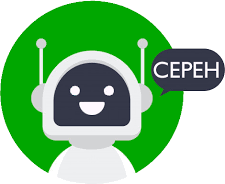
**Feasibility and Acceptance of Chatbots Embedded in Healthcare Curricula:**



Matthew Pears, Eirini Schiza, James Henderson, Natalia Stathakarou, Klas Karlgren, Panagiotis. D. Bamidis, Iraklis Tsoupouroglou, Constantinos. S. Pattichis, and Stathis. Th.

Konstantinidis

CEPEH Report

————

*December*

2022

see CEPEH.eu for more information

This work is supported by the ERASMUS+ Strategic Partnership in Higher Education “Chatbot Enhance Personalise European Healthcare Curricula (CEPEH)” (www.cepeh.eu) (2019-1-UK01- KA203-062091) project of the European Union.

The CEPEH Team

This document details the evaluation of each resource in terms of the feasibility and acceptance from the end-users. There was evidence of identifying the feasibility of such resources into formal training and studies exist on the acceptance of such resources, with promising results. However, all these studies deﬁned the need for further research in the area until the use of chatbots in healthcare education became common. Furthermore, the creation process of CEPEH resources was signiﬁcantly diﬀerent and had improvements to current methods, due to the co-creation process, and use of low cost but eﬀective technology.

List of Figures [vii](#_bookmark0)

List of Tables [viii](#_bookmark1)

List of Abbreviations [ix](#_bookmark2)

Introduction [1](#_bookmark3)

Background [2](#_bookmark4)

1. Method [4](#_bookmark5)
   1. Participants [4](#_bookmark6)
   2. Procedure [5](#_bookmark8)
   3. Design [6](#_bookmark9)
   4. Materials and Measures [7](#_bookmark10)
2. Results [11](#_bookmark17)
   1. Participants’ Characteristics [11](#_bookmark18)
   2. Chatbot Usability Questionnaire (CUQ) [13](#_bookmark20)
   3. System Usability Scale (SUS) Scores [16](#_bookmark26)
   4. Technology Acceptance Model [17](#_bookmark27)
   5. Inferential Statistics [22](#_bookmark34)
   6. Other Findings [23](#_bookmark37)
3. CEPEH Focus Group Discussion Analysis [27](#_bookmark41)
   1. Tokenizing [27](#_bookmark42)
   2. Plotting word frequencies - bar graphs [28](#_bookmark43)
   3. Word clouds [29](#_bookmark44)
   4. Sentiment analysis [31](#_bookmark45)

[Contents](#_TOC_250001)

1. Discussion [33](#_bookmark46)
   1. Quantitative Results [33](#_bookmark47)
   2. Qualitative Results [34](#_bookmark51)
   3. Limitations [34](#_bookmark52)
   4. Conclusions [34](#_bookmark53)
2. (Additional Analyses) Training Events [36](#_bookmark54)
   1. CEPEH Training Event C1 [36](#_bookmark55)
   2. CEPEH Training Event 2 [37](#_bookmark56)

**Bibliography** [39](#_bookmark57)

Appendix [40](#_bookmark58)

References [43](#_bookmark60)

##### List of Figures

* 1. Location and Profession of Participants [5](#_bookmark7)
  2. Chatbot Usage History- Pre [13](#_bookmark21)
  3. Chatbots are Useful Opinion- Pre [14](#_bookmark23)
  4. CUQ CEPEH Score [14](#_bookmark24)
  5. CUQ Scatter Plot [15](#_bookmark25)
  6. Improvements in Knowledge [18](#_bookmark28)
  7. Trust Chatbots POST use [19](#_bookmark30)
  8. Ease of Use Comparison [22](#_bookmark33)
  9. Ease of Use Comparison [23](#_bookmark36)
  10. Intend to Reuse-Post [24](#_bookmark38)
  11. Easy to Use- Post [25](#_bookmark39)
  12. Table of Results [26](#_bookmark40)
  13. TAM Model processes [34](#_bookmark49)

**CEPEH** Chatbot Enhance Personalised European Healthcare curricula

**HELM** Health E-Learning and Media team

**RLO** Reusable Learning Object

**ASPIRE** Aims Storyboarding Production Implementation Release Evaluation

**NLP** Natural Language Processing

**NLU** Natural Language Understanding

**TAM** Technology Acceptance Model

**SUS** System Usability Scale

**CUQ** Chatbot Usability Questionnaire

**HIG** HELM is Great

Personalised Healthcare Education is needed to meet growing demand and quality maintenance. There is growing evidence around chatbots, namely machine conversation systems- these programs have the potential to change the way students learn and search for information.

Chatbots can quiz existing knowledge, enable higher student engagement with a learning task, or support higher-order cognitive activities. In large-scale learning scenarios with a high student-to-lecturer ratio, chatbots can help tackle the issue of individualized student support and facilitate personalized learning. However, limited examples of chatbots in European Healthcare Curricula have been utilized to combine both the continuum of cognitive processes presented in Bloom’s taxonomy, with the idea that some repetitive tasks can be done with a chatbot- to provide greater access or to scale faculty time.

Thus, CEPEH strategic partnership has co-created open access chatbots utilizing artiﬁcial intelligence, promoting innovative practices in digital era, by supporting current curricula and fostering open education.

