Play to Improve: Gamifying Usability Evaluations in Virtual Reality

Abhijai Miglani(✉), Sairam Kidambi(✉), and Praveen Mareguddi(✉)

Philips Innovation Campus, Bengaluru, Karnataka, India
{abhijai.miglani,sairam.kidambi,praveen.mareguddi}@philips.com

Abstract. Objective: The research study focuses on evaluating how usability engineering related activities would look like in virtual reality from an exploratory point of view. Research questions for the study are: a) What quantitative impact does an environment have on the usability evaluation of software interface components? b) How is this impact different in virtual environment as compared to physical environment? c) What could be the best interaction design representation in virtual reality which could have a similar mental model for the users as having a mouse and keyboard in physical environment? d) What role does another user/a virtual mannequin and other elements/objects play on influencing the usability evaluation results in virtual reality?

Background: As per ISO 62366 and Food and drugs administration (FDA) guidelines, simulating usage environment when evaluating software components is crucial. However, with conventional lab environment usability testing sessions have no environment simulated in it.

The research focuses on how the transition path (moving from physical to virtual environment) would look like if a researcher wants to thoroughly evaluate a design concept in virtual reality.

Method: Participants (N = 8) participated in the experiment to evaluate 3 interaction design concepts in virtual reality (1) gaze timer: seeing the virtual monitor for 3 s to go to the next page in the workflow, (2) gaze click: seeing the virtual monitor and using controller to aim and go to the next page in the workflow, and (3) gaze gesture: seeing the virtual monitor and using controller to pick-drag-drop a page in the workflow stack to another location. The three interaction design concepts varied in physical workload, cognitive workload, familiarity, learning curve and readability. The experiment design was a within subject design.

Results: Participants preferred gaze click interaction design concept over gaze timer and gaze gesture concept.

Conclusion: Having a virtual environment added to a conventional lab/physical environment, transition could be possible. Replacement of controls like mouse and keyboard could be done by adding gaze click interaction.

Application: Results of the study could serve as providing design guidelines for simulation of software interfaces' usability evaluation in virtual reality.

Keywords: Virtual reality (VR) · Usability Engineering · Food and Drugs Administration (FDA) · Interaction design · Usage environment · Gamification

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12423, pp. 248–266, 2020.
https://doi.org/10.1007/978-3-030-60114-0_17

1 Introduction

The first Healthcare software components when manufactured follow ISO 62366 in the software development process they opt. As per ISO 62366, one of the major components to make the software products fail-safe from human error is to follow Usability Engineering or Human Factors Engineering. Usability Engineering requires the software manufactures to include multiple rounds of usability evaluation – formative and summative usability testing. This is also in accordance with the guidelines provided by Food and Drugs Administration (FDA). Moreover, these guidelines also suggest simulating usage environment when evaluating software components.

The type of environment or context – physical and social also effects the results of the components' usability evaluation as pointed by (Trivedi 2012). In this research, the researchers defined physical environment as refers to the environment in which user is tested and social environment as the environments having people involved. The laboratory evaluations do not simulate the context when usability testing is done with mobile phones, because laboratory settings lack the desired ecological validity. Similar claim also comes from (Park and Lim 1999) where they state that simulating the use settings is very hard, time consuming, expensive and lacks contextual factors. Field testing takes place in a more natural setting.

With respect to social context (Trivedi 2012) mentions people involved in usability evaluations can be the evaluators, the test monitors, the users, and other people who may not be directly involved with the evaluation, however, their presence can have a substantial effect on the results of usability evaluations.

(Tsiaousis and Giaglis 2008) examined the effects of environmental distractions on mobile website usability. They categorized the environmental distractions into auditory, visual and social. Results confirmed that environmental distractions have direct effect on mobile website usability.

Study by (Jacobsen et al. 1998) examines the evaluator effect in the usability tests. In their study four HCI research evaluators, all familiar with the theory and practice of usability analyzed four video tapes. The results indicate that only 20% of the 93 unique problems were detected by only a single evaluator. Severe problems were detected by more often by all four evaluators (41%) and less often by only one evaluator (22%), however, the evaluator effect remained substantial.

The product/prototype in the research study under focus offers an advantage over the conventional usability evaluation method by extending the method to include environment also as a component.

Similar work has been done by (Madathil and Greenstein 2011), where the researchers propose a synchronous remote usability testing, also using virtual environment for introducing environment in the traditional lab environment. The significance of this study is that the synchronous remote usability testing using virtual lab provides similar results as of the traditional lab method and in some respects it works better than the conventional approaches.

Interestingly, participants appeared to identify a slightly larger number of lower severity defects in the virtual lab environment than in the traditional lab and WebEx environments. The results of this study suggest that participants were productive and enjoyed the virtual lab condition, indicating the potential of a virtual world based approach as an alternative to the conventional approaches for synchronous usability testing.

Our research is novel in terms of the tools and the research methods opted. The research also has prime focus on evaluating the effect of virtual environment for usability evaluations (and less on remote usability testing). Also, due care was taken to address potential concerns that could arise due to acceptance of the prototype as a tool for usability evaluations. These concerns could arise because of user's familiarity and mental model with the tools used in the physical/lab environment for usability evaluations like mouse and keyboard for instance. Moreover, as part of exploration process, we are evaluating the contribution of different factors that would effect user's immersion in the environment and hence, acceptance of the prototype/tool.

As pointed out by (Rajanen and Rajanen 2017), for usability in gamification, there are some guidelines – define business objectives, delineate target behaviors, describe your behaviors, device activity loops, do not forget the fun and deploy the appropriate loops. The prototype designed to answer the research questions is a game (see section – prototype design) having components like instructions/demo given by a virtual mannequin, tasks being performed by the user, rewards obtained being the pleasure out of evaluation of design.

However, with the inclusion of environment and including all the components in virtual reality (that were part of physical environment), the accuracy of the results could change and could be different as when validated in physical environment.

The research focuses on how the transition path (moving from physical to virtual environment) would look like if a researcher wants to thoroughly evaluate a design concept in virtual reality. Results of the research study could serve as providing design guidelines for simulation of software interfaces' usability evaluation in virtual reality. These design guidelines if standardized by also including software as a medical device (SAMD) components, could help further in the FDA submissions.

Research questions for the study are: a) What quantitative impact does an environment have on the usability evaluation of software interface components? b) How is this impact different in virtual environment as compared to physical environment? c) What could be the best interaction design representation in virtual reality which could have a similar mental model for the users as having a mouse and keyboard in physical environment? d) What role does another user/a virtual mannequin and other elements/objects play on influencing the usability evaluation results in virtual reality?

