Can users recall their user experience with a technology? Temporal bias and the system usability scale.

Kyle Boyd
Ulster University
Belfast, BT15 1ED
ka.boyd@ulster.ac.uk

Raymond Bond
Ulster University
Jordanstown,BT37 0QB
rb.bond@ulster.ac.uk

Justin Magee
Ulster University
Belfast, BT15 1ED
jdm.magee@ulster.ac.uk

Paul McCormack
Ulster University
Belfast, BT15 1ED
p.mccormack@ulster.ac.uk

The System Usability Scale (SUS) score survey is a widely respected tool for measuring usability. While there are other surveys available such as the User Experience Questionnaire (UEQ) or the Single Ease Question (SEQ), the SUS is amongst the most popular and widely used instrument. SUS provides an easy-to-understand score with benchmarking. Generally, a SUS score is administered directly after a usability test to assess the user experience and the usability of a product, including websites and smartphone apps and more. However, some researchers have used it as a survey as part of a 'in the wild' trial which is often completed after the trial or indeed sometime after the subjects interacted with the technology. With this in mind the aim of this research was to see if a participant's user experience would change if a SUS score was administered at different times after a test to understand if recalling the usability of technology led to temporal bias for the SUS.

Usability, System Usability Scale, User Experience, Usability Testing, Human-Computer Interaction, User Interfaces

1. INTRODUCTION

User Experience (UX) as a discipline has evolved considerably over the last number of decades. The introduction of mediums such as desktop, mobile, and web including, native, audio and tactile input means that over time the process of how we conduct UX design has changed. As a discipline, designers design experiences and the aim is to make these experiences better [1]. The UX process is an iterative process of Observation → Idea Generation → Prototyping → Testing. This loop is run through multiple times to unsure assumptions are tested and designs revisited. By trying to understand users better, there has been a drive toward UX research through Usability testing.

Usability testing refers to the process of evaluating a product or service by testing it with representative users [2,3,4]. Typically, during a test, participants will try to complete tasks while observers watch, listen and take notes. The goal of the test is to identify any usability problems, collect qualitative and quantitative data and determine the participant's satisfaction with the product. To run effective usability testing development of a repeatable test protocol, appropriate participant recruitment, analysis and reporting is required.

User testing, often using incomplete or sketch prototypes, permits a process where proposed designs or individual features within a system are forced to fail early, fast and often, in order to refine the most robust user experience or effective functionality. This agile design thinking process identifies problems before a full product is designed and released for end use [5,6]. The earlier the issues are realised the quicker they can be rectified, resulting in less impact on time and cost. Typically, a usability test will assess:

- Effectiveness (the extent to which people can complete their tasks and achieve their goals successfully)
- Efficiency (the extent to which they expend resource in achieving their goals)
- Satisfaction (the level of comfort and/or enjoyment of the experience in achieving those goals)

It is important to collect the right data so that this can be analysed and re-design recommendations can be made. There are various ways to collect the above attributes but one of the most popular and widely used methods is post-test surveys such as the System Usability Scale (SUS).

1.1 SYSTEM USABILITY SCALE

SUS was created by John Brooke [7] in 1986 and allows evaluation of hardware, software, mobile devices, websites and applications. SUS consists of a 10-item questionnaire, each offering a Likert scale (normally 5 point) ranging from strongly agree to strongly disagree. Subsequently, a universal

Proceedings of British HCI 2018.
Belfast, UK

SUS score is computed. The standard SUS consists of the following ten items (odd numbered items are worded positively, even numbered items worded negatively. Questions are as follows:

(I) I think that I would like to use this system frequently.
(II) I found the system unnecessarily complex.
(III) I thought the system was easy to use.
(IV) I think that I would need the support of a technical person to be able to use this system.
(V) I found the various functions in this system were well integrated.
(VI) I thought there was too much inconsistency in this system.
(VII) I would imagine that most people would learn to use this system very quickly.
(VIII) I found the system very cumbersome to use.
(IX) I felt very confident using the system.
(X) I needed to learn a lot of things before i could get going with this system.

Typically, a SUS questionnaire is given after a participant has completed a usability test so they can rate the usability and user experience [8]. However, researchers have been using this questionnaire in various ways, for example they have used SUS after a longitudinal study involving a trial of technology, after a usability test that has specified tasks or even after a session without tasks where a user casually reviews an app or some other technology.

