Usability and user experience evaluation of Virtual Integrated Patient

Pabba Anubharath
National University of Singapore
Singapore

Yoon Ping Chui
Singapore University of Social Sciences
Singapore

Judy Sng
National University of Singapore
Singapore

Lixia Zhu
National University of Singapore
Singapore

Kai Tham
National University of Singapore
Singapore

Edmund J.D. Lee
National University of Singapore
Singapore

Most existing Virtual Patients utilize simplistic, predictable, and prescriptive approaches that limit deductive learning and the development of decision-making skills for medical students. We have designed a chat-based virtual patient for performing patient interviews, physical examinations, and investigations to help medical students develop reasoning skills. In this paper, we present results from a two-part study. In the first part, we conducted a usability evaluation with seven medical students and six clinicians. The objectives of the usability evaluation was to determine how VIP's user interface and its features affect the usability (efficiency, effectiveness and learnability) as well as the general subjective user experience associated with the use of system. Each participant completed a user experience, system usability scale, and a self-prepared survey form. A significant difference was seen between the results of students and tutors. Due to a lack of training data, the chatbot model predicted incorrect responses that led participants to feel frustrated. In the second part of study, we have retrained the chatbot model using the feedback and designed an error correction approach and engaged seven new medical students to test the chatbot intensively — a total of 2169 user interactions were performed with the chatbot. Of that, 77.4% were properly answered by the bot, 10.8 % were out-of-domain concepts, 8.6 % were unknown concepts (Li et al., 2018), 3.3 % were corrected using the error correction approach designed.

Keywords: virtual patient, medical education, usability evaluation

INTRODUCTION

A large percentage, 41-60% of medical graduates, feel clinically unprepared after university graduation (Cave et al., 2007; Goldacre et al., 2003; Ochsmann et al., 2011) due to the decreasing access of students to real patients. Training is designed based on a back-to-front approach where the student learns about the disease by analysing a diagnosed patient. There are relatively limited opportunities for the students to practice engaging with a patient, using a more natural perspective approach, beginning with a symptom, and concluding with diagnosis and management. Currently, medical schools use both mannequin-based and standardized patient simulation to overcome these limitations. However, the limitations of these approaches are that they can only involve a small number of students at one time, and the faculty have to conduct repeated sessions to cater to the cohort.

In the past two decades, medical education has placed increased reliance on simulation technologies, such as virtual patient (VP) simulations, to boost the growth of learner knowledge and to shape the acquisition of clinical skills for medical students and health professionals (Barry Issenberg et al., 2005). Most of the VPs have, however, been narrative based, using a linear or menu-driven model with preselected options, which are relatively simplistic, predictable, and prescriptive in their approaches limiting the opportunities of the student to engage the virtual patient in a more naturalistic way, to practice his/her decision-making skills. We have designed a virtual patient platform that allows a more natural and realistic way of interaction between students and the virtual integrated patient model. We believe that a well-designed VP, one that is easy and intuitive to use, will help students to remain engaged with exploratory learning, and eventually improve disease understanding and clinical reasoning skills. Therefore, a user-centred approach has been adopted for the development of the VP to ensure that the system is easy and enjoyable to use and meets the pedagogical goals. The system will be subjected to iterative user testing, evaluation and improvement design throughout the development lifecycle. This paper details the first user testing and a study conducted to improve the model performance.

In the next section, we introduce an overview of the design of the virtual integrated patient. Subsequently, the design of the usability study. In section 4, the results of the user study will be discussed. Following which, there

would be a brief overview of an improvement study that was conducted to improve the system further. Through thorough exercises, the system learned a variety of new questions from experienced medical students.

DESIGN OF VIRTUAL INTEGRATED PATIENT

Figure 1. Student performing an interview with a virtual patient

Virtual Integrated Patient (VIP), generates realistic virtual patients which students can interact with using a free-text interface, through the process of interviewing (Figure 1), conducting a physical examination (e.g., to check the temperature, blood pressure, heart rate) and order of investigations (e.g., full blood count, x-ray, MRI, etc.). The patients generated are realistic because they come with randomly generated, rich and comprehensive case details, such as gender, age, ethnicity, presenting symptoms, family, social and medical history, drug history, travel history, and genetics. The engagements allow medical students to prospectively move from presenting symptom to eventual diagnosis, thus allowing the development of clinical reasoning skills.