CEPEH Erasmus+ strategic partnership aimed to co-design and implement new pedagogical approaches and, in particular, chatbots for European medical and nursing schools. CEPEH used use participatory design to engage stakeholders (students, healthcare workforce staﬀ, lecturers, clinicians, etc.) in order to co-design eﬀective chatbots and release them as open access resources. Through CEPEH, eﬀective use of digital technologies and open education were be incorporated into healthcare curricula. This enabled students to increase their health and medical related skills through ﬂexible learning.

CEPEH expected that students adopted this new digital pedagogy and im- prove their skills and competences through ﬂexible personalized learning, while the teaching staﬀ enhanced their e-learning tool co-creation competences and make use of co-design best practices and recommendations for use. It is also expected increased cooperation between the partners. Thus, in the long term, CEPEH expects to inﬂuence the development of medical and nursing curricula with this digital innovation, foster the quality of the future healthcare workforce and further improve international competitiveness of the partners’ healthcare curricula. This document details the evaluation of the resources created by the CEPEH team.

The evaluation speciﬁcally explored the feasibility and acceptance from the end-users. These end-users are learners in European healthcare higher education institutions.

There was ﬁrstly evidence for the need to identify the feasibility of chatbots and similar resources into formal education and training, with a further need to improve access to these types of learning resources. Of course, studies exist on the acceptance of chatbots, virtual patients, and many other healthcare applications, with promising results. However, through various limitations, we believed there was further research to be completed to accelerate the design, development, implementation, and evaluation processes. These have ﬁnancial, stakeholder, time, and eﬃcacy beneﬁts. The creation process of CEPEH resources was signiﬁcantly diﬀerent to most in the literature, and this report highlights the approach of the CEPEH team towards enhancing personalized healthcare education can be achieved.

###### Background

The working practices of CEPEH are aimed at maximizing eﬃcacy of these chatbots as learning resources and provided a sense of shared development and ownership from all stakeholders. The process normally begins with workshops in which the project is scoped, and team building occurs. The CEPEH workshops involve the widest possible team of stakeholders including tutors, students, healthcare workers, learning technologists, health service users and carers depending on the materials being created. For readers who are interested in using these high-quality digital resources please access them for free at CEPEH.EU. The next section will now present the evaluation of all CEPEH chatbot re- sources.

##### Method

**1**

**Contents**

* 1. **Participants** [**4**](#_bookmark6)
  2. **Procedure** [**5**](#_bookmark8)
  3. **Design** [**6**](#_bookmark9)
  4. **Materials and Measures** [**7**](#_bookmark10)
     1. Chatbot Usability Questionnaire (CUQ) . . . . . . . . . [7](#_bookmark11)
     2. UTAUT2 (Uniﬁed Theory of Acceptance and Use of Technology) . . . . . . . . . . . . . . . . . . . . . . . . . . . [7](#_bookmark12)
     3. System Usability Scale . . . . . . . . . . . . . . . . . . . [8](#_bookmark13)
     4. Computer Self-Eﬃcacy Scale Tool . . . . . . . . . . . . [8](#_bookmark14)
     5. Technology Acceptance Model (TAM) . . . . . . . . . . [8](#_bookmark15)
     6. Qualitative Measure- Focus Group Discussions . . . . . [9](#_bookmark16)

###### Participants

This dataset had 14 males and 28 females therefore a total of 42 participants. It was a repeated measure design whereby each participant used the 4 chatbots developed by the CEPEH team. Therefore, there are 42 points of data in the condition before testing, and 126 data points after testing the chatbots- for a total of 168 row of data. There were 78 questions asked in total, therefore the full dataset had approximately 6000 cells recorded.

Chart

Description automatically generated

There were 22 females and 7 males from Greece. There were 3 females and 4 males from Cyprus. There were 2 females and 2 males from Sweden, and there were 2 participants from the United Kingdom (see ([1.1](#_bookmark7)).

The majority 36 participants were student, with 3 being learning technologists, 2 were lecturers, and 1 was a doctor. Although there could be a diﬀerence in these groups, the design was within- groups therefore each participants pre-usage metrics were the comparative control data, and participant diﬀerences did not aﬀect the evaluation.

###### Procedure

For each resource created by the Partners, the same experimental methodology was followed. For each resource created by partners, students performed a study within an online or face to face workshop or course. Student participants joined

from Greece, Cyprus, Sweden, and the United Kingdom. A repeated measures design was used as the same group measures were taken before and after usage of the chatbots. They were recruited via staﬀ members in the CEPEH group.

Participants were asked prior to the study if they agree to participate, providing them with a PIS form. Participants had the opportunity to discuss with the research team prior to the study and before consent is given. Then, participants used the chatbot resources independently and technical support was provided. Finally, post-intervention measures were recorded. Some of the participants were invited to participate in Focus Group Discussions (FGD), and each FGD lasted between 15 to 25 minutes, with 5-10 participants. Participants were asked if they would like to be informed of the ﬁndings of the study.

