Impact of Education and Experience Level on the Effectiveness of Exploratory Testing: An Industrial Case Study

Ceren Şahin Gebizli
Vestel Electronics,
Manisa, Turkey
ceren.sahin@vestel.com.tr

Hasan Sözer
Ozyegin University,
Istanbul, Turkey
hasan.sozer@ozyegin.edu.tr

***Abstract*—Exploratory Testing (ET) is a widely applied approach in practice. This approach relies on the knowledge and experience of test engineers and technicians who perform ET. Hence, we aimed at evaluating the impact of education and experience level on the effectiveness of ET. We conducted an industrial case study for this purpose. 19 practitioners, who have different educational backgrounds and experience levels, were involved in applying ET for testing a Digital TV system. We measured the number of detected failures and categorized these failures based on their severity. We evaluated effectiveness from two aspects: criticality of the detected failures and efficiency in terms of number of failures detected per unit of time. The results show that the efficiency of ET is significantly affected by both the educational background and experience. When it comes to the number of critical failures detected, we cannot observe any impact of education on results. However, experience level has a significant impact for this aspect as well.**

***Keywords*-software testing, industrial case study, exploratory testing, testing effectiveness, testing experience, testing education**

I. Introduction

Exploratory Testing (ET) [1], [2], [3] adopts a continuous learning and adaptation process for testing. Hereby, the tester iteratively learns about the product and its faults, plans the testing activities, designs and executes the tests [4]. Unlike traditional test case based testing, ET is not based on a set of predesigned test cases. Instead, tests are dynamically designed, executed and modified by testers. Hence, test design, execution and learning are all concurrent activities. Among other existing definitions [5], [6], Bach briefly defines ET as *"simultaneous learning, test design and test execution"* [7].

ET does not utilize formal descriptions or detailed methodologies. Testing activities are performed manually and testers do not strictly follow any procedures during these activities. As a result of these facts, one might consider ET as an ad-hoc approach. Nevertheless, it is known to be one of the mostly applied and one of the most effective approaches [8], [9], [10], [11] in terms of revealing failures. This recognition is also aligned with our own observations in the industry [12], [13]. We have seen that test models that are refined based on ET activities were more effective in terms of finding failures.

ET aims at exploiting human effort efficiently by utilizing intuition, experience and knowledge. This experience and knowledge can be related to the application domain, system (i.e., specific to the tested product) or background (e.g., general software engineering and testing knowledge) [14]. In particular, ET activities are supposed to be performed by testers who have both technical knowledge and accumulated unwritten knowledge on where failures most likely exist [4]. Therefore, it is believed that ET is largely dependent on the skills and experience of the tester [11]. However, to the best of our knowledge, the impact of education and experience level on the effectiveness of ET has not been formally evaluated before.

In this paper, we report a case study that is performed in an industrial context. 19 practitioners, who have different education and experience levels, were involved in applying ET for testing a Digital TV system. We measured the number of detected failures and categorized these failures based on their severity. We also measured the time spent by each subject for performing the tests. Then, we compared different groups of subjects. These groups are formed based on experience and education levels. Comparisons are performed with respect to two criteria: *i)* the number of detected failures that are critical, and *ii)* efficiency measured as the number of failures detected per unit of testing time.

Results show that the efficiency of ET is significantly affected by both the educational background and experience levels. Experience level has a significant impact on the number of detected critical failures as well. However, we can not observe an impact of education on the number of critical failures detected.

The remainder of this paper is organized as follows. In the following section, we summarize related work. In Section III, we explain our case study and we discuss the results. Finally, we conclude the paper with a discussion of future work in Section IV.

II. Related Work

Although ET is commonly applied in practice, scientific research and empirical studies on ET are scarce, especially when we consider those that are conducted in an industrial context [15], [11]. In the following, we summarize existing publications on related experimental studies.

978-1-5090-6676-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICSTW.2017.8

There have been experimental studies on testing techniques in general [16]. An evaluation of these techniques [11] suggests that skills and experience of the tester turn out to be an important factor for testing effectiveness even in test case based testing. Similarly, the importance of both testing knowledge and domain knowledge was confirmed by industrial case studies before [17]. In another study [18], the impact of experience was evaluated for the effectiveness of test case design. Hereby, results show that neither the experienced nor the inexperienced testers performed better with respect to each other in all aspects. The two groups had both relative strengths and weaknesses with respect to different aspects.

An experimental study [19] for comparing ET and test case based testing revealed that these two approaches do not differ in terms of failure detection effectiveness. However, they have both pros and cons with respect to efficiency, the number of false positives and management overhead.