1.2 RESEARCH QUESTIONS

The research questions are as follows:

(I) Does the memory and recollection of a past user experience change over time?

(II) What is the users' memorability of user experience over a three week period.

2. PREVIOUS WORK VALIDITY AND RELIABILITY

Bangor *et al.* [9] conducted usability studies on various products and services using the SUS score. They conducted over 200 studies with 2300 surveys and found that the mean SUS score was 70 and the median was 75. Bangor *et al.* [10] also analysed the interpretation of SUS and added new descriptors. They compared the SUS score and perceived levels of usability. Over 85 was excellent, 70-85 was good to excellent, 50-70 is acceptable but has issues that need addressed and below 50 is unusable and unacceptable.

Tulis and Stetson [11] measured the usability of two websites using a range of different surveys including the Questionnaire for User Interaction Satisfaction (QUIS), SUS and the Computer System Usability Questionnaire (CSUQ). It was found that the SUS provided the most reliable results across a range of samples.

We are interested in recollection and ones reflection of a past User Experience event. Koon *et al*. [12] also explores the utilitarian, hedonic and social aspects of smartphones to measure people continually engage in smartphone activity. We wish to further this work and differentiate by looking at the user experience of an application over time using SUS survey [13,14].

User opinions in the moment and retrospectively are likely to be different. In order to understand if a participant's recollection and memory of a user experience changes over time a suitable protocol for repeatable usability testing was required. This study design was approved by the ethical approval by the Art & Design Research Ethics Committee (Ulster University) on 28th February 2018. The researchers conducted a usability test on a web application and invited participants to complete a SUS score immediately after the test and then over the following two weeks. The latter two time points involved the user completing the SUS survey using their memory of their past user experience. The idea is to measure any recall or temporal bias in completing SUS.

3. METHODOLOGY

The following section outlines the methodology of the study with details of participants and data analysis.

3.1 DATA COLLECTION

Participants were asked to complete a series of tasks (See Table 1) on a Web Application called Virtuagym (http://www.virtuagym.com) a publicly available web application which promotes healthy living (See Figure 1). It was our intention to have rudimentary tasks, that was perceived to have a neutral emotive experience. This was to focus participants to determine design inconsistencies and usability problem areas within the user interface and content areas.

After each participant completed the tasks they completed a SUS survey. Participants were sent another SUS survey via email both one week and two weeks after the test. Once all the SUS questionnaires had been completed the data was