###### Design

The data captured from the participants were their initials and numerical day of birth, used as anonymous identiﬁer for pre-post analysis. Their institution was captured (Aristotle University of Thessaloniki, CYENS Centre of Excellent, Karolinska Institute, and The University of Nottingham), and Sex (Male/Female/Other). Before any interaction with the learning resources, various perceptions of chat- bot such as conﬁdence and ease of use, usefulness, Inﬂuence from others, and current learning resources (videos, textbooks, friends etc.), were captured. Descriptive data was produced alongside t-tests. Repeated measures t-tests were the appropriate test to use as this explores diﬀerences between groups, there were no covariates, and we did not have several dependent variables. There was one independent factor being Chatbot use having 2 levels (pre/post). There were 3 chatbots therefore there was option for ANOVA to determine where diﬀerences lie if statistical diﬀerences were found however this was not wholly appropriate for the data type and not necessary for pre-post comparison.

###### Materials and Measures

The measures used ﬁt within a newly developed Chatbot Evaluation Framework- which takes the best measures of 5 previous frameworks. Denecke and Warren

[2] derived several quality dimensions and attributes from previous chatbot literature. They formed six perspectives from their review of articles and mobile health applications.

These six perspectives were: 1) Task-oriented, 2) Artiﬁcial intelligence, 3) System quality perspective, 4) Linguistic perspective, 5) UX Perspective, 6) Healthcare quality perspective.

To capture these perspectives, we used several validated materials that can distinguish these elements of the CEPEH chatbots.

* + 1. Chatbot Usability Questionnaire (CUQ)

The Chatbot Usability Questionnaire (CUQ) [4] is a new questionnaire speciﬁcally designed for measuring the usability of chatbots by an interdisciplinary team from the Ulster University. CUQ can be used alongside the prevalent System Usability Scale Score (SUS) [5]. Multiple metrics are more appropriate when measuring usability of chatbots [6] therefore a combination of two scores can provide an all-inclusive overview.

* + 1. UTAUT2 (Uniﬁed Theory of Acceptance and Use of Technology)

The underpinning theory of the UTAUT2 is that there are four key constructs to the intentions of using technology-based resources: 1) performance expectancy, 2) eﬀort expectancy, 3) social inﬂuence, and 4) enabling conditions.

The TAM and the UTAUT2 have cross over in measuring technology acceptance, however the UTAUT2 has more applied probing questions. Few studies exist that use technology acceptance theories for the intention to use products that explicitly incorporate AI. A recent extension of the UTAUT2 model added ﬁve (health,

convenience comfort, sustainability, safety, security, and personal innovativeness) additional inﬂuencing factors to accommodate for AI [7]. This can be used for products in either health, household use, or mobility and can help to explain behavioural intention and use behaviour of chatbots.

* + 1. System Usability Scale

The System Usability Scale (SUS) was used [10] and is a widely used and adopted usability questionnaire. It is popular due to its unbiased and agnostic properties, a nonproprietary, and a quick scale of 10 questions. However, as there are the CUQ and parts of the UTAUT2 we have selected only 2 questions which do not cross-over with the other measures. These are important statements however and good indicators of usability when assessed with the other results.

* + 1. Computer Self-Eﬃcacy Scale Tool

The 10 question CSEST is based on the 32-item questionnaire by Murphy, Coover, and Owen (1989). It can be adapted for any technology, and we have selected only a few pertinent questions. Participants are asked to think about using the CEPEH chatbots and answer that they would use the chatbots if I had never used a product like it before; If they could call someone for help if I got stuck, or if someone showed them how to do it ﬁrst, and other similar usage questions.

* + 1. Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM) [1] was speciﬁcally developed with the primary aim of identifying the determinants involved in computer acceptance in general; secondly, to examine a variety of information technology usage behaviours; and thirdly, to provide a parsimonious theoretical explanatory model.

TAM suggests that attitude would be a direct predictor of the intention to use technology, which in turn would predict the actual usage of the technology. The only modiﬁcation to the nine sub-scales of the questionnaire consists of applying

the items to the context of chatbots. All the items, except those measuring attitudes, utilize a seven-point Likert scale ranging from “strongly agree” to “strongly disagree” with a middle neutral point [2].

The nine sub-scales of the questionnaire:

Ease of use of chatbots Perceived usefulness of chatbots Intention of use. Attitude toward usage of chatbots. Perception of personal eﬃcacy to use a chatbot resource. Perception of external control toward chatbots. Anxiety toward chatbot use. Intrinsic motivation to use chatbot resources. Perceived costs of chatbots.

* + 1. Qualitative Measure- Focus Group Discussions

Focus groups are a pervasive means of market research and provides credible acceptance evaluators regarding the penetration that a product or service will have on a target demographic. Focus groups are a form of qualitative research consisting of interviews or structured discussions, in which a group of people are asked about their perceptions, opinions, beliefs, and attitudes towards a product, service, concept, advertisement, idea, or packaging.

Questions are asked in an interactive group setting where participants are free to talk with other group members. During this process, the researcher either takes notes or records the vital points he or she is getting from the group. Researchers select members of the focus group carefully for eﬀective and authoritative responses. Relevant stakeholders, then, can use the information collected through focus groups to receive insights on a speciﬁc product, issue, or topic focus [7]. A series of short focus group sessions identiﬁed the feasibility of CEPEH re- sources for formal curricular integration. These sessions, spanning no more than 1-1.5 hours and consisting of no more than 5-7 persons each explored all axes of curricular integration such as accessibility in the classroom, use case scenarios, technology requirements for curricular integration etc. These axes were formalized by the research team, in each evaluation site, to consider the curricular details of each institution.