For answering the research questions, two research hypotheses were framed. For the first research hypothesis, independent variable was chosen as environment realism and dependent variable was chosen as task success rate. Environment realism could be further segregated into components like social presence, objects in virtual reality, lighting – textures – materials and audiovisual cartography. Second hypothesis describes different constructs for the interaction modalities/prototype - independent variable and its behavior with respect to task success rate - dependent variable.

Figure 1 shows the first hypothesis - difference in task success rate for different environments with respect to environment realism.

Figure 2 shows the second hypothesis - difference in task success rate for different interaction modalities/prototypes with respect to modalities' properties.

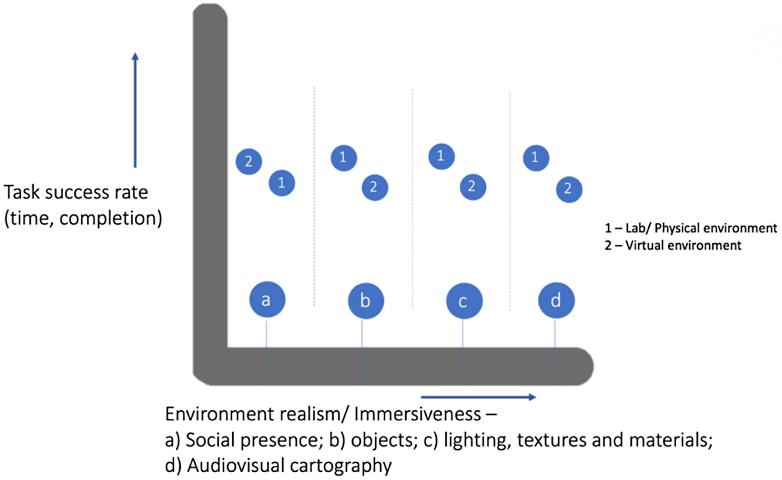

Fig. 1. First hypothesis graph showing effect of environment realism on task success rate

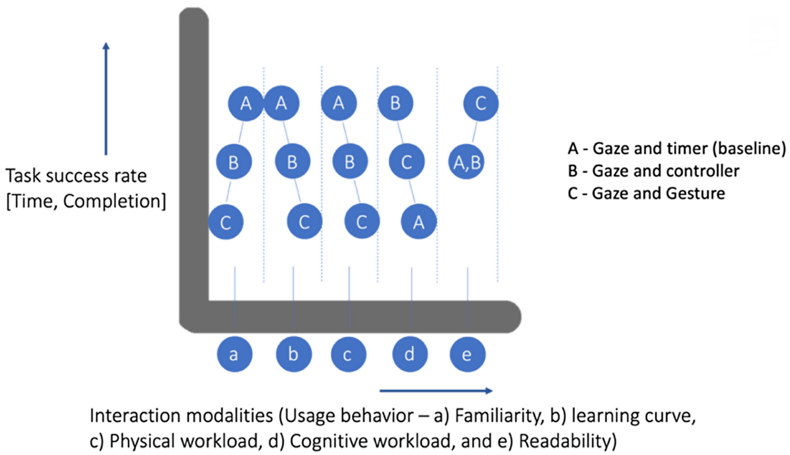

Fig. 2. Second hypothesis graph showing effect of interaction design concepts' properties and task success rate

The research hypothesis was framed considering the unified technology acceptance model.

Environment realism variance is correlated to habit, effort expectancy, performance expectancy factor, social influence, hedonic motivation, age, gender, experience and facilitating conditions.

Interaction modality design intuitiveness is related to facilitating conditions and effort and performance expectancy, age, gender, experience, habit and hedonic motivation.

The independent variables are related to performance and effort expectancy and hence, user behavior (task success rate – dependent variable).

2 Method

2.1 Apparatus

A stock Oculus Rift was used with the two controllers. However, for the experiment, only right hand controller was used. The Oculus Rift also comes with two sensors for setting the space/boundary, which was also used.

The computer used for simulating the experiment was HP ZBook 15 G3 with 16 GB RAM and i7, 2.7 GHz.

In the experiment, Philips Ingenia 3.0 T MRi machine was simulated with appropriate noise coming from the machine.

2.2 Ethnography Studies

To study the environment that has to be simulated in virtual reality, a site visit was done where Philips Ingenia 3.0 T MR system is installed.

Philips Ingenia 3.0 T MR machine is a magnetic resonance machine by Royal Philips. The machine has a software user interface also to control the machine.

Position and movement of the objects in the room, movement of hospital staff, lighting and noise conditions were studied.

More importantly, tasks performed by the hospital staff were analyzed. In the room next to the room where MR system is installed, the technologist is responsible for initiating, capturing and analyzing patient's scans. The administrator is responsible for assisting the technologist with patient data entry and printing the scans.

2.3 Participatory Design Workshop

For ideation of the ethnographic research results, a participatory design workshop - Storyply was conducted. Storyply method was developed by (Atasoy and Martens n.d.). The method describes 9 steps to convert a research problem into design ideas. When ethnographic research results were presented to the participants in form of research problem, 3 design ideas came out, out of one was included in the prototype design. Participants suggested to have a demonstration of the usability session in virtual reality. This idea was included in the prototype design in form of stage-1 of the game - instructions. Please refer to section - prototype design for more details.

2.4 Prototype Design

Three set of prototypes were designed in virtual reality and each set was categorized as 3 stages for the participant to complete – introduction, performance and exit. The introduction and the exit was common across the 3 sets. Whereas, the performance stage varied based on the research hypothesis (interaction design in this case) to be evaluated. The prototype concepts were designed considering the heuristics suggested by (Sutcliffe and Gault 2004).

The three sets consisting of 3 stages each are shown as a storyboard in Fig. 3. As shown in the figure, based on the interaction design – gaze timer or gaze click or gaze gesture, the participants had to perform the task of undergoing through the workflow of MR Ingenia 3.0 software interface. The workflow of the application consists of pages - patient selection, exam routine selection, exam routine start, print area/layout mapping and end acquisition. Other than gaze timer and gaze click, in gaze gesture, the workflow pages were arranged as separate blocks (as a stack) on top of each other which the participant had to pick, drag and drop to another location. In gaze timer and gaze click, participants had to look at the pages (to initiate the workflow) and click using the right controller.