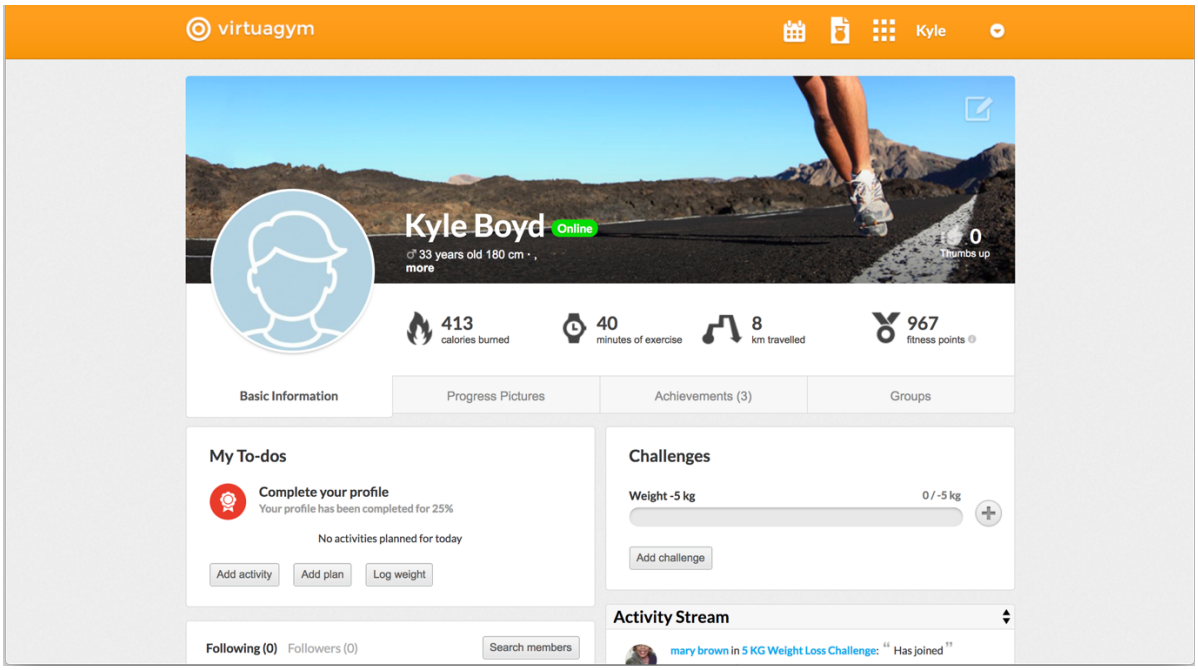

Figure 1: *The Website Virtuagym.com which was used in the study*

Table 1: *The tasks that were completed by participants using virtuagym*

Task Number	Task to Complete
1	Sign up to http://www.virtuagym.com
2	Go through the setup process
3	Add activity calendar and workouts to your portfolio
4	You want to go running each Saturday add a running activity
5	Each Tuesday and Thursday you go to the gym add a gym workout to the calendar
6	You would like to tone arms for the summer. Include a dumbbell weekly workout
7	You would like to raise money for charity – you are going to do 150 sit-ups a day. Add this challenge to your workout
8	Find out how many calories will be burned with this exercise regime?

collated and analysed using R Studio. The findings of the study can be found in the results section.

3.2 PARTICIPANTS

Thirty participants from Ulster University, were chosen to undertake the usability test. The test took place in public buildings in Northern Ireland. Public buildings are chosen specifically as they are required by law to be accessible for those with disabilities ensuring participant inclusivity [15]. This was an evaluative study and therefore no statistical analysis was used to model participant sample size. Within usability testing sample sizes of between 5 and 15 are deemed appropriate, with the 5 yielding 80% of usability issues [16]. The participants were given an information sheet and consent form to prove an opportunity to review the study and ask any questions before the test. Written informed consent was obtained before commencing the study.

The participants were made up of 18 Male and 12 Female. Of those one was aged between 25-34, the remaining subjects were aged 18-24. When the participants were asked to self-evaluate their computer literacy (1 being novice and 5 being expert) 83% responded between 4 and 5. 50% of the participants felt learning a new technology was easy and 93% used technology like smartphones and tablets very often. Of the thirty participants, 63% felt that technology was important to accomplish tasks of daily living. Participants were recruited from the BDES (Hons) Interaction Design course at the Belfast School of Art. . There will be a context of IT proficiency bias in this group.

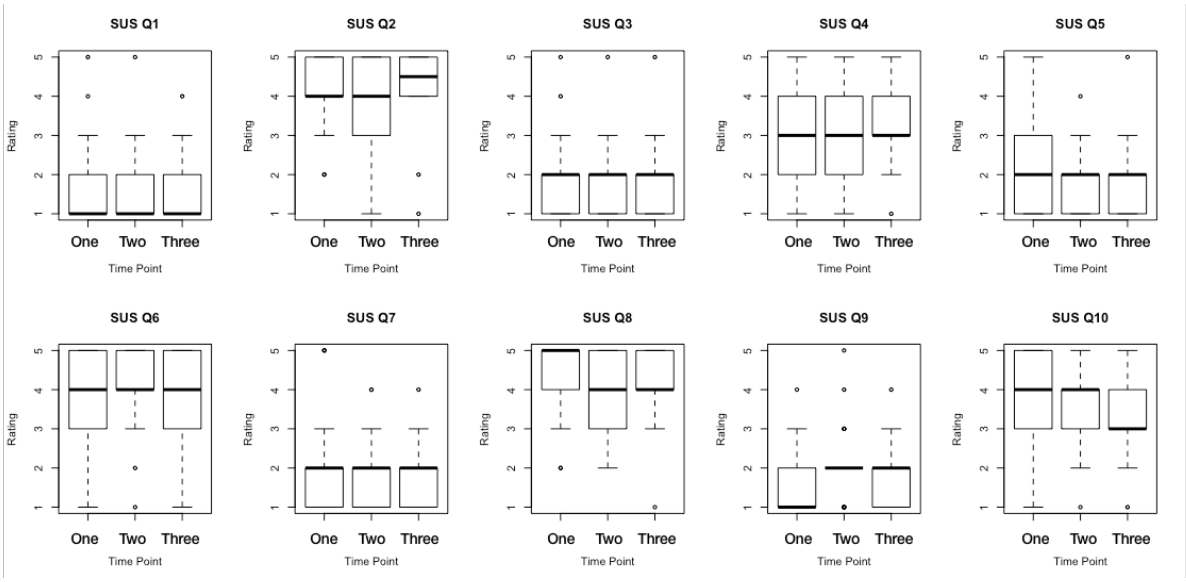

Figure 3: The Boxplots showing ratings for each SUS question at each time point.