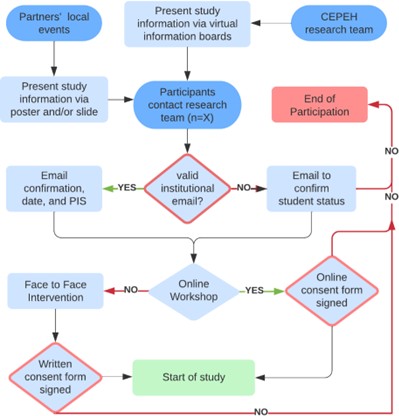


Figure 1: Flow diagram of the recruitment process

##### Results

**2**

**Contents**

* 1. **Participants’ Characteristics** [**11**](#_bookmark18)
  2. **Chatbot Usability Questionnaire (CUQ)** [**13**](#_bookmark20)
     1. CUQ Calculation tool [13](#_bookmark22)
  3. **System Usability Scale (SUS) Scores** [**16**](#_bookmark26)
  4. **Technology Acceptance Model** [**17**](#_bookmark27)
     1. Knowledge and Trust after Use [18](#_bookmark29)
     2. Personality and Interactions [21](#_bookmark31)
     3. Ease of Use and Seeking Support [21](#_bookmark32)
  5. **Inferential Statistics** [**22**](#_bookmark34)
     1. Repeated Measures T-test results [22](#_bookmark35)
  6. **Other Findings** [**23**](#_bookmark37)

###### Participants’ Characteristics

When participants were asked if they have used a chatbot before, 23 stated they had never used a chatbot. Further, 19/42 stated having used a chatbot at least once for between 0-4 hours of use in total. These are likely commercial/website- based assistant chatbots however there are some medical/healthcare resources used in anatomy and/or patient interactions. One individual had spent much longer time with usage- this was the mature student.

**Table 2.1:** Previous Chatbot Usage of Participants

|  |  |
| --- | --- |
| Previous Chatbot Usage | n |
| 1-4 hours | 15 |
| 10-19 hours | 1 |
| 20+ hours | 1 |
| 5-9 hours | 2 |
| Never | 23 |

In short, approximately 50% had never used a chatbot, and 45% had used a chatbot, at some period over the years, for a short period of time. Most learners use books or online books as resources. They may use multiple sources however they were asked to note the primary source. Only 6 stated their primary sources were *Online videos/interactive materials* which includes such tools as chatbots. The boxplot next (2.1) shows learners perceptions of ease of use of mobile app and other educational mobile resources.

***Chart, waterfall chart

Description automatically generated***

(2.2) shows the opinions of all participants on the usefulness of chatbots. Many had not had experience with them yet had positive rating. This positive opinions of chatbots may be from colleagues, friends, media, tutors, or other social information of the beneﬁts in healthcare education. Around 25% were neutral or disagreed that healthcare chatbots were useful.

***Chart, waterfall chart

Description automatically generated***

***The participants then used the 4 chatbots and completed the post- usage survey after each chatbot. Results after use are as followed:***

###### Chatbot Usability Questionnaire (CUQ)

* + 1. CUQ Calculation tool

The CUQ was developed by researchers at Ulster University and as the calculation can be complex, a dedicated calculation tool has been created. Please download the CEPEH CUQ calculation tool which has all of the data entered, so you can see the CEPEH CUQ scoring

[Click here to download CUQ calc tool](file:///E:\CEPEH%20Rmarkdown\data\CUQ-Calculation-Tool.xlsx)

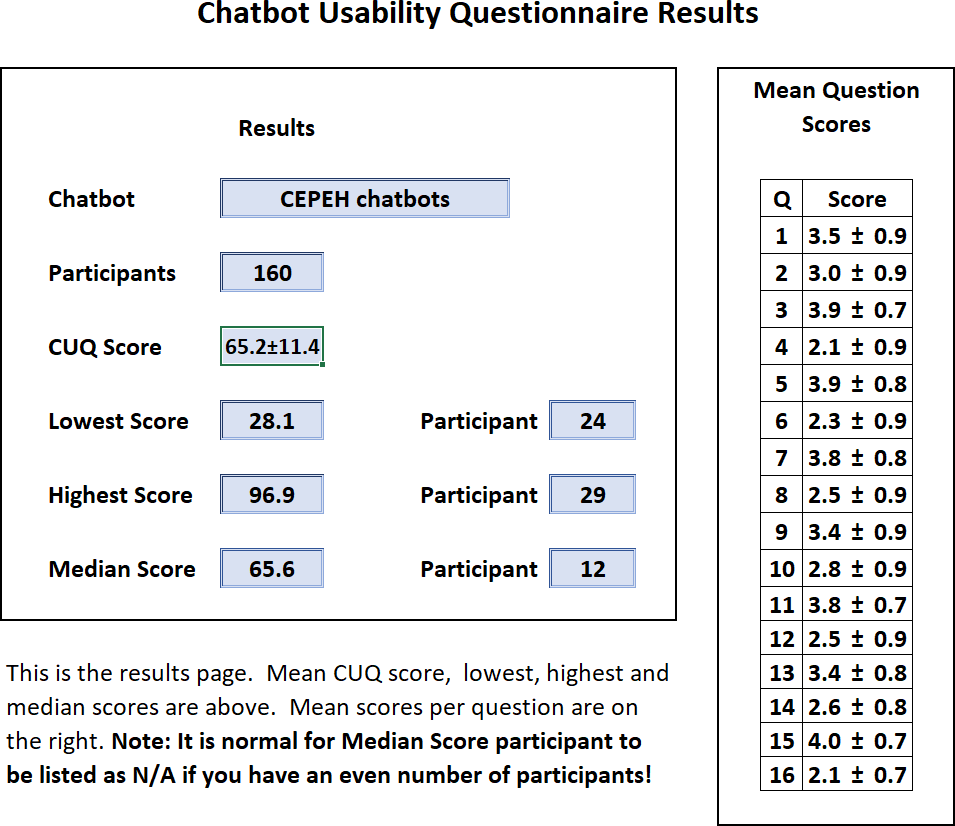
[Click here to download CEPEH CUQ score result](file:///E:\CEPEH%20Rmarkdown\cuq.png)