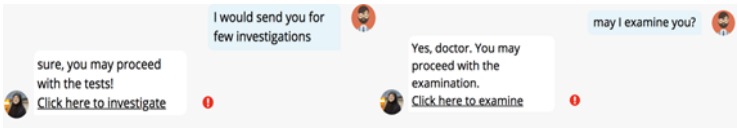

Figure 2. Taking patient consent to perform physical examination and investigation

In clinical practice, clinicians are trained to inform the patients or seek consent from the patients before they perform a physical examination or send patients for investigations. Similarly, on the VIP, students are reminded that they can only get access to perform a physical exam or to order investigations when they have informed the patient or have sought consent from the patient, as described in Figure 2.

The participant subsequently will be allowed to navigate to an examination interface by clicking on the link shown in the patient response within the chatbot. In the examination module, the participant clicks on any part of the human anatomical figure and types in the specific examination they wish to perform. As demonstrated in Figure 3, to check the patient temperature, participants have to input “temperature”. Similarly, to order an investigation, the participant has to inform the patient or seek consent from the patient before they input the specific investigation name.

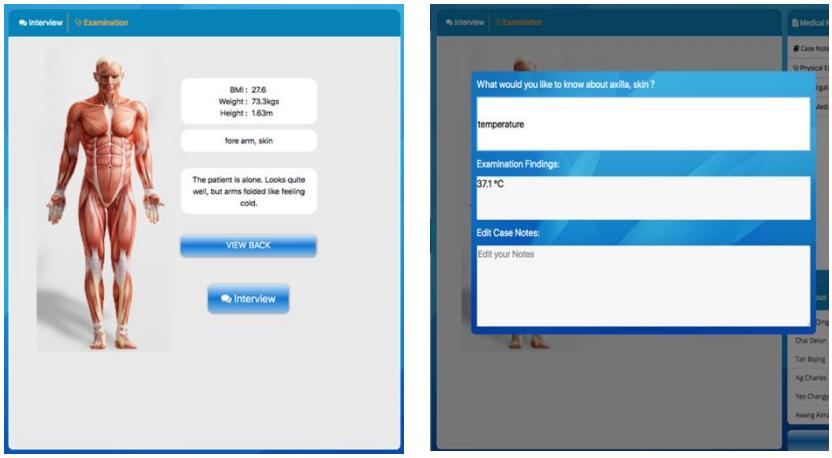

Figure 3. Physical exam: Click on human anatomical figure and search for a test name

Currently, the VIP provides feedback for the cost incurred to the patient based on investigations ordered and efficiency of the student's engagement with the VIP. The latter is calculated based on the number of interview questions asked, examinations, and investigations negotiated before a diagnosis is made. The norms for these assessments are empirically made at the moment but can be better calibrated as more data is harvested according to the seniority of the student. Clinicians are trained to start the interview by greeting the patient, verifying patient name, identity, and encouraged to ask for the patients' consent while asking any sensitive or personal information. The participant is assigned a penalty if they ignore any of these steps and rewarded for compliance. When the participant is ready to make a diagnosis, they simply indicate by clicking *Ready to Diagnose* button (Figure 1) and asked to re-enter brief case notes from the patient interview, physical examination, and investigations. They are then able to make a diagnosis and prescribe a treatment plan for the patient. Subsequently, the patient's actual diagnosis and treatment will be shown for the student to do personal reflection and review.

USABILITY EVAULATION

AIM OF USABILITY EVALUATION

Usability is defined as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" (ISO 9241-11:2018). The objectives of the usability evaluation was to determine how VIPS's user interface and its features affect the usability (efficiency, effectiveness and learnability) as well as the general subjective user experience associated with the use of system.

This was achieved through a user-based evaluation, where the usability and user experience were evaluated using targeted end users of the system, which in this case were medical students. The study also aimed to seek feedback from clinicians (tutors) on the usefulness of the VIP platform as a tool for teaching and training for clinical reasoning skills.

PARTICIPANTS DEMOGRAPHICS

In Phase 1, a total of 13 participants (seven medical students and six clinicians) were recruited to evaluate the usability of VIP. All of them were Singapore citizens, and majority were Chinese (12/13). Eight out of 13 participants were male. Of 7 students, 6 were the second-year medical students, and one was a fourth-year medical student, and their ages ranged from 20 to 23 with an average age of 21. Of 6 clinicians, their ages ranged from 40 to 52 years old, with an average age of 46.5 years old, and their clinical working experience ranged from 15 to 27 years. All the participants are informed to sign a consent form before starting the study. There was an omission of a 4th-year participant during analysis to make a precise comparison between only second-year students and clinicians.

