# A Comparison of Quantitative and Qualitative Data from a Formative Usability Evaluation of an Augmented Reality Learning Scenario

Dragoş Daniel IORDACHE, Costin PRIBEANU
National Institute for Research and Development in Informatics - ICI Bucharest
{iordache, pribeanu}@ici.ro

*The proliferation of augmented reality (AR) technologies creates opportunities for the development of new learning scenarios. More recently, the advances in the design and implementation of desktop AR systems make it possible the deployment of such scenarios in primary and secondary schools. Usability evaluation is a precondition for the pedagogical effectiveness of these new technologies and requires a systematic approach for finding and fixing usability problems. In this paper we present an approach to a formative usability evaluation based on heuristic evaluation and user testing. The basic idea is to compare and integrate quantitative and qualitative measures in order to increase confidence in results and enhance the descriptive power of the usability evaluation report.*
***Keywords****: augmented reality, multimodal interaction, e-learning, formative usability evaluation, user testing, heuristic evaluation.*

## 1 Introduction

The proliferation of augmented reality technologies are creating opportunities for the development of new learning scenarios able to promote new teaching methods and to enhance students' motivation to learn. The interest in AR-based educational systems that are providing with interaction techniques able to support a collaborative and active learning significantly increased in the last years.

According to Azuma [1], Augmented Reality is a variation of Virtual Reality (VR) that supplements reality, rather than completely replacing it. From the point of view of e-learning systems, AR is a new type of multimedia content featuring an integration of real and virtual (computer generated images) into real environments, real time 3D interaction and targeting all senses (visual, auditory and haptical).

Several configurations of augmented reality technologies exist based on the integration of real and virtual. AR systems based on head mounted displays (HMD) are integrating specific AR devices into a real life environment. Conversely, desktop AR configurations are bringing real life objects into a computing environment. As such, they make possible the integration of real objects used in the traditional didactics with computer generated images that are providing with explanations given in real time via a multimodal user interface. Touching, holding and manipulating a real object make it possible learning by doing which is an effective alternative to the traditional learning by reading.

The tight integration of real and virtual into a single interaction space is challenging designers to create new interaction techniques. From a usability point of view, these interaction techniques are a corner stone of AR systems. AR interaction components are often poorly designed, thus reducing the usability of the overall system [5]. An explanation is the lack of specific user-centered design methods and usability data for AR-based systems [3], [4].

The ISO standard 9241-11:1994 [8] defined usability as the extent to which a product can be used by specified users to achieve specified goals effectively, efficiently and with satisfaction in a specified context of use. Depending on its purpose, the evaluation can be formative or summative [14].

Formative usability evaluation is performed in an iterative development cycle and aims at finding and fixing usability problems as early as possible [16]. The earlier these problems are identified, the less expensive is the development effort to fix them. In order to ensure usability the system has to be tested as

early as possible in the development process.
Formative usability evaluation can be carried on by conducting an expert-based usability evaluation (sometimes termed as heuristic evaluation) and / or by conducting user testing with a small number of users. In this last case, the evaluation is said to be user-centered, as opposite to an expert-based formative evaluation.
A formative evaluation report should be both reliable and useful for designers. A general approach to increase confidence in results is to conduct both heuristic evaluation and user testing then to analyze and compare results. In this respect, Gabbard et al. [5] proposed a user-centered design approach based on 4 main activities: user task analysis, expert-based formative evaluation, user-centered formative evaluation and summative usability evaluation.
In this paper we present an approach to the formative usability evaluation of an AR-based learning scenario for biology. Both a heuristic evaluation and user testing were conducted. Then the results were analyzed in order to identify the causes and possible ways to fix them. In order to increase their usefulness for designers we grouped the usability problems on categories and documented each category with the qualitative data collected from user testing.
The rest of this paper is structured as follows. In the next section we will present the research framework. Then we describe the method used. The measures of effectiveness and efficiency are presented in section 4. In section 5 we compare the evaluation results from heuristic evaluation and use testing. The paper ends with conclusion in section 6.