Although the design and development were similar, each chatbot CUQ score was calculated to understand how the topic content may affect usability:

The breakdown of the chatbots was:

* Aristotle University of Thessaloniki CUQ score = 63/100
* CYENS Centre of Excellence CUQ score = 67/100
* Karolinska Institute CUQ score = 63/100
* University of Nottingham CUQ score = 68/100

The score for all 3 chatbots grouped was 65/100. See Discussion CUQ section for interpretation.



**Figure 2.3:** CUQ CEPEH Score

Figure ([2.4](#_bookmark25)) shows the CUQ scores as a scatter plot to highlight how there was a moderate distribution of results. Further exploration is required to understand which elements are causing this spread, and if it was due to problems within a small group of learners.

Chart, scatter chart

Description automatically generated

###### System Usability Scale (SUS) Scores

*Note= The amount of ‘agreement’ is deﬁned as the addition of ‘Agree’ and ‘Strongly agree’ responses.*

The SUS score should consist of 10 items. However, some SUS questions were improved upon by 1 or more CUQ questions, speciﬁcally to this Chatbot study. The SUS results would be obscured by the CUQ scores, expect 2 that did not have cross-over. The two questions were:

* I would like to use the CEPEH chatbot I tested, more frequently (SUS1) (post)
* I felt conﬁdent using the CEPEH chatbot (SUS2) (post)

This meant the score of the SUS was not created, however the CUQ score better represented the Learners’ perceptions of the CEPEH chatbot in terms of feasibility of use and acceptability in healthcare curricula.

|  |  |
| --- | --- |
| Keep Using CEPEH Chatbot | Responses |
| Agree | 66 |
| Disagree | 15 |
| Neutral | 17 |
| Not Applicable | 3 |
| Strongly Agree | 23 |
| Strongly Disagree | 2 |

The table above shows the results for agreement participants may continue to use the CEPEH chatbots: 89/126 (70%) agreed or strongly agreed. However, there were 23 records that learners were neutral or disagree they would continue use.

|  |  |
| --- | --- |
| Conﬁdence using CEPEH Chatbot(s) | Responses |
| Agree | 71 |
| Disagree | 11 |
| Neutral | 21 |
| Not Applicable | 4 |
| Strongly Agree | 19 |

Conﬁdence when using the chatbots is in table (**??**)- it shows the distribution of agreement for participants for all 4 chatbots. The table shows 90/126 records that participants feel they are conﬁdent in using the chatbots. However, 21/126 (16%) were neutral and 11/126 (8.5%) disagreed and this was explored in the qualitative analysis section.

###### Technology Acceptance Model

The TAM questions were analyzed according to their subsets. The subsets were Perceived Usefulness (PU) and Perceived Ease of Use (PEU)

The questions were: Perceived Usefulness (PU):

1. Using CEPEH chatbots would enable me to accomplish tasks more quickly
2. Using CEPEH chatbots would increase performance
3. Using CEPEH chatbots would increase my productivity
4. I would ﬁnd CEPEH chatbots useful on my course Perceived Ease of Use (PEU):
5. Learning to use CEPEH chatbots would be easy to me
6. It would be easy for me to be skillful at using CEPEH chatbots
7. My interactions with CEPEH chatbots would be clear and understandable
8. I would ﬁnd CEPEH chatbots easy to use

*Results*

The scores as a percentage of agreement, were calculated by averaging the subsets and interpreted as:

* + Before using the CEPEH chatbots, there was 66% (2.2/5) agreement for the Perceived Usefulness of chatbots in healthcare education, and after 48% (2.6/5) agreed.
  + Before using the CEPEH chatbots, there was 64% (2.3) agreement for Perceived Ease of Use of chatbots in healthcare education, and after 51% (2.56) agreed.

The justiﬁcation for this may be due to being early versions of applications with limited functionality and functions which can be diﬃcult for user to experience the intended further range of features and learning exercises.

* + 1. Knowledge and Trust after Use

CYENS chatbot had around 10 more participants stating that they were neutral on gaining knowledge of the topic. The ﬁgure 2.6 shows the ratings by participants of the CEPEH Chatbots to provide them with the necessary course information.

Chart, waterfall chart

Description automatically generated

This is an integral element in learners’ motivational and educational choices to reuse the learning resources. As previously described, the trust of the information is also a factor in these responses.

