–68, 2003.

[5] L. D. Goodwin. Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, 5(1):13–34, 2001.

[6] E. Guzman and W. Maalej. How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)*, pages 153–162. IEEE, 2014.

[7] A. J. Ko, T. D. Latoza, and M. M. Burnett. A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering*, 20(1):110–141, 2015.

[8] D. Leonard-Barton and D. K. Sinha. Developer-user interaction and user satisfaction in internal technology transfer. *Academy of Management Journal*, 2017.

[9] J. Oh, D. Kim, U. Lee, J.-G. Lee, and J. Song. Facilitating developer-user interactions with mobile app review digests. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1809–1814. 2013.

[10] D. Tesch, M. G. Sobol, G. Klein, and J. J. Jiang. User and developer common knowledge: Effect on the success of information system development projects. *International Journal of Project Management*, 27(7):657–664, 2009.

APPENDIX

A - Cognitive Walkthrough Template

INSTRUCTIONS

What is a scenario?
The scenario explains in what situation the user is in (context), how experienced the user is (skills) and what the user is trying to achieve (goal). You can find the scenario below the instructions.

What is a subgoal?
In order to achieve the goal described in the scenario, the user will form certain subgoals, i.e. steps/tasks the user needs to execute. First, use the attached form to describe the subgoals you think the user will set for himself.
Examples for subgoals are: "Log in", "Filter search results", "Find latest news article", etc.

What is an action?
The subgoals consist of the specific actions a user needs to perform, in order to achieve the subgoals and consequently the goal. Try to think of the smallest action of the user. For each subgoal use a form to specify the actions.
Examples for actions that you would fill in the form for the subgoal "Log in" are: "Click on first input field ("Username")", "Type in username", "Click on second input field ("Password")", "Type in password", "Click on button ("Sign in")"

Leave any subgoal or action empty if you don't need it.

SCENARIO

[description according to company and use case]

SUBGOALS

Subgoal #1: ______
Subgoal #2: ______
...
Subgoal #11: ______

ACTIONS FOR SUBGOAL [#]

Action #1: ______
Action #2: ______
...
Action #11: ______

B - Interview Guide

- How old would you estimate your average user to be? What are the reasons that you think of that age?
- What gender did you think of when you estimated the user's behavior? Is there a reason why you thought of a male/female?
- What else did you take into account when you estimated the user's behavior?
- What did you think about the motivations of the user to use technology? (E.g. did you think of a user that regularly uses technology – also for fun – or a user that rather infrequently uses it – mainly to accomplish tasks?)
- What did you think about the computer self-efficacy of the user? (E.g. did you think of a user that has high confidence when doing unfamiliar tasks or a user with low confidence?)
- What did you think about the attitude towards risk of the user? (E.g. did you think of a user that likes to try out features with unknown outcomes or one that would be hesitant?)

C - Demographics

- Age
- Gender: male, female, other (specify), don't want to answer
- Educational level
- Job title
- Job role
- Professional (development) experience in years

D - Open Coding Categories

- Age
  - Younger
  - Older
- Gender
  - Gender independent
  - Male
  - Pronoun confusion
  - Stereotyping
- Target audience
- Task difficulty
- Design
  - Improvement of UI
  - Discoverability
  - Usability
- Motivations
  - Frequent technology use
  - Other computer uses
  - Discomfort
- Computer self-efficacy
  - High computer self-efficacy
  - Medium computer self-efficacy
- Risk attitude
  - Risk-averse
  - Risky
- Behaviours
  - Bias
    - Remove bias
    - Positive bias
  - Personas
  - User patterns
  - Experienced user
  - Technical user
  - Path variance
  - Curiosity

E - Quantitative Data Overview

Pair	Gender difference	Age difference	Gender x age difference	Accuracy (%)
Project A				
D1-D2	2	0	0	64.0
D1-U3	2	0	0	59.2
D1-U4	1	2	2	51.2
D1-U5	2	9	18	51.6
D2-U3	1	0	0	47.0
D2-U4	2	2	4	40.0
D2-U5	1	9	9	36.4
U3-U4	2	2	4	43.8
U3-U5	1	9	9	50.0
U4-U5	2	7	17	39.1
Project B				
D7-U8	1	5	5	46.2
D7-U9	1	10	10	50.0
D7-U10	1	2	2	45.8
D7-U11	2	8	16	47.8
D7-U12	2	8	16	39.3
D7-U13	1	10	10	34.5
U8-U9	1	5	5	70.4
U8-D10	1	3	3	48.3
U8-U11	2	3	6	82.6
U8-U12	2	3	6	46.9
U8-U13	1	5	5	64.3
U9-D10	1	8	8	38.7
U9-U11	2	2	4	82.6
U9-U12	2	2	4	51.6
U9-U13	1	0	0	58.6
D10-U11	2	6	12	44.4
D10-U12	2	6	12	30.3
D10-U13	1	8	8	59.3
U11-U12	1	0	0	48.3
U11-U13	2	2	4	68.0
U12-U13	2	2	4	62.1

Table 3: Overview including gender difference, age difference, gender x age difference and accuracy for each pair of each project