PROTOCOL

Each participant had to complete two patient cases that covered interviewing patient, examining, and ordering lab investigations. The participants were briefed on the aim of the study and viewed the tutorial before starting. Guided

by the facilitator, participants were asked to think aloud while performing the tasks - what they are trying to achieve, how they think they will achieve them, unexpected systems responses, i.e. if something happened which they did not expect or if something they were expecting to happen did not happen. Each participant's interaction with the VIP were logged and video recorded. Thereafter, participants completed a user experience questionnaire, system usability scale, and a self-developed questionnaire. A semi-structured interview was conducted to follow-up on the observations during task performance (such as difficulties encountered, expressions of frustrations, etc) and capturing all participants' subjective feedback on the VIP system.

USER EXPERIENCE QUESTIONNAIRE

The User Experience Questionnaire (UEQ) contains 6 scales (e.g., attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty) with 26 items. The attractiveness scale has 6 items, and all other scales have 4 items. Each item is scaled from -3 to +3. Thus, -3 represents the most negative answer, 0 a neutral answer, and +3 the most positive answer (Schrepp, 2017).

SYSTEM USABILITY SCALE

The System Usability Scale (SUS) consists of 10 items to examine "usability" and "learnability" of a product. All items are measured on a 1 to 5 (1–Strongly disagree, 5–Strongly agree). Calculation of the SUS score is achieved by converting the 1 to 5 scale to a 0 to 4 scale. For items 1,3,5,7 and 9, the score contribution is the scale position minus 1. For item 2,4,6,8 and 10, the contribution is 5 minus the scale position. The overall SUS value is calculated by multiplying the sum of the scores by 2.5 (Brooke, 1996). Thus, the total score ranged from 0 to 100.

SELF DEVELOPED QUESTIONNAIRE

The self-developed questionnaire includes 15 questions to score patient responses, labels on menu items, and popup boxes, consistency of icons, colours used, navigation, organization of items, performance metrics, physical examination and placeholders in input boxes. All questions are measured on a 1-5 scale (1–Strongly disagree, 5–Strongly agree).

RESULTS

In the UEQ, students rated all items positively and the mean score of each item ranged from 0.29 to 2.29, with especially high mean scores ($\geq$ 2.00) on items of enjoyable, valuable, interesting, good, practical and meet expectations. The students rated the low mean scores ($<$ 1.00) on predictable, fast, and leading-edge. The scores of all items were rated lower in clinicians than students. The clinicians rated more negatively on some items, namely fast-slow (-1.33), pleasing-unlikable (-0.67), efficient- inefficient (-1.33) (Figure 4).

The mean scores of the scales of attractiveness, efficiency, and dependability were negative in clinicians (Figure 5). An Independent sample t-test between the students and clinicians showed that there were statistically significant differences in attractiveness, efficiency, perspicuity, dependability, and stimulation between them ($p < 0.05$) as shown in table 1 below.