Chart, bar chart

Description automatically generated

* + 1. Personality and Interactions

Chart, bar chart

Description automatically generated

*The chatbot seemed too robotic* results had the largest mix of responses, and for all 4 chatbots evaluated. The University of Nottingham Cybersecurity chatbot had more deterministic pathways with exploitation of the NLP modelling to provide illusion of realism. This may explain why there was less agreement. However, Neutrality and/or agreement was not desired.

Chart

Description automatically generatedThere were mixed results for the chatbot used being realistic and engaging. This question has two descriptive terms however based on the other results we under- stand that the chatbots’ NLP logic, or ability to respond required improvement to be more ‘smooth’ in replying. The primary limitation was found in the ‘robotic’ interactions (See Figure x). This was investigated further in the ‘Text Mining’ and ‘Sentiment Analysis’ sections.

* + 1. Ease of Use and Seeking Support

After usage, there was only agreement in Ease of Use- as shown in ([2.7](#_bookmark33) as there are no ‘Neutral’ or disagree columns. Any learners with disagreement before using the CEPEH chatbots, after believed they were easy to use.

Those who disagreed or were neutral in the pre usage measure, improved their understanding that help was available with the CEPEH chatbots. After usage, 40

participants agreed they could get help if they had diﬃculty using the resources.

Chart, bar chart

Description automatically generated

Chart, bar chart, waterfall chart

Description automatically generated

After using the CEPEH chatbots, majority of participants stated they would reuse the chatbots. However, there was 6 counts of *disagree* or *strongly disagree* for all 4 chatbots. Further, there were 17 counts of neutral in reuse, which was approximately 4 participants per chatbot (see ([2.9](#_bookmark38)).

For CYENS, even though the knowledge of the topic was not perceived to improve by some participants, this box plot shows how 34/42 stated they would reuse the chatbot developed by CYENS. There was only 1 ‘Strongly Disagree’ response. The agreement options counted for the majority of the data.

###### Inferential Statistics

Paired t-test involves matching the same participants on a variable before intervention with, ideally, the equal measure of variable after intervention. For this study, we used several metrics from the various questionnaires to facilitate pre-post comparisons. The CUQ was only asked after chatbot usage, however most other questions were able to have a pre-post comparison.

Importantly, there is value in the modes which indicate majority consensus, rather than mean driven t-tests. Being an initial single session with technical orientation for Users as well as practical usage, there is high change that a minority will experience technical or functional problems. Although the majority may benefit, the measures from this minority can significantly impact the measures. For example, if a sample of 42 have the results 5/5 (27), 4/5 (11), and 1/5 (5), the mode and median are 5 however the mean is 4.38. This can affect parametric or non-parametric results that infer equal experience from participant. We are not overlooking participants but factoring in their experiences as a minority and assessing their results in other ways- i.e., the focus group discussions. This better reflects each participants experience and the accuracy of efficacy of the chatbots.

Because of the experimental setup and the 4 different chatbots, the meaning of the t-tests has small power and effect size. We intended for all metrics to improve, but to have significant findings with the setup does not provide much more information in addition to increased means/scores. We performed paired t-tests, however many paired samples had significant Shapiro-Wilk results indicating the normality assumption were violated. The test is robust to violations, and the same participants are being tested over Time, rather than different participants. This also indicated that the meaning of the Wilcoxon has limited strength.

* + 1. Paired sample t-tests and Wilcoxon signed rank test

Two-tailed paired t-tests were conducted to examine whether the mean difference is of the following groups were significantly different:

* Confidence
* Daily usefulness
* Increasing achievements
* Accomplishing things quickly
* Increased productivity
* Ease of use
* Clear and understandable interactions
* Use chatbots more frequently

The results showed there were no significant differences in these comparisons. For each comparison the results were:

* Confidence- *t*(24) = -0.35, *p* = .731 (*pre*m=1.96, *post*m=2.04)
* Daily usefulness- *V* = 48.50, *z* = -0.27, *p* = .790 (*pre*m=2.04, *post*m=2.12)
* Increasing achievements- *t*(24) = -0.18, *p* = .857 (*pre*m=2.28, *post*m=2.32)
* Accomplish tasks quickly- *V* = 36.00, *z* = -0.25, *p* = .805 (*pre*m=2.12, *post*m=2.16)
* Increased productivity- *V* = 96.00, *z* = -1.51, *p* = .131 (*pre*m=2.6, *post*m=2.24)
* Ease of use- *V* = 72.00, *z* = -1.32, *p* = .186 (*pre*m=2.36, *post*m=2.12)
* Clear and understandable interactions- *V* = 101.50, *z* = -1.82, *p* = .068 (*pre*m=2.2, *post*m=2.61)
* Use chatbots more frequently- *t*(24) = 0.45, *p* = .657 (*pre*m=2.2, *post*m=2.08)

As predicted, the sensitivity of the t-test meant Wilcoxon test was more appropriate for some measures. The results show minor increases in means for some, but minor decreases in others. These results have high standard deviations which are from a minority of participants scoring low, and explored in the focus group discussions.

###### Other Findings

###### Chart, waterfall chart Description automatically generated

Table

Description automatically generated

This rather large table presents all of the descriptive statistic measures AFTER chatbot usage. The Mode is important to show that the majority of participants stated their perceptions, experience, and acceptance increased.

##### CEPEH Focus Group Discussion Analysis

**3**

The focus group discussions provided a lot of feedback for how the participants experienced their interactions with the chatbots, and how the CEPEH team can improve them, improve the design and development processes, and improve up- take and sharing.