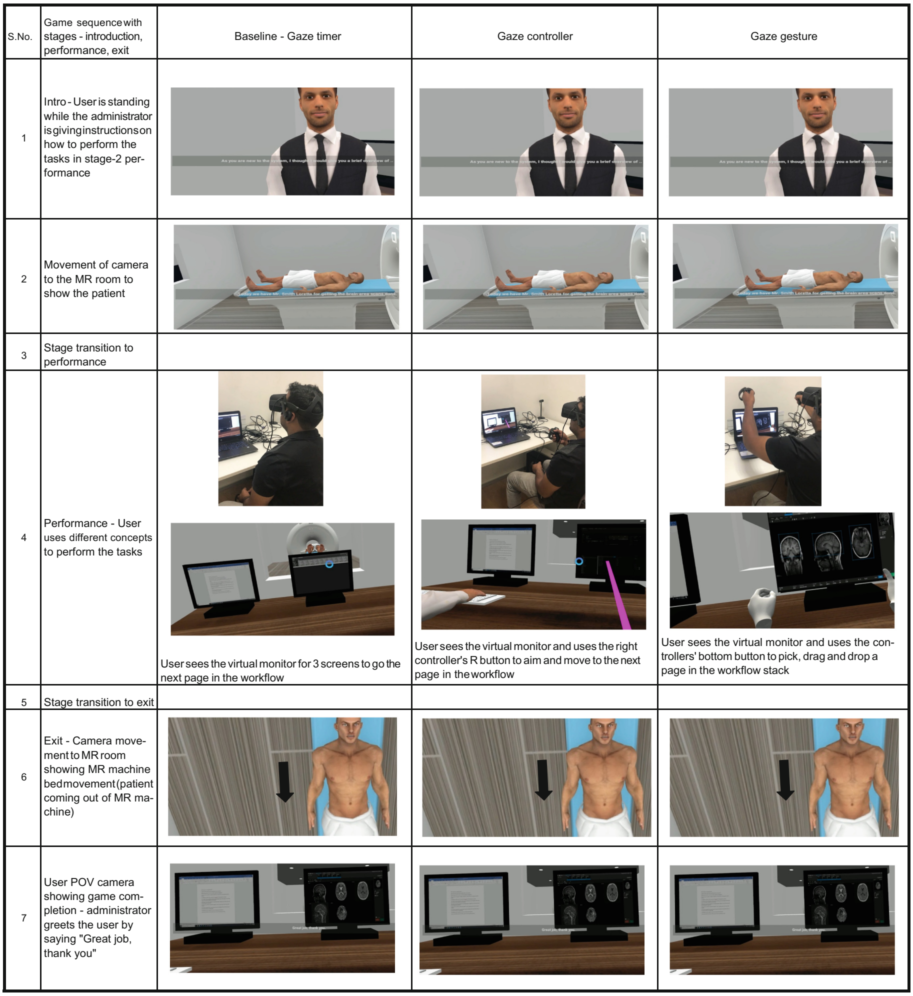

Fig. 3. Storyboard showing timeline for actions in Concept A: Gaze timer, Concept B: Gaze click and Concept C: Gaze gesture

For gaze timer, the gaze was denoted as a circular cursor/crosshair as shown in Fig. 3. For gaze click, the controller was used to aim at a point of any page followed by clicking on controller's 'A' button. The aiming was aided by showing a pink colored ray. In gaze gesture, virtual hands were present to aid picking of pages from the stack.

The interaction design concepts varied (refer to section - introduction) in terms of familiarity, learning curve, physical workload, cognitive workload and readability. For Concept – A: Gaze timer, familiarity and task success rate was hypothesized to be the highest followed by Concept B: Gaze click and Concept C: Gaze gesture. The hypothesis was based on the fact that more the familiarity, more would be the task success rate. As in Concept C, gestures were involved, it could come as something new to the participants. Whereas, in Concept A, participant just had to look at the virtual monitor to go to the next page in the workflow. Participants would be more familiar with the seeing task than the inclusion of gestures in the task. Learning curve for Concept C would be highest (with least task success rate) as participants would be least familiar with the gesture system.

With respect to Concept C, physical workload would be highest as it involves physical movements for picking-dragging-dropping the pages of a stack to another location. With physical workload highest for Concept C, task success rate would be least as it would take maximum time with respect to Concept A (just seeing would involve least physical workload) and Concept B.

With respect to cognitive workload, Concept A had the highest cognitive workload (with least success rate) followed by Concept C and B. For Concept A, participants would be looking at the virtual monitor for 3 s with 'proper focus' to go to the next page in the workflow. Concept B and C did not require participants to focus at the virtual monitor. However, concept C involved pick-drag-drop functionality which requires participants' attentional resources to complete the task.

In Concept-C pick-drag-drop functionality is provided, the participants can move a page of the workflow closer to have a better look at the content of the page (unlike in Concept A and B). Readability improves with having affordance of having the functionality to read the content of a page.

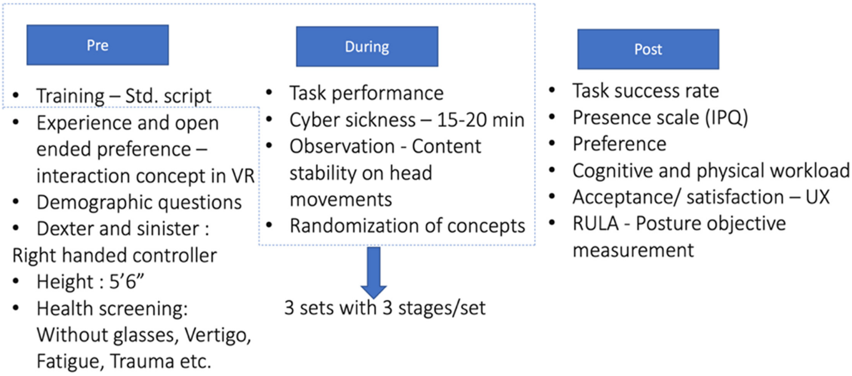

Fig. 4. Experiment design containing pre, during and post experiment sections

The prototype designed is modular in nature i.e. the same prototype could be used to evaluate any other system with the required adjustments in some components. There is a base component to the prototype which could be kept as common across any system evaluation.

Also, the 3 sets of game were designed keeping a time frame of 15–20 min into consideration. A game set having this time duration would ensure the participants do not feel cyber sick due to their continuous exposure to virtual reality.

As shown in the Fig. 3, the introduction stage had the assistant explaining the tasks to be performed by the participant in stage 2 - performance. As mentioned by (Junglas et al. 2007) trust in digital identities is crucial for immersion in virtual reality. To maintain the immersion levels, photorealistic mannequins were used with appropriate body joint movements and synchronization with the speech. Also spatial auditory noise was simulated to be coming from the MR room to enhance the immersion.

Furthermore, following the site visit done in ethnography studies stage, the usability room was designed in virtual reality taking the site visited as reference – with appropriate lighting, textures and materials applied on the objects in the room.

As pointed out in section – introduction, the virtual reality room was designed keeping the concept of gamification into consideration and designing different stages of the game accordingly. Each stage has a transition in between to make the movement between one stage to another a smooth one. With respect to the first and the third stage, some animations were designed including camera positions with its movements and subtitles on the footer.