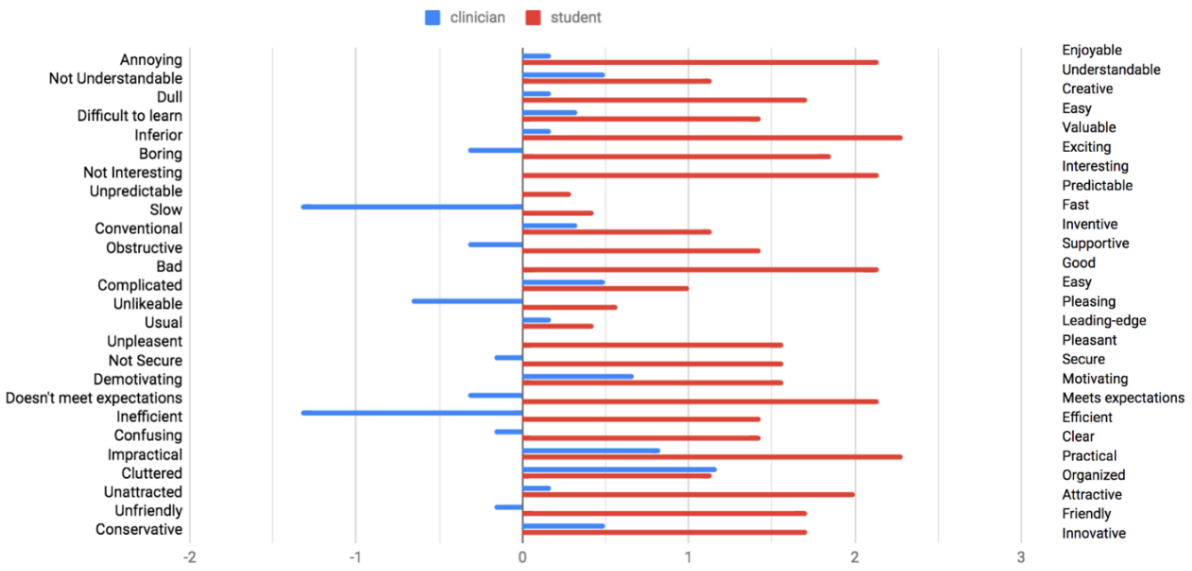

Figure 4. Mean scores of UEQ items between students and clinicians

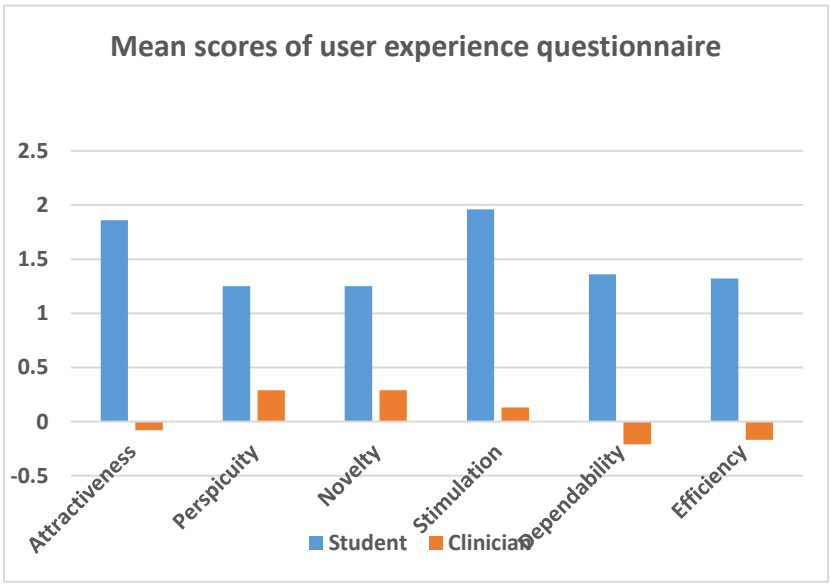

Figure 5. Mean scores of UEQ scales between students and clinicians

Table 1: Independent sample t-test on UEQ scales between students and clinician

	t	df	p-value
Attractiveness	5.698	11	0.000**
Efficiency	4.795	11	0.001**
Perspicuity	2.276	11	0.044*
Dependability	3.923	11	0.002**
Stimulation	5.229	11	0.000**
Novelty	1.983	11	0.073

In System Usability Scale (SUS), the mean score of each item ranged from 2.43 to 3.43 in students and 1.50 to 2.67 in clinicians. The students rated higher mean scores of all items than clinicians. When the comparison of the mean score of each item was performed, the students rated significantly higher scores on items of "I think that I would like to use this platform frequently", "I found the platform unnecessarily complex", "I thought the platform was easy to use", and "I found the various functions in this platform were well-integrated" than those in clinicians. The mean score of total SUS score was 71.43 in students and 53.75 in clinicians, and there was a statistically significant difference in the total mean SUS score between students and clinicians ($p < 0.05$) as shown in Table 2.