4. RESULTS

A total of 76 SUS survey completions were collected. This comprises of 33 SUS survey completions at time point 1 (immediately after the test), 25 at time point 2 (one week after the test) and 18 at time point 3 (two weeks after the test). Hence there was subject dropout as time progressed.

SUS distributions at time point 2 and 3 are not normally distributed (Shapiro test, $p<0.05$) whilst SUS distribution at time point 1 maybe normally distributed ($p=0.1141$) perhaps due to sample size.

Figure 2 shows that median SUS score remained similar across all three time points. Median scores did increase slightly (22.50, 25, and 23.75)

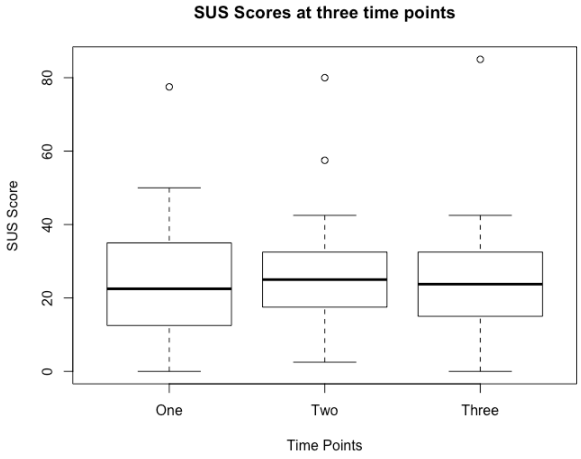

Figure 2: *Boxplots of SUS scores across three time points.*

However, Wilcoxon signed rank test showed that there was no statistical significance between the 3 SUS distributions at the three time points.

Significance was tested where $p<0.05$ (all p values were above 0.3). Interquartile Ranges (IQRs) across three time points are 22.5, 15 and 14.375 respectively. In agreement with the median, the mean SUS scores slightly increased from time point 1 to time points 2 and 3 (mean SUS scores were 24.92, 26 and 25.13 respectively). However, there is no statistical significance between these distributions and the subtle change would have no inferential changes, i.e. all average SUS scores across all three time points yield the same interpretation regarding the usability of the system.

Standard deviation of SUS scores across time points is as follows: 16.81, 17.31, 19.24 respectively. This shows a slight increase in variance as time progresses. Levene's test for homogeneity of variance indicated a statistical difference between the variance at time point 1 and the variance at time point 3 ($p<0.001$). However, whilst the variance is different, this is perhaps due to outliers and the change in variance would not be sufficient in effecting the interpretation of system usability based on SUS scores.

Figure 3 shows the boxplots of each SUS question at each test time. Questions 2, 8, 9 and 10 seem to have different medians at the different time points.

5. DISCUSSION

The SUS score survey is a widely respected tool for measuring usability [17]. While there are other surveys available such as the User Experience Questionnaire (UEQ) or the Single Ease Question

(SEQ) the SUS was chosen because of its widespread use and popularity. Therefore, the authors intended to understand the recall of a usability test and the SUS. Generally, a usability test happens in three parts. Firstly, the participants are briefed on what is to take place and background information is recorded. Secondly the test is conducted a series of tasks completed. Thirdly, the participants then complete a usability questionnaire, in this instance a SUS score to record how they used the application. The SUS score survey is administered straight after the test. In the current research two further and identical tests were conducted at one week intervals to verify the hypothesis that a participant's user experience could change over time.

The nature of this intentional rudimentary tasks chosen to be completed, was not particularly enjoyable for this group. However, this may have effected the dropout rate. Participants reported some usability challenges [14]. This potentially answers why many of the SUS scores were low.

In relation to the aims of the current research, the analysis shows that the memory and recollection of a past user experience does not change over a short period of time (3 weeks) nor does the users' memorability of user experience change.

To build up a body of work which informs choices of which usability tool to use on particular tests [18,19], future work includes further stress testing of the SUS survey by answering the following questions:

(I) Task Orientation: Is there a variation in the memorability of SUS scores when comparing a structured schedule of user tasks against casual browse and retrieval methods?