2.5 Procedure

To evaluate the research hypothesis framed, a within experiment design was framed as shown in Fig. 4.

All participants had to fill a pre-experiment survey before they started participating in the experiment. The survey had questions related to participants' demographics (including height and if they wear eye glasses), experience with virtual reality technology, dexterity etc.

Due to the virtual mannequin of the assistant/administrator incorporated in the game and the eye point of the participant in virtual reality, it was crucial to ask the participants about their height. Also, the contrast of the overall prototype designed in virtual reality could have varied if the participants had eyeglasses. Hence, these parameters were covered as part of pre-experiment survey.

After filling the survey, the participants proceeded with wearing the Oculus Rift headgear. As mentioned before, the tasks to be performed by the participants were incorporated as a 'virtual speech script' in the game's introduction stage. The script also described how to go about performing the tasks using the three interaction modalities.

As mentioned above, the three interaction design concepts had variation in physical workload, cognitive workload, readability ease, familiarity/learnability. After the experiment was performed, the participants had to fill questionnaires pertaining to all these parameters/dependent variables. For workload, NASA – TLX was consulted and for acceptance/satisfaction – SUS scale was referred to.

Preference was also asked – which out of the 3 concepts would you prefer using?

Also, change in posture due to movements of head, elbow, hands, fingers and upper back portions of participants' body when experiencing the interaction design concepts was captured using rapid upper limb assessment (RULA) method.

Moreover, task completion rates were also captured.

Last but not the least, behavioral validity of the prototype having components like presence, realism etc. was captured using igroup presence questionnaire (IPQ).

3 Results

A convenience sample of 8 user interface/experience designers working at Philips Innovation Campus was taken. Out of the 8 participants, 5 participants wore glasses. It was critical to ask participants if they wear glasses as the virtual reality Oculus head-gear they had to wear did not allow glasses. Without the glasses, readability of the content could have been an issue.

Participants were also screened for health. All but one participant was found to be healthy as per the questionnaire designed. A participant had a history of migraines, had ear and balance problems and was claustrophobic.

Health screening questionnaire was essential to prevent participants feeling cyber sick. For the same reason, the content in virtual reality restrained to 15–20 min.

With respect to the demographics questionnaire, 7 out of 8 participants belonged to the age group 25–40 and one participant belonged to the age group 41–60. Gender wise, 5 participants were males and 3 were females.

All participants were working on different products and had no overlap in the work they were doing.

Participants were also asked to mention any other design idea (other than the design concepts) by which they would like to create a digital version of usability evaluations.

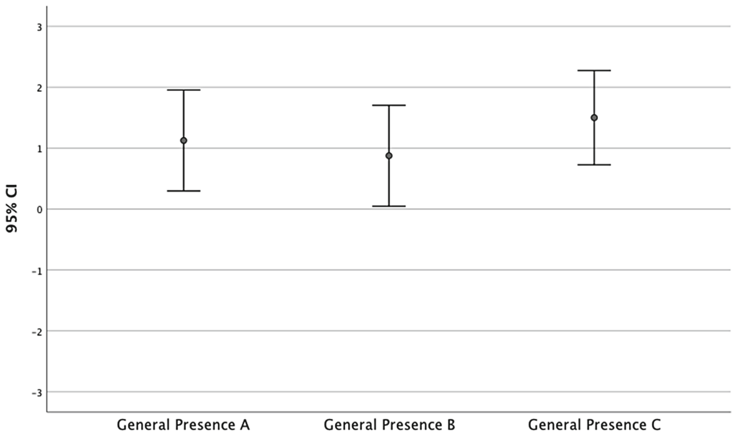

Fig. 5. General presence construct of IPQ scores for Concept A gaze timer, B gaze click and C gaze gesture

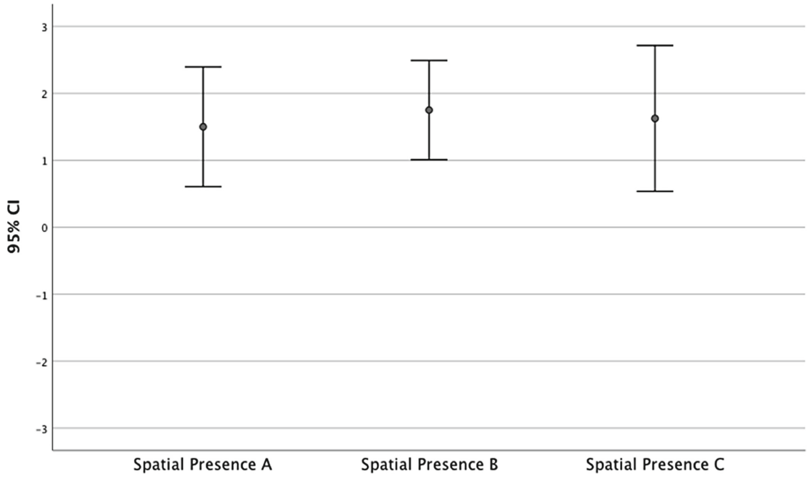

Fig. 6. Spatial presence construct of IPQ scores for Concept A gaze timer, B gaze click and C gaze gesture

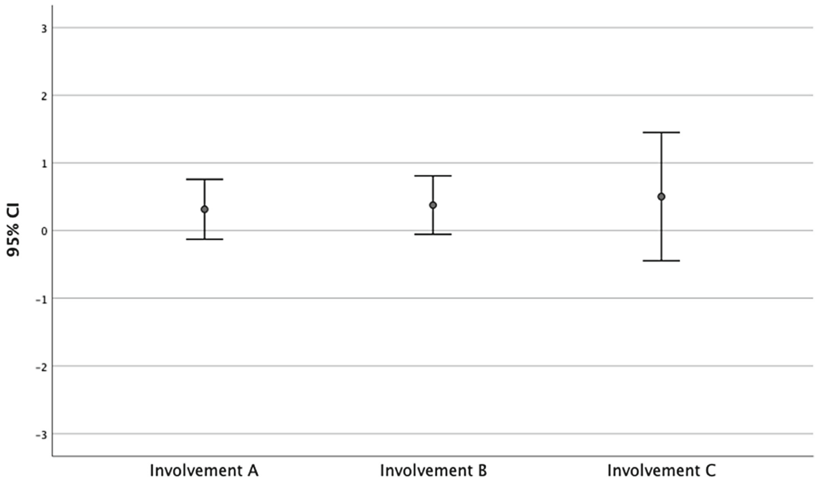

Fig. 7. Involvement construct of IPQ scores for Concept A gaze timer, B gaze click and C gaze gesture

Most of the participants had varied feedback varying from having voice feedback commands to having augmented reality based mobile application. Also, as per the participants, they considered having a digital version of usability evaluations as highly useful (75%) and neutral (25%).