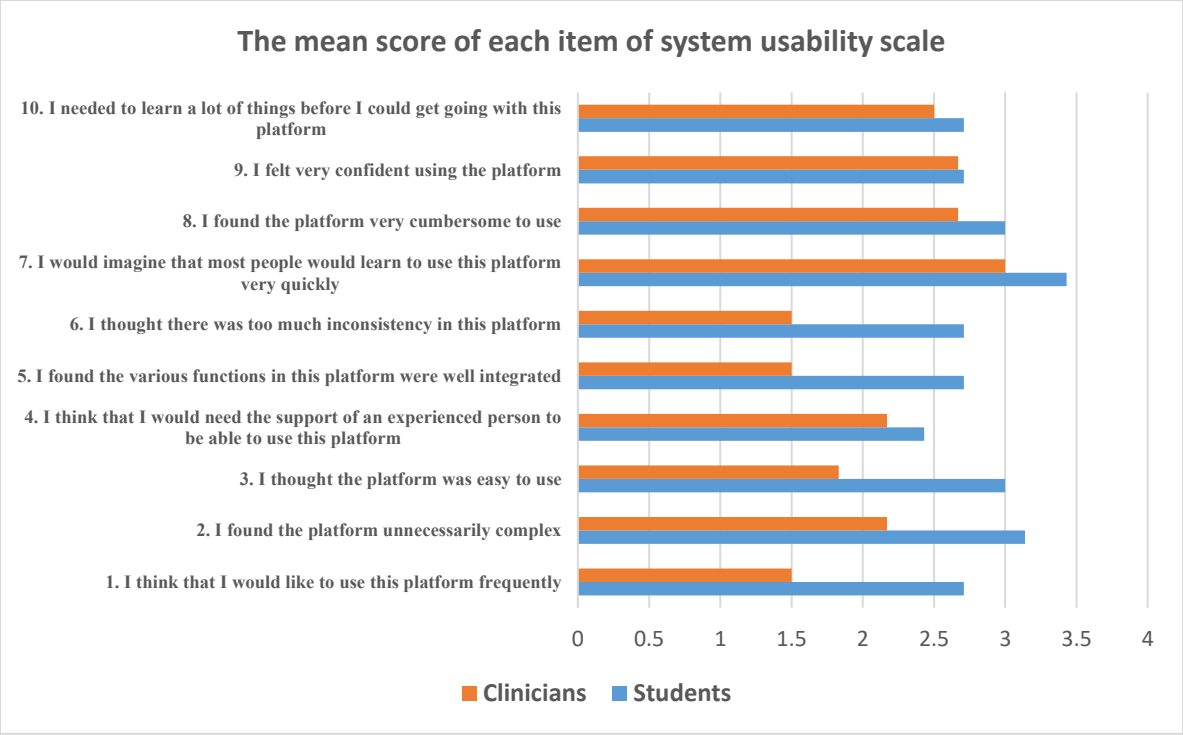

Figure 6. Mean scores of SUS items between students and clinicians

Table 2: Independent t-test on SUS scales between students and clinicians

Variables	t	df	P-value
1. I think that I would like to use this platform frequently	3.261	11	0.008**
2. I found the platform unnecessarily complex	3.029	11	0.011*
3. I thought the platform was easy to use	3.164	11	0.009**
4. I think that I would need the support of an experienced person to be able to use this platform	0.363	11	0.724
5. I found the various functions in this platform were well integrated	3.261	11	0.008**
6. I thought there was too much inconsistency in this platform	1.874	11	0.088
7. I would imagine that most people would learn to use this platform very quickly	1.326	11	0.212
8. I found the platform very cumbersome to use	1.088	11	0.300
9. I felt very confident using the platform	0.109	11	0.915
10. I needed to learn a lot of things before I could get going with this platform	0.355	11	0.729
SUS score	2.541	85	0.027*

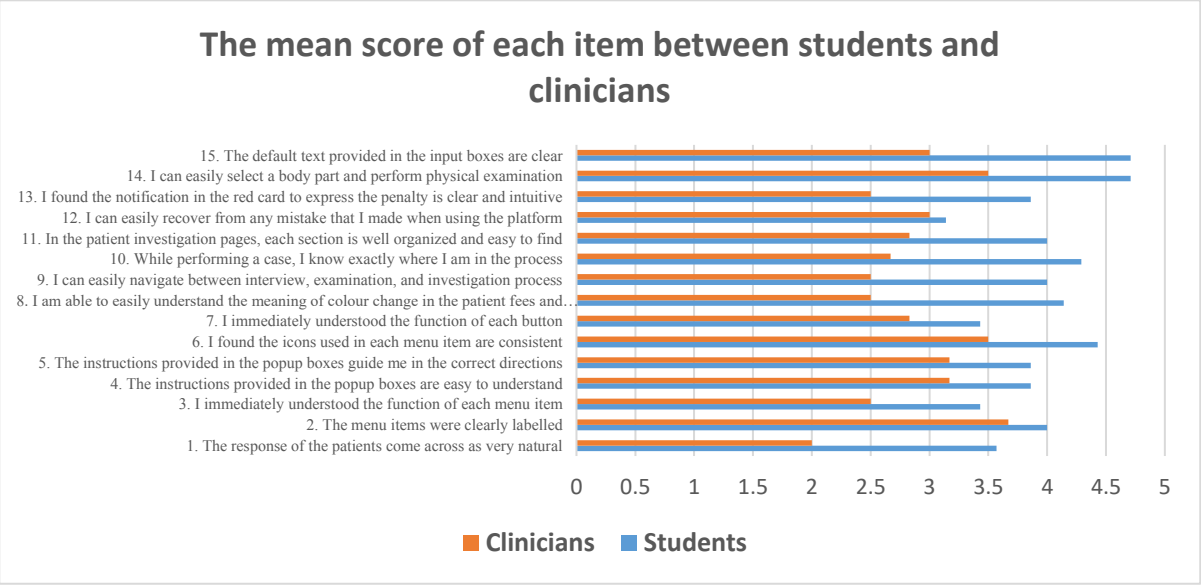

Figure 7. Mean scores of self-developed questionnaire items between students and clinicians