A controlled experiment [20] was conducted to compare the failure detection effectiveness of model based testing and model-based exploratory testing. Results show that the overall test effectiveness was improved though the two approaches detect different types of failures. We have also conducted case studies before for evaluating the effectiveness of model-based testing, when it is supported by ET [12]. Results show that the test effectiveness can be significantly improved when the test models are refined based on execution traces collected during ET activities.

An empirical study [21] that focuses particularly on the effectiveness of ET studied the impact of personality traits as a factor. Results show that testers having extrovert personality might be more likely to be good at ET. In this study, we evaluate the impact of education and experience level on the failure detection effectiveness of ET.

III. Industrial Case Study and Discussion

Our case study is performed for testing Smart TV systems developed by Vestel[1], which is one of the largest TV manufacturers in Europe. Smart TV systems are highly cost sensitive and they are subject to short development time periods. The market is highly competitive and end users are less tolerant to failures [22]. Hence, effective testing methods are essential. ET is known to be one of the methods that has been applied for years [12], [13]. It has been applied by different employees in the company over time. In this study we aimed at evaluating the impact of the educational backgrounds and experience levels of these employees on the effectiveness of ET. Hereby, we differentiate experience regarding the domain or the system under test from the experience in testing activities in general. In terms of educational background, we differentiate between those who have higher education (college or university) on a relevant subject (Computer Science/Engineering or Software Engineering) and those who do not have (high school graduates or graduates of two-year educational programs). Testers who have higher education and do not have higher education participated in software testing training which was given internally in the company for 2 days.

[1]http://www.vestel.com.tr

A. Research Questions

We evaluate the effectiveness of ET from two aspects. First, we consider test efficiency based on the effort and the number of detected failures. Second, we consider how critical the detected failures are. Therefore, we defined the following research questions.

RQ1: How domain and testing experiences are affecting the test efficiency in terms of number of failures detected per unit of time?

RQ2: How domain and testing experiences are affecting the number of critical failures detected?

RQ3: How educational background is affecting the test efficiency in terms of number of failures detected per unit of time?

RQ4: How educational background is affecting the number of critical failures detected?

Accordingly, we defined the following 4 hypothesis:

- H_o^1 The level of domain experience and testing experience do not have any effect on the ET efficiency.
- H_o^2 The level of domain experience and testing experience do not have any effect on the the criticality of detected failures during ET.
- H_o^3 The level of education does not have any effect on the ET efficiency.
- H_o^4 The level of education does not have any effect on the the criticality of detected failures during ET.

. In the following, we explain the experimental setup we used for addressing the research questions.

B. Experimental Setup

There is a dedicated software testing group in the company who is performing tests of different consumer electronics products such as Digital TVs, refrigerators, washing machines, dishwashers, air conditioners, cookers and smart phones. Most of the tests are automated but there are also manual tests being performed and we focused on such tests in this study. The testing group is composed of either test engineers who studied at a college/university or test technicians who completed two-year educational programs. We refer to both test engineers and test technicians as practitioners in the rest of the paper.

In total, 19 practitioners were involved in the case study as subjects. These practitioners were instructed to apply ET for a particular feature of a real Digital TV system. This feature was explained to all the subjects for 15 minutes before the case study. Some of the subjects have already had domain knowledge, i.e., they have been previously working on testing Digital TVs. Some other subjects lacked this knowledge, i.e., they were involved in the testing of products other than Digital TVs.

The list of all the subjects are provided in Table I. For each subject, 3 properties are listed in the 2^{nd}, 3^{rd} and 4^{th}

columns, respectively. These properties also define the factors we consider in the case study:

- *Domain Experience*: measured in ratio scale in terms of the number of years.
- *Testing Experience*: measured in ratio scale in terms of the number of years.
- *Higher Education*: measured in nominal scale, at two levels: exists, not exists.

In Table I, it can be seen that, some practitioners have more domain experiences than testing experiences. These practitioners are testers who do not have higher education and they had worked for different products other than Digital TVs.

Practitioner ID (PId)	Domain Experience (# of years)	Testing Experience (# of years)	Higher Education (Yes/No)
1	11	11	No
2	8	8	No
3	16	14	No
4	7	5	No
5	7	7	No
6	11	10	No
7	8	10	No
8	1	1	Yes
9	6	6	Yes
10	1	8	Yes
11	1	1	Yes
12	1	1	Yes
13	1	1	Yes
14	1	1	Yes
15	4	4	Yes
16	4	4	Yes
17	12	12	Yes
18	0	3	Yes
19	0	2	Yes

TABLE I: The whole list of subjects.