Hands' dexterity was also asked from participants and majority were right-handed. One participant was ambidextrous. This question was essential to ask as participants had to use their right hand for controlling the right controller while evaluating concept B and C.

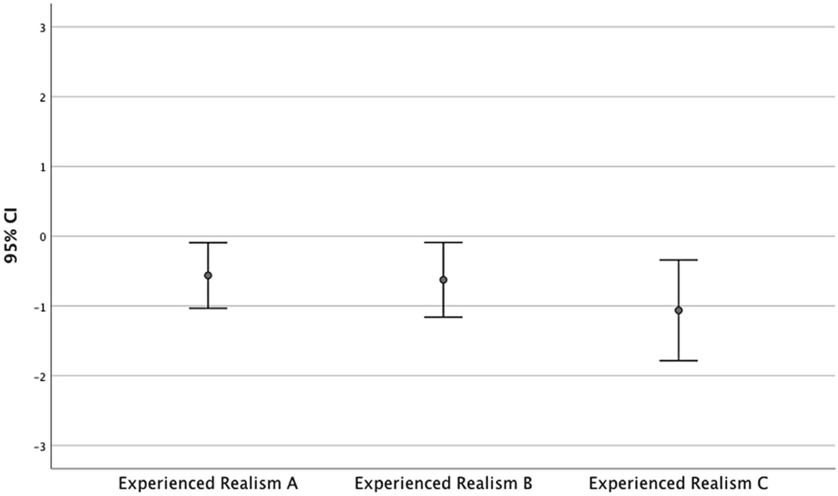

Fig. 8. Experienced realism construct of IPQ scores for Concept A gaze timer, B gaze click and C gaze gesture

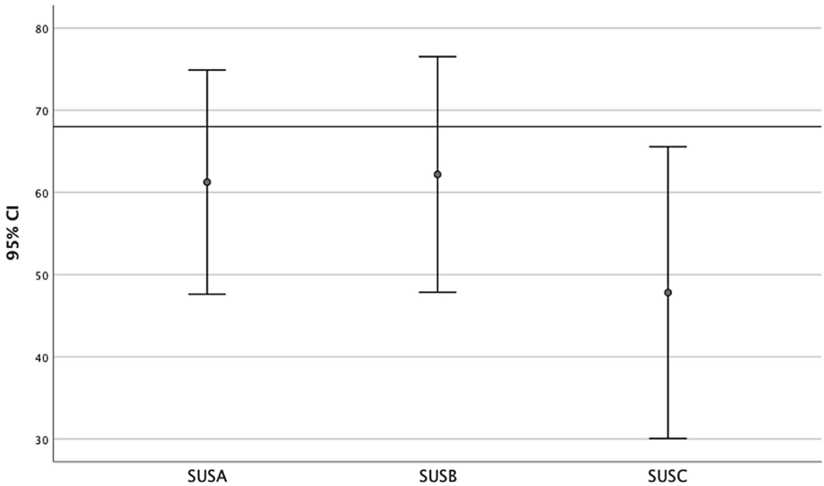

Fig. 9. System usability scale (SUS) scores with threshold value of 68 for Concept A gaze timer, B gaze click and C gaze gesture

Moreover, the most important need and most critical task if participants were playing a radiologist or a technologist was also asked. The tasks given as options were reading content, scrolling up/down or zooming in/out and using mouse to go to the next screen. Whereas the needs given as options were learnability of the application, readability of the content and comfortable workflow. Reading the content and comfortable workflow were the options majority of the participants marked as most important task and need respectively.

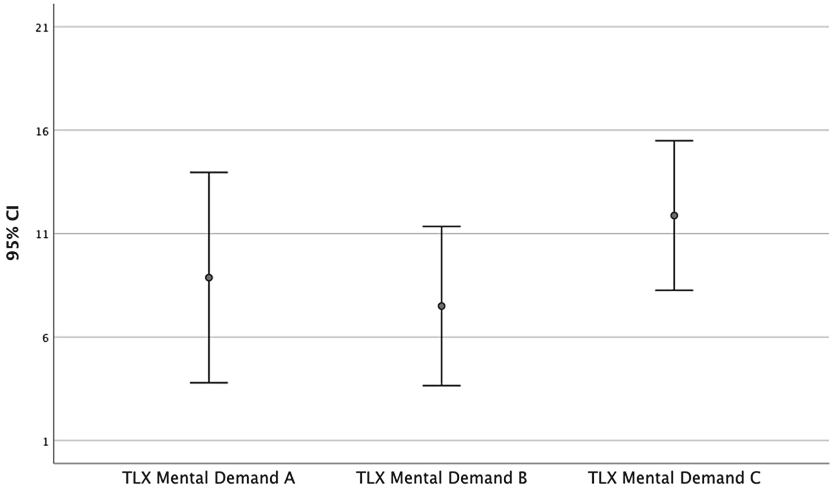

Fig. 10. Mental demand construct scores (Task load index – TLX) for Concept A gaze timer, B gaze click and C gaze gesture

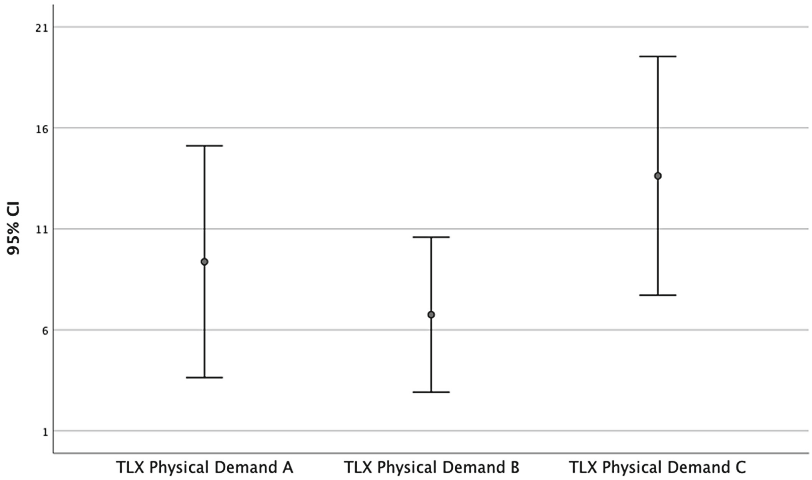

Fig. 11. Physical demand construct scores (Task load index – TLX) for Concept A gaze timer, B gaze click and C gaze gesture