Table 3: Independent t-test on Self developed questionnaire between students and clinicians

Variables	t	df	P-value
1. The response of the patients come across as very natural	3.006	11	0.012*
2. The menu items were clearly labelled	0.734	11	0.504
3. I immediately understood the function of each menu item	1.419	11	0.184
4. The instructions provided in the popup boxes are easy to understand	1.322	11	0.213
5. The instructions provided in the popup boxes guide me in the correct directions	1.726	11	0.112
6. I found the icons used in each menu item are consistent	3.088	11	0.010*
7. I immediately understood the function of each button	0.930	11	0.372
8. I am able to easily understand the meaning of colour change in the patient fees and efficiency	2.422	11	0.034*
9. I can easily navigate between interview, examination, and investigation process	2.901	11	0.014*
10. While performing a case, I know exactly where I am in the process	3.712	11	0.010*
11. In the patient investigation pages, each section is well organized and easy to find	2.340	11	0.039*
12. I can easily recover from any mistake that I made when using the platform	0.196	11	0.848
13. I found the notification in the red card to express the penalty is clear and intuitive	2.302	11	0.042*
14. I can easily select a body part and perform physical examination	3.261	11	0.008**
15. The default text provided in the input boxes are clear	3.750	11	0.003**

Similar to the findings of the UEQ and SUS, the students rated higher scores on all 15 items than the clinicians in the self-developed questionnaire (Figure 7). An Independent t-test comparison between the students and the clinicians indicated that there was statistically significant difference in more than half of items ($p < 0.05$) as shown in Table 3.

The difference between the students' and tutors' usability score can be attributed to the differences in technical proficiency. Medical students comprise mainly of young adults who are more at-home with a chat-based interface, whereas the tutors are less familiar in engaging with online (virtual) patients. Medical students had more positive perceptions of the system as a learning tool and found the system to be engaging and simple to use. The tutors on the other hand found it less useful to them as they were trying to engage with the platform as a student. However, during the interview sessions, they conceded the system as a useful tool to support students' learning.

OBSERVATIONS FROM TASK PERFORMANCE AND FEEDBACK FROM SEMI-STRUCTURE INTERVIEW

Most participants had positive experiences using the platform. One of their positive experiences was that they felt that VIP was a good, fun and an easy and interesting platform to use, even though it could be further improved. Quite a number of participants liked the format and layout of the platform as they felt that it was simple and clear to learn and use. Some participants thought that VIP was a good platform to practice history-taking in a safe online

environment without feeling stressed, as there were no limitations to the questions to be asked. Another positive experience, pointed out by one participant, was that the responses were quite practical and realistic. Thus, the participant felt that the platform was quite flexible. With the platform providing specific responses to each of the participants' questions, participants liked how the programme released one information at a time and further information must be probed by the participants themselves. In addition, for the lab investigations, the participants liked how the programme gave realistic results displaying descriptions of the investigations. The cases seemed more personalised, rather than generic, to the participants.

Despite the positive experiences, some participants also expressed some negative experiences when using the platform. The default model was not able to answer open-ended, long sentence and follow-up questions. Some participants had the conception that the platform was intended to replace the simulated patient or real patient and stated that the interview section was unrealistic. They also complained that they often felt frustrated from the inaccurate responses.

We classified the errors encountered in 3 different categories: unknown concept, out-of-domain and context errors. If the intent of the question asked by the student is already defined in the chatbot model but not predicted correctly, then it is classified as an unknown concept. If the model was not aware of any input, then it is classified as an out-of-domain concept (Li et al., 2018), and if the patient's response shown to the user doesn't follow the context of the current state of the conversation, then it is classified as a context error. Analysing the student logs, we found that 49.7% of the questions were predicted correctly, 18.3% were unknown concepts, 30% of out-of-domain concepts and 2% of issues were due to context. Due to the missing data in the system, the computer cannot provide appropriate responses to the participants' questions, which made the participants feel frustrated. A common issue also occurred when participants presented questions in different ways. In this case, although the system has the data relating to the question, the computer could not recognize the different presentations of the questions, which resulted in many unknown concepts. There were some mapping issues due to the context of the conversation and thus, the computer could not recognise the context questions or follow-up questions, which resulted in the generation of wrong answers.