We asked all the participants to perform ET on the same system. They were just observed without interference throughout this process. The system under test was also not altered (no bug fixes) throughout the study. We measured/calculated the following variables:

- *Test duration*: measured for each subject in ratio scale in terms of the number of days.
- *Number of failures detected*: measured for each subject in absolute scale for each of the failure categories listed as *Critical*, *Major*, *Minor* and *Trivial*.
- *Efficiency*: calculated for each subject as the ratio of the total number of failures detected and test duration.

In the following, we present and discuss the results.

C. Results and Discussion

The overall results are summarized in Table II. The first column lists the ID of each subject just like in Table I. The second column lists the test duration. For instance, we can observe that $PId-18$ and $PId-19$ (Practitioner ID 18 and 19) completed the test in 16 days, where $PId-9$ and $PId-17$ completed the test in 6 days. The last column lists the number of failures detected. This list is provided separately for the 4 different failure categories; *Critical*, *Major*, *Minor* and *Trivial*. For instance, we can see that $PId-2$ detected the maximum total number of failures (12); however $PId-10$ could only find 5 failures.

Practitioner ID (PId)	Test Duration (# of days)	# of Failures Detected			
		Critical	Major	Minor	Trivial
1	10	2	2	5	1
2	10	2	2	6	2
3	9	2	2	5	0
4	9	2	2	5	0
5	10	2	2	4	0
6	10	2	2	5	0
7	10	2	2	5	0
8	15	2	2	5	2
9	6	2	2	6	0
10	10	1	2	2	0
11	15	1	2	5	1
12	14	1	2	5	0
13	15	1	2	5	1
14	12	0	2	5	1
15	8	2	2	6	0
16	9	2	2	6	0
17	6	2	2	6	0
18	16	1	2	4	2
19	16	1	2	4	2

TABLE II: The list of overall results.

We compared different groups of subjects with respect to efficiency and the number of detected failures that are of type *Critical*. These groups are formed based on experience and education levels. Education level is already provided in nominal scale and as such it is trivial to separate the two groups. To be able to separate subjects with respect to experience, we set a threshold for the number of years of experience as 2. Subjects, who have at least 2 years of (both domain and testing) experience are considered *experienced*, while the others are considered *inexperienced*.

In the following, we list the results for different groupings to answer the 4 research questions.

1) Impact of domain and testing experience on test efficiency: To be able to evaluate the impact of experience, we have separated the results into two; *i)* those for subjects who do not have higher education, and *ii)* those for subjects who have higher education. These results are listed in Table III and IV, respectively.

When we look at the subjects that do not have higher education (Table III), we see that they all have high experience levels. Therefore, we could not evaluate the impact of experience for this group. However, experience levels vary for the subjects who have higher education (See Table IV). We can also observe that the efficiency is higher for those subjects that have experience. We performed a t-test to validate this observation. We formed two groups, Group A and Group B from the

Practitioner ID (PId)	Domain Experience	Testing Experience	#Critical Failures	Efficiency
1	11	11	2	1.00
2	8	8	2	1.20
3	16	14	2	1.00
4	7	5	2	1.00
5	7	7	2	0.80
6	11	10	2	0.90
7	8	10	2	0.90

TABLE III: Results for subjects who do not have higher education.

Practitioner ID (PId)	Domain Experience	Testing Experience	#Critical Failures	Efficiency
8	1	1	2	0.73
9	6	6	2	1.67
10	1	8	1	0.50
11	1	1	1	0.60
12	1	1	1	0.57
13	1	1	1	0.60
14	1	1	0	0.67
15	4	4	2	1.25
16	4	4	2	1.11
17	12	12	2	1.67
18	0	3	1	0.56
19	0	2	1	0.56

TABLE IV: Results for subjects who have higher education.

list of subjects in Table IV. Group A consists of experienced subjects (who have 2 or more years of experience), whereas Group B consists of inexperienced subjects (who have less than 2 years of experience). We compared the efficiency for these groups. Results[2] are listed in Table V.

	Group A	Group B
Mean	1.423	0.599
Variance	0.081	0.005
Observations	4	8
Hypothesized Mean Difference	0	
df	3	
t Stat	5.668	
P(T¡=t) one-tail	0.005	
t Critical one-tail	2.353	
P(T¡=t) two-tail	0.01	
t Critical two-tail	3.182	

TABLE V: T-test results regarding the comparison of experienced (Group A) and inexperienced (Group B) groups by means of test efficiency.