First dependent variable being general presence, the order varied with different concepts as shown in Fig. 5. Concept C had the highest general presence followed by Concept A and B. Other than Concept A, normality tests (Shapiro wilk) showed the significance as more than 0.05. Hence, both ANOVA and Friedmann tests were conducted. Friedmann tests showed $x^2 = 2.381$, df = 2, p = 0.304 > 0.05. Furthermore, Mauchy's test of sphericity showed approx. x^2 value as 0.325 with p as 0.850. Hence, Huynh Feldt test was conducted to study effect of different concepts on general presence variable. It was

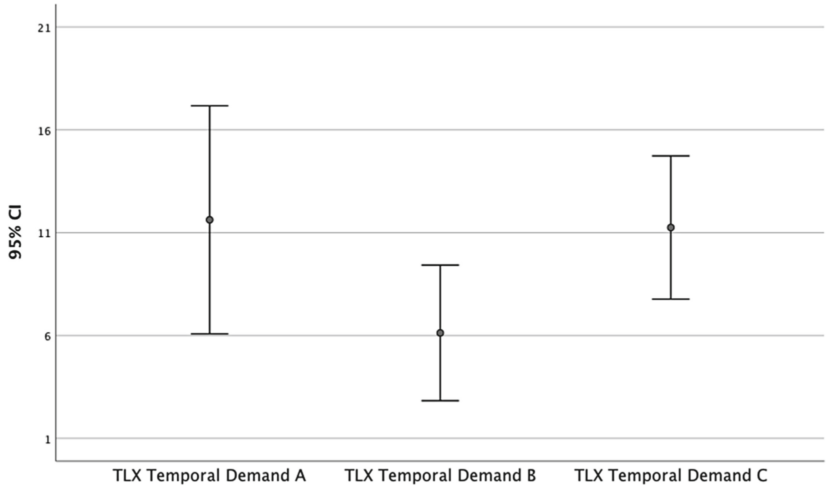

Fig. 12. Temporal demand construct scores (Task load index – TLX) for Concept A gaze timer, B gaze click and C gaze gesture

found out that F(2,14) – 1.727, $p = 0.214 > 0.05$ and quadratic trend was non-significant with F(1,7) – 2.054, $P = 0.195 > 0.05$.

Spatial presence being the second dependent variable has the order variation with different concepts shown in Fig. 6. Concept B had the highest spatial presence followed by Concept C and A. As per Shapiro-wilk test of normality, no concept had normally distributed data as significance for all concepts were more than 0.05. Hence, Friedmann test was done further. Friedmann tests showed the effect of different concepts on spatial presence as non-significant - $x^2 = 0.471$, df = 2, $p = 0.79 > 0.05$.

For involvement, the order variation with different concepts is as shown in Fig. 7. Concept C had the highest score followed by Concept B and A. Shapiro Wilk's test of normality did not show any significance for any concept and hence, Friedmann test was followed to test significance of effect of different concepts on Involvement scores. Friedmann test showed $x^2 = 0.4$, df = 2, $p = 0.819 > 0.05$.

Normality tests – Shapiro Wilk showed no significance for experience realism as per different concepts. Friedmann test was followed to test significance of effect of different concepts - $x^2 = 2.471$, df = 2, $p = 0.291 > 0.05$. The order for values of experienced realism as per different concepts was Concept A had the highest value followed by Concept B and C (Fig. 8).

For system usability, system usability score (SUS) was the highest for Concept B followed by Concept A and C. Normality tests – Shapiro Wilk showed significance for no concept and hence, Friedmann test were conducted to test effect of different concepts on system usability - $x^2 = 1.742$, df = 2, $p = 0.419 > 0.05$ (Fig. 9).

For task load index, Mental demand had the highest score for Concept C followed by Concept A and B. Normality tests – Shapiro Wilk showed significance for no concept and hence, Friedmann test were conducted to test effect of different concepts on Mental Demand - $x^2 = 6.467$, df = 2, $p = 0.039 < 0.05$ (significant) (Fig. 10).

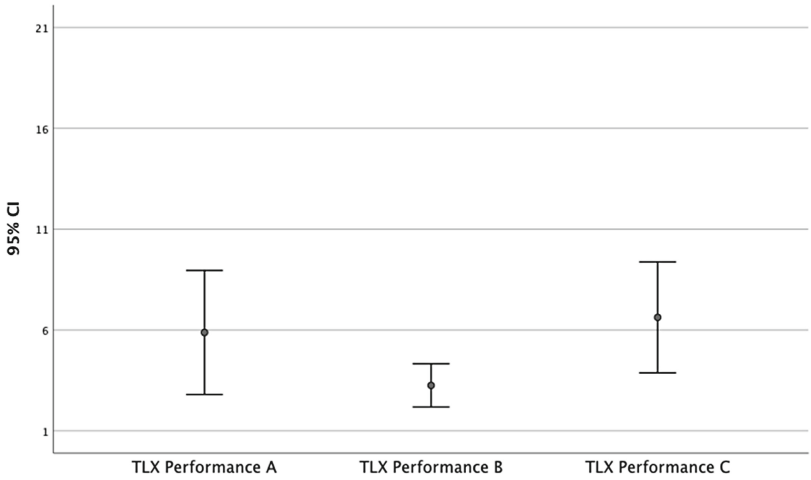

Fig. 13. Performance construct scores (Task load index – TLX) for Concept A gaze timer, B gaze click and C gaze gesture

On the other hand, Physical demand scored highest for Concept C followed by Concept A and B. Normality tests – Shapiro Wilk showed significance for no concept and hence, Friedmann test were conducted to test effect of different concepts on Physical demand - $x^2 = 6.870$, df = 2, $p = 0.032 < 0.05$ (significant) (Fig. 11).

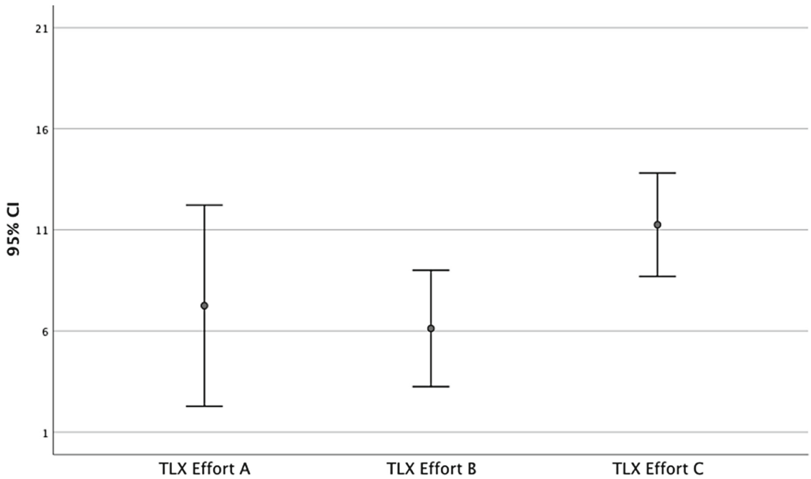

Fig. 14. Effort construct scores (Task load index – TLX) for Concept A gaze timer, B gaze click and C gaze gesture

Temporal demand construct had the highest score for Concept A followed by Concept C and B. Shapiro Wilk test for normality showed no significance for any concept and hence, Friedmann test was conducted to test effect of different concepts on Temporal Demand - $x^2 = 5.067$, df = 2, p = 0.079 > 0.05 (Fig. 12).