## 2 The ARiSE research project

The AR-based learning scenario was developed in the framework of the ARiSE research project (Augmented Reality for School Environments). The project was carried on in a consortium of five research partners and two school partners.
The main objective of the ARiSE project is to test the pedagogical effectiveness of introducing AR in schools and creating remote collaboration between classes around AR display systems. ARiSE developed a new technology, the Augmented Reality Teaching Platform (ARTP) in three stages thus resulting three research prototypes. Each prototype is featuring a new application scenario based on a different interaction paradigm.
ARTP is a desktop AR environment: users are looking to a see-through screen where virtual images are superimposed over the perceived image of a real object placed on the table [17].
In the biology scenario, the real object is a flat torso of the human body showing the digestive system. The test was conducted on the platform of ICI Bucharest. The real object and the pointing device could be observed in Figure 1 which is showing, two students staying face-to-face and sharing the same torso.

**Fig. 1.** Students testing the biology scenario

A pointing device having a colored ball at the end of a stick and a remote controller Wii Nintendo as handler is used as interaction tool that serves for three types of interaction: pointing on a real object, selection of a virtual object and selection of a menu item. The tasks as well as user guidance during the interaction are presented via a vocal user interface.
The application implemented 4 tasks: a demo program explaining the absorption / decomposition process of food and three exercises: the 1st exercise asking to indicate the organs of the digestive system and exercises 2 and 3, asking to indicate the nutrients absorbed / decomposed in each organ respectively the or-

gans where a nutrient is absorbed / decomposed. In order to make possible a self evaluation of their knowledge, the application counted and displayed after each exercise the errors made by students.

## 3 Method and procedure

There are several approaches to usability evaluation and, consequently many usability evaluation methods [6]. In the last decade, many usability studies compared the effectiveness of various usability evaluation methods [7]. As pointed out in [9], the trend is to delineate the trade-offs and to find ways to take advantage of the complementarities between different methods. In order to increase confidence in results mixed research approaches that are undertaken as well as comparison between quantitative and qualitative data [15].

Another key concern is to increase the downstream utility of usability evaluation results, i.e. to make them useful for designers [11]. Downstream utility was defined as the degree to which the plus or minus in usability could be directly related to the evaluation results and recommendations [10]. This can be achieved by finding suitable ways to report usability problems and prioritizing them following their importance for the software system.

Two evaluation methods are widely used in the current usability practice: *user testing* and *heuristic evaluation*. User testing is based on testing the system with a group of participants representing, as closely as possible, the users of the target product. Heuristic evaluation is conducted by a small group of evaluators who examine a user interface and assess its compliance with a set of usability principles or heuristics [13].

Usability problems (UP) identified during heuristic evaluation are ranked for their potential impact into severe, moderate and minor problems. Heuristic evaluation provides two kinds of measures:

- Quantitative: number of usability problems in each category.
- Qualitative: detailed description of individual usability problems.

Nielsen & Molich [13] proposed 10 heuristics for the evaluation of a user interface: visibility of system status, compatibility with the activity, user freedom and control, consistency, error prevention, recognition instead of recall, flexibility, aesthetics and minimalist design, quality of error messages.

Bastien and Scapin [2] proposed a set ergonomic criteria consisting of 18 elementary criteria: prompting, immediate feedback, grouping / distinction by location, grouping / distinction by format, legibility, concision, minimal actions, information density, explicit user actions, user control, user experience, flexibility, error protection, quality of error messages, error correction, significance of codes, consistency, compatibility. These criteria are grouped into 8 categories (general principles). For each ergonomic criterion the prescription is providing with definition, rationale, comments, and examples of guidelines.

A user testing of the biology scenario was conducted with 42 students (2 classes) from two schools in Bucharest. Students came in groups of 6-8 accompanied by a teacher, so testing has been organized in 2 sessions. The students were assigned 3 tasks: a demo lesson, the 1st exercise and one of the exercises 2 or 3.

Effectiveness and efficiency measures from user log files were collected. After testing the students were asked to mention three most positive and three most negative aspects regarding their interaction with the ARTP.

A heuristic evaluation was carried on in the same period. Two experts in usability evaluation tested the application by performing all tasks in order. The usability was assessed against the ergonomic criteria defined by Bastien and Scapin [2] and further refined and adapted for mixed reality environments by Bach & Scapin [3]. Usability problems identified were recorded by using the template described in [9].