We can see that P(T¡=t) one-tail and P(T¡=t) two-tail values are very low. They are well below the commonly accepted threshold (0.05) [23], which means that the difference is significant. That also means that the null hypothesis H_o^1 can be rejected and the impact of experience on ET efficiency can be confirmed. We also evaluated the impact of experience on the criticality of the detected failures. This is discussed in the following.

2) Impact of domain and testing experience on the criticality of the detected failures: We can observe in Table IV that experienced subjects detected more failures of type *Critical*. We performed a separate t-test to evaluate the significance of this difference. We performed the comparison between the same groups, Group A and Group B as formed in the previous test.

The results are listed in Table VI. Again, the P values turn out to be well below 0.05, which points out the significance of the difference. Hence, we conclude that the null hypothesis H_o^2 can be rejected. The level of experience has a significant impact on the criticality of failures detected during ET.

In fact, we can observe from Table III that all the subjects detected all the 2 critical failures. Recall that all of these subjects have high experience though they do not have higher education. Their efficiency (0.97) is also higher than those listed in Table IV (0.87) on average. We evaluate the impact of education in more detail in the following.

	Group A	Group B
Mean	2	1
Variance	0	0.285
Observations	4	8
Hypothesized Mean Difference	0	
df	7	
t Stat	5.291	
P(T¡=t) one-tail	0.0005	
t Critical one-tail	1.894	
P(T¡=t) two-tail	0.0011	
t Critical two-tail	2.364	

TABLE VI: T-test results regarding the comparison of experienced (Group A) and inexperienced (Group B) groups by means of number of detected critical failures.

3) Impact of higher education on test efficiency: To be able to evaluate the impact of higher education, we have separated the results into two; *i)* those for subjects who have 2 or more years of domain and testing experience, and *ii)* those who have less than 2 years of experience. These results are listed in Table VII and VIII, respectively.

When we look at the inexperienced subjects (Table VIII), we see that they all have higher education. Therefore, we could not evaluate the impact of education for this group. However, education level varies for the subjects who are experienced (See Table VII). On average, the efficiency of those who have higher education (1.42) is more than the efficiency of others (0.97). We also performed a t-test to evaluate the significance of this difference. We formed two groups, Group C and Group D from the list of subjects in Table VII. Group C consists of subjects with higher education, whereas Group D consists of subjects without higher education. We compared the efficiency for these groups. Results are listed in Table IX.

We can see that P(T¡=t) one-tail and P(T¡=t) two-tail values are 0.019 and 0.039, respectively. These values are below 0.05,

[2]We used Microsoft Excel (2010) to obtain the results.

Practitioner ID (PId)	Higher Education	# of Critical Failures	Efficiency
1	No	2	1.00
2	No	2	1.20
3	No	2	1.00
4	No	2	1.00
5	No	2	0.80
6	No	2	0.90
7	No	2	0.90
9	Yes	2	1.67
15	Yes	2	1.25
16	Yes	2	1.11
17	Yes	2	1.67

TABLE VII: Results for subjects who have at least 2 years of experience.

Practitioner ID (PId)	Higher Education	# of Critical Failures	Efficiency
8	Yes	2	0.73
10	Yes	1	0.50
11	Yes	1	0.60
12	Yes	1	0.57
13	Yes	1	0.60
14	Yes	0	0.67
18	Yes	1	0.56
19	Yes	1	0.56

TABLE VIII: Results for subjects who have less than 2 years of experience.

	Group C	Group D
Mean	0.971	1.423
Variance	0.015	0.081
Observations	7	4
Hypothesized Mean Difference	0	
df	4	
t Stat	-2.998	
P(T_i=t) one-tail	0.019	
t Critical one-tail	2.131	
P(T_i=t) two-tail	0.039	
t Critical two-tail	2.776	

TABLE IX: T-test results regarding the comparison of groups who have higher education (Group C) and who do not have higher education (Group D) by means of efficiency.

suggesting that education has a significant impact on efficiency among the experienced subjects. Hence, the null hypothesis H_o^3 can also be rejected although, P values are closer to the threshold in this case. However, we can not conclude the same for the criticality of the detected failures as discussed in the following.

4) Impact of higher education on the criticality of detected failures: We can see in Table VII that all the subjects detected 2 critical faults. Hence, we can not observe any impact of higher education in that respect. There is no difference at all. So, we conclude that higher education does not have an impact on the criticality of detected failures. As such, we have to accept the null hypothesis H_o^4.