One method of analyzing this data is with use of text mining and data manipulation, creating word clouds, sentiment analysis, and using a model which can distinguish the unique themes in text, and highlights for us what text is used to create these themes.

Therefore, we have created a model to allow eﬃcient and intelligent analysis of this open/free focus group data.

###### Tokenizing

Firstly, we tokenized the words from the FGDs. A Token is “a meaningful unit of text, most often a word, that we are interested in using for further analysis”. For each word we give it a property that we can call upon later.

The data manipulation for this included removing punctuation, converting to lower-case, and setting word type to word (and not such types as “characters”, “ngrams”, “sentences”, “lines” etc.)

1. *CEPEH Focus Group Discussion Analysis*
   * 1. Stop Words

The model then removed words with meaningless function. These are called stop words. Words like “the”, “of” and “to” are the most frequent words found, technically, but are of little interest to us.

We also created a custom list of stop words for CEPEH. We know participants may mention other objects, and the list was as followed: found; chatbot; chatbots; presentation.

The data was ready for analysis by the model. We ordered it to ﬁnd the most frequent words. Below is a table with the 6 frequently occurring words, showing how Stop words have now been ﬁltered.

|  |  |
| --- | --- |
| word | n |
| information | 11 |
| helpful | 8 |
| understand | 8 |
| idea | 7 |
| ideas | 7 |
| lot | 7 |

This word list can then be used for sentiment analysis, (see *Sentiment Analysis*

section), in addition to frequency of words.

###### Plotting word frequencies - bar graphs

* + 1. Normalized frequency

With this information a list of top words from the participants in the FGD can be rendered and after some modiﬁcations, a graph of the top 20 words is produced, with better aesthetics. This is a better way to understand this data, and the axis can be normalized for the frequency of occurrences in accordance with the source text. The raw text had 2827 words in total. Therefore, we can mutate the ratios to reﬂect this.

* + 1. Plotting Normalized frequency

Now we can plot, for example, the 20 most frequent words when normalized by the source text.

Chart, bar chart

Description automatically generated

In summary, this understanding of frequent words can help to understand common concurrences and extrapolate to a larger audience. If scope and impact of CEPEH chatbots increased we can understand the type of themes and trends may occur, based on such FGD analysis.

###### Word clouds

To visualize the most frequent words in another format, below is a word cloud which presents the word size to indicate the frequency- words that occur more often being displayed in a larger font size. This has a normalized data frequency in accordance with the FGD source document analysis.

Text

Description automatically generated

We understand the context has been reduced for each word. However, in general there can be categorized positive/negative words from the word cloud: Positive words are- beneﬁt, practical, nice, helpful, learn, ideas, and enjoyed Negative words are- diﬃcult, test (who likes a test?), don’t, and ‘lot’ may be negative if there is a ‘lot’ of information.

* + 1. The vocabulary of Texts

Here is a graph that has plotted the words in places depending on the word frequencies. Additionally, colour hotspots shows how diﬀerent the frequencies are

- darker items are more similar in terms of their frequencies, lighter-colored ones more frequent in one text compared to the other.

Chart, scatter chart

Description automatically generated

###### Sentiment analysis

What is the sentiment of all participants? What are types of emotional words are being used? The preparation of these words has some use in understanding the frequencies, but their emotional valence are not compared. The table above has the word *‘helpful’* which has a positive connotation, however there are 386 words, with many having several occurrences.

As the table below shows. the FGD data has been analyzed for sentiment of each word and has been calculated to have 62 positive emotional valences of words, with 24 negative valences of words. These are from a **Bing sentiment lexicon** which is the most popular English language dictionary.

|  |  |  |
| --- | --- | --- |
| negative | positive | Total score |
| -24 | +62 | +38 |

There is little research using sentiment analysis for chatbot related focus group results that can help to understand the scoring found. However, on a basic interpretation the higher the score the better the chatbots were discussed in the FGD’s. A score of 72% (62/ (24+62))) would be in 3/4th quartile in distribution of sentiment distribution. Alternatively, 62/24 = 2.58 would state the ratio that for every 1 negative word recorded, there were 2.58 positive words recorded.

##### Discussion

**4**

Contents

* 1. **Quantitative Results** [**33**](#_bookmark47)
     1. CUQ [33](#_bookmark48)
     2. TAM [34](#_bookmark50)
  2. **Qualitative Results** [**34**](#_bookmark51)
  3. **Limitations** [**34**](#_bookmark52)
  4. **Conclusions** [**34**](#_bookmark53)

##Summary of Findings

The Chatbots were beneﬁcial. Learners have lots of other choices such as YouTube, but there is a certain need for personalized information gathering, this can save time and prevent learning incorrect information. This was one reason why they were rated positive as they are able to streamline data ﬁnding for learners in a format that is understandable and easy to them.

###### Quantitative Results

* + 1. CUQ

Holmes et al. [<https://dl.acm.org/doi/10.1145/3335082.3335094>]

designed the CUQ to be comparable with the system usability scale (SUS).

We have calculated both these scores out of 100 to allow the same benchmark, which is 68. A score of 68 is at the center of the range is thought of as “C”. The average benchmark for CUQ is 68, and in initial/pilot studies 68 may be considered higher than expected when considering technical issues, less developed user interfaces etc.