In terms of interaction design, one of the major problems was navigating from history taking (interview) to physical examination. Participants were required to seek permission from the patients for physically examination in order for the system to move to the physical examination segment. However, most participants looked for a button to click to move to physical examination.

Another common problem encountered was most participants clicked on either the ear, forehead/face and mouth for the measurement of body temperature. However, the temperature was designed to be taken under the axilla. Participants found it difficult to navigate between the physical examination and the interview. When participants tried to order laboratory investigations, they also encountered some problems as there were some unregistered investigations in the system and also costing had an implementation problem where the system provided doubled costs when the student ordered the same investigation twice. Participants expressed their displeasure over some issues in the programme, such as the need to re-type or rewrite a brief of the patient case notes before diagnosing, as they found it tiring to cut and paste the words from the case notes.

REDESIGNING VIP

We have retrained the chatbot model using the usability study feedback. Following are few changes we have incorporated in the design.

1. Consider the student asked the patient *How long you have fever?* for which the platform is not been trained and so the program matched the response for the question *How long ?* as it was similar and the patient responded as *Since 3 days ago* which is for the symptom cough the patient may be suffering from. To avoid these inconsistencies, we have tailored the response more specific to the question i.e., *I got cough since 3 days ago.*

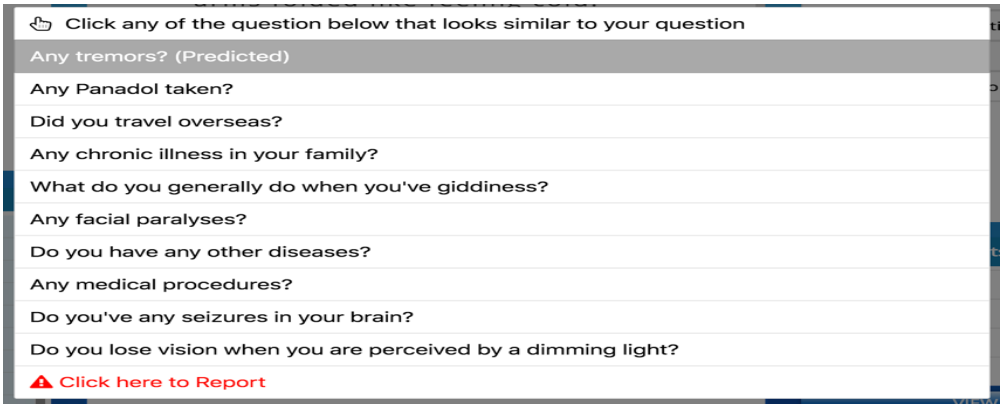

Figure 8. Nearby list of questions given to correct unknown concepts

2. We have used nearby predictions made by the natural language understanding model to correct unknown concepts. Whenever a user asks a question in the chat, the question is detected using a trained language model. Rasa NLU (Rasa Documentation) has been used to deploy the chatbot language model. Rasa NLU predicts a list of intents, each associated with a confidence level. By default a intent with high confidence level is chosen and a patient response for that intent is shown to the user. For example, If the question asked is *Did you take any panadol?* which is focussing on patient medications, then the chatbot incorrectly predicts a intent named *any tremors* with an confidence level 0.97, *medications intake* with an confidence level 0.94. So the intent shown to the user will be *I don't have any tremors* which is incorrect. We have provided a report option next to the patient response, on clicking this icon a nearby list of predictions made by the language model are shown in a popup as shown in figure 8. When user clicks on *Any Panadol taken?* the response *I haven't taken any Panadol* is shown to the user.

USER STUDY TO IMPROVE HISTORY TAKING

In phase 2 of the study, we focused mainly on thoroughly allowing the system to begin learning new questions that clinicians would often use. Ensuring the robustness of the system and allowing usage by people with different levels of medical experience. A total of 9 fourth year medical students were recruited to test the system. Students were instructed to be as thorough as they can be in this experiment. They were given instructions that they should cover only the interview portion, something learned in their second year and using it as a skeletal guideline in their "quest". The viewing of the tutorial was done at the beginning after their consent was taken. Each participant was tasked to complete 4 cases thoroughly. There was an improvement in the performance of the model after incorporating the feedback of the usability study. A total of 2169 user interactions were performed while interviewing the patients. Of them, 1678 (77.4%) were correctly predicted, 71 (3.3 %) unknown concepts can be corrected using the error correction approach described above, 186 (8.6 %) were the unknown concepts not be able to correct using the error correction approach, and 234 (10 %) were out-of-domain concepts.