In fact, we can also observe from Table III that all the subjects who do not have higher education could find all the critical failures. On the other hand, subjects who do have higher education could not (See Table IV). However, we can not compare these directly since the level of experience is different between the groups.

In the following, we discuss validity threats for our case study.

D. Threats to Validity and Limitations

Our study is subject to an external validity threat [24] since it is based on a single case study. Internal validity threats are mitigated by using a real system and involving real participants from the industry to our study. Conclusion and construct validity threats are mitigated by observing the activities of participants without interfering with them. The number of participants (19) can also lead to conclusion and construct validity threats. However, we performed statistical tests to evaluate the significance of the results.

We also performed Anova analysis [25] to test the significance of differences among different groups of participants. 4 different groups can be considered based on the 2 factors we evaluate in this study. These groups are shown in Table X.

Factors	Higher Education	No Higher Education
Experience	Members of both Group A and C	Members of both Group A and D
No Experience	Members of both Group B and C	Members of both Group B and D

TABLE X: 4 different groups of subjects based on 2 factors: experience and education level.

Hereby, we are missing one of the groups as highlighted in the table since we do not have any subjects in that category. We performed one-way (single factor) Anova analysis on the remaining 3 groups in terms of test efficiency. Table XI lists sum, average and variance values for each group, whereas Table XII lists the analysis results.

	Group A & D	Group A & C	Group B & C
Count	7	4	8
Sum	6.8	5.69	4.79
Average	0.97	1.42	0.59
Variance	0.015	0.081	0.005

TABLE XI: Descriptive statistics regarding the test efficiency of groups listed in Table X.

We can see that the F value is much greater than F-critical value. We can also see that the P-value is much smaller than 0.05. These results also indicate significant difference among the groups of subjects.

The significance of the results can also be observed with the box plot depicted in Figure 1. Hereby, we compare the distributions of test efficiency values for the 3 groups on which

	Between Group	Within Group	Total
SS	1.852	0.376	2.228
df	2	16	18
MS	0.926	0.023	0.59
F	39.4		
P-value	0.0000007		
F crit	3.63		

TABLE XII: Anova analysis results regarding the comparison of groups listed in Table X in terms of test efficiency.

we applied Anova analysis. We can see that error bars (i.e., whiskers) are short and variance within each group is small. This also increases our confidence regarding the significance of the results.

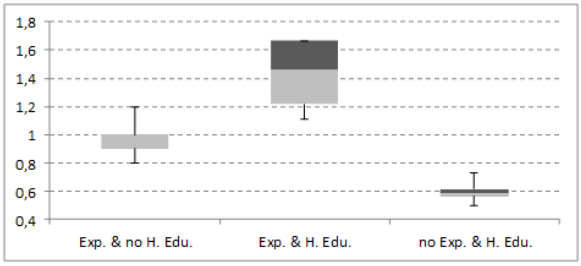

Fig. 1: Box plots regarding the test efficiency of the 3 groups listed in Table X; Exp. & no H. Edu (Group A & D), Exp. & H. Edu. (Group A & C), no Exp. & H. Edu. (Group B & C)

IV. Conclusions and Future Work

We performed an industrial case study to evaluate the impact of education level and experience level on the effectiveness of ET. 19 practitioners, who have different education and experience levels, were involved in applying ET for testing a Digital TV system. The results show that efficiency in terms of the number detected failures per unit of time is significantly affected by both the educational background and experience. Experience level has also a significant impact on the number of detected critical failures, whereas education level has not. As future work, we aim at extending the case study and conducting a controlled experiment to evaluate the impact of different factors on the effectiveness of ET.

Acknowledgment

This work is supported by the joint grant of Vestel Electronics and the Turkish Ministry of Science, Industry and Technology (909.STZ.2015). The contents of this article reflect the ideas and positions of the authors and do not necessarily reflect the ideas or positions of Vestel Electronics and the Turkish Ministry of Science, Industry and Technology. We would like to thank software test engineers and technicians at Vestel Electronics for supporting our case study.