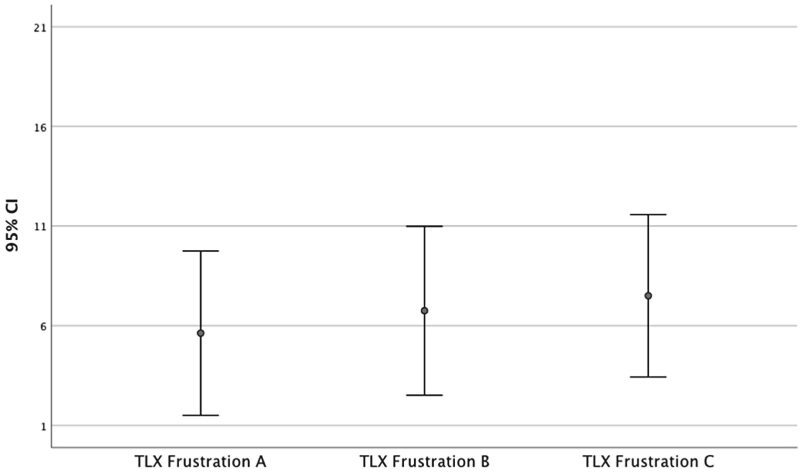

Fig. 15. Frustration construct scores (Task load index – TLX) for Concept A gaze timer, B gaze click and C gaze gesture

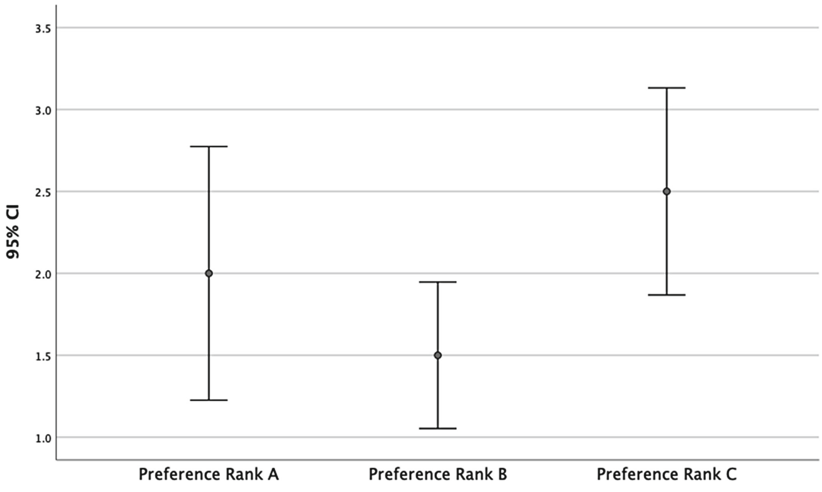

Fig. 16. Preference rank for Concept A gaze timer, B gaze click and C gaze gesture

Performance scored highest for Concept C followed by A and B. Shapiro wilk, the test for normality showed no significance for any concept. Friedmann tests concluded $x^2 = 3.267$, df = 2, p=0.195 > 0.05 (Fig. 13).

Effort scored highest for Concept C followed by A and B. Shapiro wilk showed no significance for normality of any of the concepts. Friedmann test scored $x^2 = 5.429$, df = 2, p = 0.066 > 0.05 (Fig. 14).

Frustration had highest score for Concept C followed by Concept B and A. Shapiro wilk showed significant normality for just Concept B. Friedmann test scored $x^2 = 1.786$, df = 2, p = 0.409 > 0.05. Whereas in ANOVA Mauchy's test for sphericity showed approx. x^2 value as 0.788 with p as 0.674. The effect of concepts on frustration was non-significant, Huynh Feldt scored F(2,14) – 0.549, p = 0.59 > 0.05. Linear trend of the frustration scores was also non-significant F(1,7) – 1.305, p = 0.291 > 0.05 (Fig. 15).

Rapid upper limb assessment (RULA) scores for concept B and C were measured and it came out to be 2 and 4–6 respectively showing actions needed for concept C. However, posture was acceptable for Concept B.

Most importantly, Preference scored rank 1 for Concept B, rank 2 for Concept A and rank 3 for Concept C. Normality test – Shapiro Wilk showed significance for all concepts. In ANOVA Mauchy's test of sphericity showed approx. x^2 value as 2.433 with p as 0.296. The effect of concepts on preference rank was non-significant, Huynh Feldt – F(1.818,12.727) – 2.333, p = 0.140 > 0.05. The quadratic trend for different concepts' preference ranking was significant – F(1,7) – 7, P = 0.033 < 0.05 (Fig. 16).

For Concept A, qualitative feedback included comments about the timer – "timer was less, it should be more than what it is right now (3 s)", easiness – "The task was easy, no physical demand was required", control – "no control was sensed, I was not aware when I was controlling". A word cloud is shown in Fig. 17.

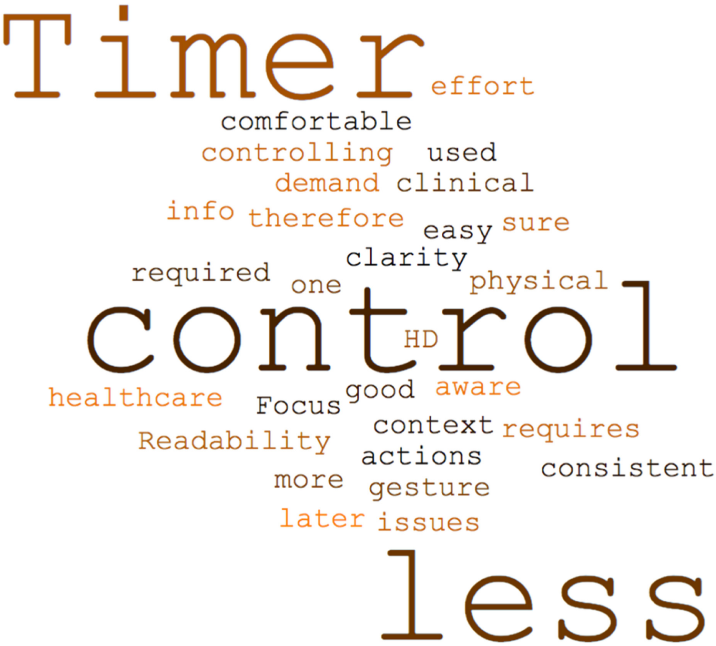

Fig. 17. Word cloud for comments and improvement suggestions received for Concept A gaze timer