(II) Is there a variation in SUS scores when using different usability questionnaires for the same task? We would also like to conduct the same test with the range of usability questionnaires.

(III) Considering emotional design factors (Desmet. & Hekkert) Does enjoyable or desirable user interfaces result in improved memorability?

(IV) Does age and/or IT proficiency effect the recall, due to increase cognitive load during completion of the user test?

There may also be a need to consider a similar test but with significant time delay between retest.

6. CONCLUSION

There is no evidence that there is a temporal bias when completing a SUS survey, at least over a short period of time (3 weeks). As such, there is no recall bias hence researchers should not be concerned about the time at which subjects complete the SUS survey. However, a limitation in this study is that there was subject drop out across the last two time points. Some insignificant findings include the that SUS scores increased very slightly along with the variance of SUS scores as subjects relied on their memory to recall the usability of a technology interaction, which may be due to repetitive reinforcement to memory.

7. ACKNOWLEDGMENT

The researchers who conducted the study would like to thank Year 1 and 2 students on the BDes Interaction Design students at the Belfast School of Art, Ulster University who took part in the study.

8. REFERENCES

1. Smashing Magazine. (2018) *A Comprehensive Guide To User Experience Design* [online]. Available at: https://www.smashingmagazine.com/2018/02/comprehensive-guide-user-experience-design/ (Accessed: 9 April 2018).
2. Nielsen Norman Group. (2012) *Usability 101: Introduction to Usability* [online]. Available at: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (Accessed: 9 April 2018).
3. Boyd, K., Bond, R., Gallagher, S., Moore, G. and O'Kane, E. (2017), Usability and Behaviour Analysis of Prisoners using an Interactive Technology to Manage Daily Living Kaufmann.
4. Bond, R.R., Finlay, D.D., Nugent, C.D., Moore, G. and Guldenring, D., 2014. A usability evaluation of medical software at an expert conferencesetting. *Computer methods and programs in biomedicine*, *113*(1), pp.383-395.
5. Tullis, T., Albert, B. and Albert, W. (2013) *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. USA: Morgan Kaufmann
6. Sauro, J. and Lewis, J.R., 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.
7. J. Brooke, (2013) "SUS : A Retrospective," J. Usability Stud., vol. 8, no. 2, pp. 29–40. DOI?

8. S. Krug, (2009) *Don't Make Me Think! A Common Sense Approach to Web Usability*, vol. Second Edi. Berkley: Newriders.
9. Bangor, A., Kortum, P., & Miller, J. (2008). An empirical evaluation of the System Usability Scale. International Journal of Human-Computer Interaction, 24(6), pp574–594. doi:10.1080/10447310802205776
10. Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS Scores mean: Adding an adjective rating scale. Journal of Usability Studies, 4(3), pp114–123 DOI?
11. Tullis, T.S. and Stetson, J.N., 2004, June. A comparison of questionnaires for assessing website usability. In *Usability professional association conference* (pp. 1-12).
12. Y. H. Kim, D. J. Kim, and K. Wachter, 2013 "A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention," Decis. Support Syst., vol. 56, no. 1, pp. 361–370.
13. McLellan, S., Muddimer, A. and Peres, S.C., (2012). The effect of experience on System Usability Scale ratings. *Journal of usability studies*, *7*(2), pp.56-67. DOI?
14. Følstad, A., (2017). Users' design feedback in usability evaluation: a literature review. *Human-centric Computing and Information Sciences*, *7*(1), p.19. DOI ?
15. Hepple, B., 2010. The new single equality act in Britain. *The Equal Rights Review*, *5*, pp.11-24.
16. Nielsen, J., 2003. Usability 101: Introduction to usability.
17. Philip Kortum & Mary Sorber (2015) Measuring the Usability of Mobile Applications for Phones and Tablets,International Journal of Human–Computer Interaction, 31:8, 518-529, DOI: 10.1080/10447318.2015.1064658
18. Orfanou, K., Tselios, N. and Katsanos, C., (2015). Perceived usability evaluation of learning management systems: Empirical evaluation of the System Usability Scale. *The International Review of Research in Open and Distributed Learning*, *16*(2). DOI?
19. James R. Lewis (2014) Usability: Lessons Learned … and Yet to Be Learned, International Journal of Human–Computer Interaction, 30:9, 663-684, DOI: 10.1080/10447318.2014.930311