References

[1] G. J. Myers and C. Sandler, *The Art of Software Testing*. John Wiley & Sons, 2004.
[2] W. C. Hetzel and B. Hetzel, *The Complete Guide to Software Testing*, 2nd ed. New York, NY, USA: John Wiley & Sons, Inc., 1991.
[3] J. Itkonen, M. V. Mntyl, and C. Lassenius, "Test better by exploring: Harnessing human skills and knowledge," *IEEE Software*, vol. 33, no. 4, pp. 90–96, 2016.
[4] J. A. Whittaker, *Exploratory Software Testing: Tips, Tricks, Tours, and Techniques to Guide Test Design*, 1st ed. Addison-Wesley Professional, 2009.
[5] C. Kaner, "Exploratory testing," in *Quality Assurance Institute Worldwide Annual Software Testing Conference*, 2006.
[6] A. Tinkham and C. Kaner, "Exploring exploratory testing," in *Proceedings of the Software Testing and Analysis and Review East Conference*, 2003.
[7] J.Bach, "Exploratory testing explained," Tech. Rep., 2003. [Online]. Available: http://www.satisfice.com/articles/et-article.pdf
[8] J. Itkonen, M. V. Mantyla, and C. Lassenius, "Defect detection efficiency: Test case based vs. exploratory testing," in *First International Symposium on Empirical Software Engineering and Measurement*. IEEE Computer Society, 2007, pp. 61–70.
[9] J. Itkonen, "Empirical studies on exploratory software testing," Ph.D. Thesis, Aalto University, 2011.
[10] A. Naseer and M. Zulfiqar, "Investigating exploratory testing in industrial practice : A case study," Master's thesis, Blekinge Institute of Technology, 2010.
[11] W. Afzal, A. Ghazi, J. Itkonen, R. Torkar, A. Andrews, and K. Bhatti, "An experiment on the effectiveness and efficiency of exploratory testing," *Empirical Software Engineering*, vol. 20, no. 3, pp. 844–878, 2015.
[12] C. Gebizli and H. Sozer, "Automated refinement of models for model-based testing using exploratory testing," *Software Quality Journal*, 2016, published online, DOI: 10.1007/s11219-016-9338-2.
[13] H. Sozer and C. Gebizli, "Model-based testing of Digital TVs: An industry-as-laboratory approach," *Software Quality Journal*, pp. 1–18, 2016, published online, DOI: 10.1007/s11219-016-9321-y.
[14] J. Itkonen, M. V. Mantyla, and C. Lassenius, "The role of the testers knowledge in exploratory software testing," *IEEE Transactions on Software Engineering*, vol. 39, no. 5, pp. 707–724, 2013.
[15] J. Itkonen and K. Rautiainen, "Exploratory testing: a multiple case study," in *Proceedings of International Symposium on Empirical Software Engineering*, 2005, pp. 84–93.
[16] N. Juristo, A. Moreno, and S. Vegas, "Reviewing 25 years of testing technique experiments," *Empirical Software Engineering*, vol. 9, no. 1-2, pp. 7–44, 2004.
[17] A. Beer and R. Ramler, "The role of experience in software testing practice," in *Proceedings of the 34th Euromicro Conference Software Engineering and Advanced Applications*, 2008, pp. 258–265.
[18] P. Poon, T. H. Tse, S. Tang, and F. Kuo, "Contributions of tester experience and a checklist guideline to the identification of categories and choices for software testing," *Software Quality Journal*, vol. 19, no. 1, pp. 141–163, 2011.
[19] J. Itkonen and M. Mäntylä, "Are test cases needed? replicated comparison between exploratory and test-case-based software testing," *Empirical Software Engineering*, vol. 19, no. 2, pp. 303–342, 2014.
[20] C. Schaefer and H. Do, "Model-based exploratory testing: A controlled experiment," in *Proceedings of the 2014 IEEE International Conference on Software Testing, Verification, and Validation Workshops*, 2014, pp. 284–293.
[21] L. Shoaib, A. Nadeem, and A. Akbar, "An empirical evaluation of the influence of human personality on exploratory software testing," *Multitopic Conference and IEEE 13th International*, pp. 1–6, 2009.
[22] I. de Visser, "Analyzing user perceived failure severity in consumer electronics products incorporating the user perspective into the development process," Ph.D. dissertation, Eindhoven University of Technology, The Netherlands, 2008.
[23] S. Boslaugh and P. Watters, *Statistics in a Nutshell: A Desktop Quick Reference*, ser. In a Nutshell (O'Reilly). O'Reilly Media, 2008.
[24] C. Wohlin, P. Runeson, M. Host, M. Ohlsson, B. Regnell, and A. Wesslen, *Experimentation in Software Engineering*. Springer-Verlag, 2012.
[25] B. Winer, D. Brown, and K. Michels, *Statistical Principles in Experimental Design*, ser. McGraw-Hill series in psychology. McGraw-Hill, 1991.