CONCLUSION AND FUTURE WORK

Medical students feel clinically unprepared after graduation due to lack of access to real patients. Virtual integrated patient allows the student to interact with the patient more naturally and realistically. The VIP generates virtual patients that students can use to rehearse and practice skills to engage a patient through the interview, examinations, and ordering of investigations. The primary interface is through free text. We have evaluated this tool with seven medical students and six clinicians (tutors). Each participant completed three questionnaires (user experience questionnaire, system usability scale, and a self-developed questionnaire). In all the surveys, there was a significant difference between the students' and clinicians' scores. We also noticed that around 30% of the questions asked were new to the platform, resulting to frustration in participants. We incorporated the usability study feedback and designed an error correction for unknown concepts, and found an improvement in the

performance of chatbot after engaging 9 new medical students to intensively test the history taking component of the website.

We believe that the VIP Interview chatbot can be further improved with increased participants interaction. We can use this tool to develop virtual populations of patients according to relevant demographics as well as even more complex patients with multiple clinical episodes, and clinical sequel necessitating follow-ups. We are building this platform to include more customisable features accommodating to the specific needs of different teaching environments, medical schools, and end-users.

ACKNOWLEDGEMENTS

We want to express our sincere gratitude to Professor Teo Boon See, Kong Shu Min Juanita, Lee Yueh Jia in helping to conduct the user study, to all the medical students, tutors signed up for the user study and to Mercy Steven who helped us in the generation of the patient database.

REFERENCES

Barry Issenberg, S., Mcgaghie, W., Petrusa, E., Lee Gordon, D., & Scalese, R. (2005). Features and uses of high- fidelity medical simulations that lead to effective learning: a BEME systematic review. Medical Teacher, 27(1), 10-28. https://doi.org/10.1080/01421590500046924

Benedict, N. (2010). Virtual Patients and Problem-Based Learning in Advanced Therapeutics. American Journal Of Pharmaceutical Education, 74(8), 143. https://doi.org/10.5688/aj7408143

Brooke, J. (1996). SUS: A quick and dirty usability scale. Usability Evaluation in Industry. Cave, J., Goldacre, M., Lambert, T., Woolf, K., Jones, A., & Dacre, J. (2007). Newly qualified doctors' views about whether their medical school had trained them well: questionnaire surveys. BMC Medical Education, 7(1), 38. https://doi.org/10.1186/1472-6920-7-38

ISO 9241-11:2018, Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts. Janet Fleetwood, Wayne Vaught, Debra Feldman, Edward Gracely, Zach Kassutto, and Dennis Novack. 2000.

MedEthEx Online: A Computer-Based Learning Program in Medical Ethics and Communication Skills. Teaching and Learning in Medicine, 12(2):96-104. https://doi.org/10.1207/S15328015TLM1202_7

Rasa Documentation. Rasa NLU for Intent Classification and Entity extraction. https://legacy-docs.rasa.com/docs/nlu/0.13.4/.

Ochsmann, E., Zier, U., Drexler, H., & Schmid, K. (2011). Well prepared for work? Junior doctors' self-assessment after medical education. BMC Medical Education, 11(1), 99. https://doi.org/10.1186/1472-6920-11-99

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). International Journal Of Interactive Multimedia And Artificial Intelligence, 4(6), 103-108. https://doi.org/10.9781/ijimai.2017.09.001

Toby Jia-Jun Li, Igor Labutov, Brad A Myers, Amos Azaria, Alexander I Rudnicky, and Tom M Mitchell. 2018. Teaching Agents When They Fail: End User Development in Goal-Oriented Conversational Agents, pages 119-137. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-95579-7_6

Please cite as: Anubharath, P., Chui, Y.P., Sng, J.C.J., Zhu, L., Tham, K. & Lee, E.J.D. (2019). Usability and user experience evaluation of Virtual Integrated Patient. In Y. W. Chew, K. M. Chan, and A. Alphonso (Eds.), *Personalised Learning. Diverse Goals. One Heart. ASCILITE 2019 Singapore* (pp. 18-28).