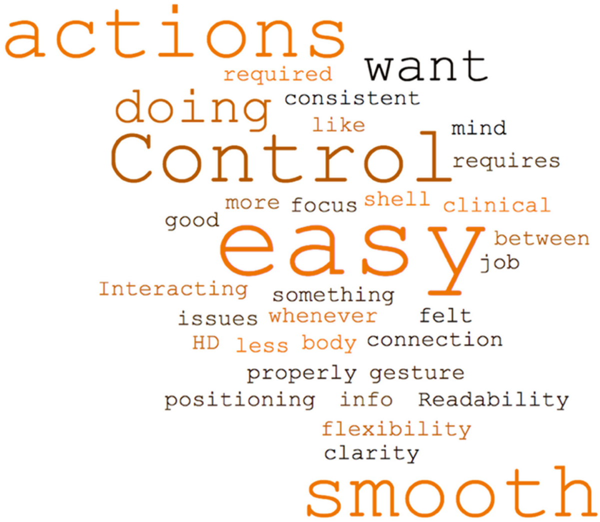

Fig. 18. Word cloud for comments and improvement suggestions received for Concept B gaze click

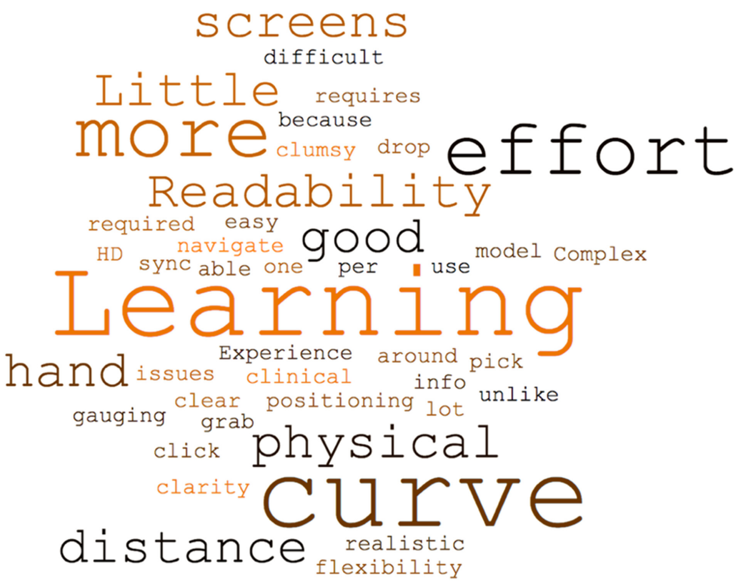

Fig. 19. Word cloud for comments and improvement suggestions received for Concept C gaze gesture

For Concept B, feedback was positive – “I am able to control my actions, easy and smooth”, “I am doing what I want to do” and “No focus required to complete the tasks” (Fig. 18).

For Concept C, feedback was mixed – "It takes effort to complete the tasks", "Readability is good", "Requires learning", "Tasks were not realistic with respect to physical model of screens in a monitor", "Experience was good as I can have more screens in front of me". Other improvement suggestions included having the functionality of grabbing the screens from a distance to avoid bending in the posture (Fig. 19).

4 Discussion

In this study, we used a prototype to simulate environment for the conventional usability lab environment. The prototype was designed to answer the research questions and were inclined towards the research hypothesis framed.

With respect to the first research question, as per (Trivedi 2012), physical and social environment can influence results in usability evaluation of a software component.

With respect to influence of different factors like visual, auditory and social on usability evaluation results, (Tsiaousis and Giaglis 2008) confirmed that environmental distractions have direct effect on direct impact on usability. Also, (Jacobsen et al. 1998) showed the impact of having another evaluator (social presence) on usability evaluation results.

For virtual environments, system usability scores (SUS) for the three interaction concepts designed namely, gaze timer, gaze click and gaze gesture had mean scores falling below the threshold value of 68. This shows that virtual environment does affect the usability when virtual environment interaction concepts are used (as compared to mouse or keyboard used in physical environment).

Moreover, positive correlations were found between gaze gesture concept and involvement construct (spearmann rho – 0.723, p-0.043) and experienced realism construct (spearmann rho – 0.704, p-0.051). Gaze gesture concept having lowest usability, had highest involvement and lowest experienced realism as compared to the other two interaction concepts.

For the third research question, participants preferred concept B: gaze click the most. Evidently, participants prefer to have control on the actions they are performing – similar to mouse and keyboard. Moreover, RULA scores also showed an acceptable posture for the concept. This is also in line with the usability scores (highest), mental demand scores (lowest), physical demand scores (lowest), temporal demand scores (lowest) and effort scores (lowest).

When making a transition from a physical environment or a conventional lab environment to having a virtual environment, gaze click concept can be used to control the workflow of pages. This guideline can be used further by other researchers working on usability engineering in virtual reality theme.

References

Trivedi, M.: Role of context in usability evaluations: a review. ACIJ (2012)

Park, K.S., Lim, C.H.: A structured methodology for comparative evaluation of user interface designs using usability criteria and measures. Int. J. Ind. Ergon. **23**, 379–389 (1999)

Tsiaousis, A.S., Giaglis, G.M.: Evaluating the effects of the environmental context-of-use on mobile website usability, pp. 314–322 (2008)

Jacobsen, N., Hertzum, M., John, B.: The evaluator effect in usability studies: problem detection and severity judgments. In: Human Factors and Ergonomics Society 42nd Annual Meeting, pp. 1336–1340 (1998)

Madathil, K.C., Greenstein, J.S.: Synchronous Remote Usability Testing - A New Approach Facilitated By Virtual Worlds. *CHI*. ACM, Vancouver (2011)

Rajanen, M., Rajanen, D.: GamiFIN, Pori, Finland (2017)

Atasoy, B., Martens, J.-B.: STORYPLY: designing for user experiences using storycraft. In: Markopoulos, P., Martens, J.-B., Malins, J., Coninx, K., Liapis, A. (eds.) Collaboration in Creative Design, pp. 181–210. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29155-0_9

Junglas, I.A., Johnson, N.A., Steel, D.J., Abraham, D.C., Loughlin, P.M.: Identity formation, learning styles and trust in virtual worlds. DATA BASE Adv. Inf. Syst. **38**, 90–96 (2007)

Sutcliffe, A., Gault, B.: Heuristic evaluation of virtual reality applications. Int. Comput. **16**, 831–849 (2004)