Previous studies evaluating chatbots have had similar score. For example, in 2022 Link found a physical activity promotion chatbot received 64.5/100, with lowest score at 40.6

* + 1. **TAM**

Diagram

Description automatically generated

**Figure 4.1:** TAM Model processes

###### Qualitative Results

###### Limitations

###### Conclusions

Those results can be interpreted that the learning objectives of the training event was chosen appropriately for the diverse audience including clinicians, academics, researchers, and learning technologists/IT specialist resulting to a successful training event that enable participants to take the acquired knowledge back to their organisations in order to co-design and implement. As it was expected and can be depicted from self-conﬁdence statements that some participants being very conﬁdent before the event, not all the objectives expected to be reached by everyone, since the training was targeting both technical and non-technical participants. However, on both average and individual matched responses participants self-statements showed that they improved their knowledge and understanding in using co-creation approaches to develop digital education resources and in designing and developing chatbots as educational resources.

##### (Additional Analyses) Training Events

**5**

**Contents**

* 1. **CEPEH Training Event C1** [**36**](#_bookmark55)
  2. **CEPEH Training Event 2** [**37**](#_bookmark56)
  3. **bibliography:**  [**39**](#_bookmark57)

###### CEPEH Training Event C1

The CEPEH training event C1 held at the premises of University of Nottingham aiming to prepare participants for the practical elements of co-creation and implementation of chatbots as an educational resource. It combined both theoretical and hands-on training. 15 participants were from RISE, AUTH, UoN.

Project managers of partners signposted the person involved, and relevant announcements were made though social media channels to the wider public. External to the project speakers were from University of Leeds, and Computer Science Department of University of Nottingham. It included academics, medical doctors, and researchers with focus both on clinical research and digital innovations in healthcare education and IT specialist/learning technologists 11.18 years of experiences (SD=7.2). A balance between male and female participants achieved. Participants were asked to highlight what they liked for each day and how each day can be improved. Findings are described below per day of the training event.

Day 1

The participants comment that they liked the design method for educational re- sources presented using a co-creation approach, they liked the interactions with other groups, and they liked the overview of existing chatbot resources of the partners. On the areas that can be improved, more media material were requested.

Day 2 Participants enjoyed the presentation from the invited speaker from another faculty of the University of Nottingham, the CEPEH resources presented and the storyboarding process. Participants highlighted that the participation of more clinicians in the event would be an added value in regards with the storyboarding process.

Day3 Participants liked the hands-on activities of the day also enjoyed the creativity of the groups on the online chatbot development tool. As an area of improvement, participants wanted more time on hands on sections.

###### CEPEH Training Event 2

**Pre-Training Event survey May 9th-13th 2022 Thessaloniki, Greece**

Twenty-six participants attended the Training Event, along with approximately 10 staﬀ members. There were 21 undergraduate students and 5 postgraduate students, who completed the survey for a total of 26 responses. There were 86% of participants who stated they had not been to a similar event like the training event CEPEH facilitated. There were 90% of students who found the event schedule very organized, and 70% agreed most of the planned sessions were relevant to that interest with the remaining 30% not having enough experience to understand the context to determine if they are interested in the training event. There were 95% of students agreeing or strongly agreeing the training event location is great, the remaining person did not leave additional comments.

Table 1 suggested attendees had minimal intention to share their own ideas due to lack of previous experience of attending such events, or due to lack of knowledge on the area. However, most were interested in listening to other groups and hearing contextual cases in healthcare.

There were 77% of participants stated they were novices in experience with chatbots in healthcare and were attending to learn more. The remaining 23% (7 students) stated they were competent and had limited experience with chat- bots in healthcare.

One day had several events regarding cybersecurity in healthcare. When asked before these events, 83% stated they were neutral or disagreed that they felt conﬁdent about their cybersecurity knowledge in healthcare. In addition, 80% stated they when neutral or disagreed that they felt they had strong cybersecurity safety in healthcare. Table 2 shows the main pre and post results suggesting a positive experience for more than 75% of attendees on all measures.

There were 90% (23) of students who heard about the event through a lecturer or a professor, the CEPEH newsletter (2), and 1 person was informed through the anatomy tutoring system at Karolinska Institute. Additionally, 60% suggested the training event to somebody else before the course started.

There were six individuals who stated neutral or disagree when asked if having issues on registration or ﬁnding the information for the event. This may have been due to being dependent on emails to receive the information, instead of a dedicated website where the information is available anytime.

As this was face-to-face, participants were asked about suﬃcient Covid-19 precautions in place at the facility, 94% agreed with suﬃcient precautions, two in- dividuals stated no but did not give further information in the additional input box provided. In summary, most participants were undergraduate students with novice experience, happy with the training event location, felt the sessions were relevant to them, and most shared the event with their colleagues. The values of co-creation,

chatbots in healthcare and taking patient history were bestowed to students in an engaging and well-received manner. Notably, the highest ratings were for staﬀ friendliness which is key to engagement and consistent interaction throughout the intense and long 5-day duration. The sessions were recorded there for the online recordings may be viewed with higher numbers over the subsequent weeks.

###### bibliography:

##### Appendix

##### References

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. John Murray.

Goethe, J. W. von. (1829). *Wilhelm Meisters Wanderjahre oder die Entsagen- den*. Cotta.