## 4 User testing results

The most negative aspects mentioned by students after user testing were analyzed in order to extract key words (attributes). Some

students only described one or two aspects while others mentioned several aspects in one sentence thus resulting a master list of 79 attributes. In fact, these attributes represent inductive codes [*] generated by a direct examination of data during the coding process. The attributes are face sheet codes applied to the whole list of students' comments. Attributes were then grouped into hierarchical categories as illustrated in Table 1.

**Table 1.** Categories of negative aspects

| Category | Frequency | Percent |
|---|---|---|
| Selection problems | 25 | 31.0 |
| Eye pains and glasses | 18 | 22.8 |
| Real object too big | 15 | 19.0 |
| Visualization problems | 5 | 6.3 |
| Wrong superposition | 2 | 2.5 |
| Other | 14 | 17.7 |
| Total | 79 | 100.0 |

Essentially, the negative aspects mentioned by students are qualitative descriptions of usability problems they experienced during the interaction with the ARTP.

Most of the negative aspects are related to selection problems (31%). Students also complained about eye pains after using the 3D stereo glasses as well as about the difficulty to accommodate glasses. A number of 15 attributes of 79 (19%) are related to the difficulty to manipulate the real object which is shared by two students. Next two categories account for 7 attributes (8.8%) that are related to the accuracy of the visual perception. This issue is often reported in AR systems where computed generated images are superimposed on the see-through screen. Finally, other attributes (17.7%) are related to various technical problems, such as: difficulty to accommodate the headphones, clarity of sound, difficulty to understand and operate the application.

In Table 2, the measures of effectiveness (completion rate and number of errors) and efficiency (mean execution time) for the biology scenario are presented that are based on the data collected in log files during user testing.

The number of observations is varying because not all tasks have been assigned to each student and, in some cases it was not possible to perform all assigned exercises because of technical problems.

**Table 2.** Effectiveness and efficiency measures

| Task | Success | Failure | Rate | Errors | Time |
|---|---|---|---|---|---|
| 1 | 33 | 8 | 80% | 6.88 | 455.8 |
| 2 | 32 | 3 | 91% | 6.28 | 318.4 |
| 3 | 16 | 1 | 94% | 15.90 | 401.8 |

The first exercise was easier to solve but more difficult to use. The lower rate of success is due both to the lack of knowledge in biology and difficulty to use. Errors (min=0, max=19, SD=4.83) were mainly due to the difficulties experienced with the selection of an organ. The mean time on task was 455.8 sec (SD=193.7).

As shown in Table 1, most mentioned negative aspects are related to selection problems. Students often complained about the difficulty to select small organs, such as: oral cavity, duodenum or pancreas. Table 3 shows the number of errors in exercise 1.

**Table 3.** Errors in selecting organs

| Organ name | No. Errors | No. Students | Mean no. of errors |
|---|---|---|---|
| Oral cavity | 42 | 19 | 2.21 |
| Duodenum | 40 | 18 | 2.22 |
| Pancreas | 36 | 15 | 2.40 |
| Esophagus | 33 | 16 | 2.10 |
| Other organs | 76 | 28 | 2.71 |
| Total | 227 | | |

As it could be observed, 151 errors of 227 (66.5%) were encountered while students tried to select small organs (oral cavity, duodenum, pancreas and esophagus), i.e. a mean of 37.7 errors / organ for all students. The rest of organs (lambs, stomach, liver, small intestine, large intestine) accounted for 76 errors with a mean of 15.2 errors / organ for all students.

The last two exercises were more difficult to solve (there is a many-to-many relationship between organs and nutrients). From a pedagogical point of view, the higher rate of success as compared to the first exercise is due to the gain in knowledge during the first task, when students rehearsed the position of each organ. From a usability point of view, the

second exercise was easier to use since the nutrients were selected with the remote controller.
Three students from 35 failed to solve the second exercise. All students made errors and 4 students made over 10 errors. The mean execution time was 318.4 sec. (SD=220.1)
Only one student from 17 failed to solve the third exercise. All students made errors and 7 students made over 20 errors. In this case, errors are due to the lack of knowledge and to the difficulties in selecting organs. The mean execution time was 401.8 sec (SD=226.8).
As regarding the mean value of errors there is a small difference between the first two exercises (6.28 respectively 6.88) and the last one (15.90). It seems that students found more difficult to use the pointer for indicating in which organs the nutrients are absorbed / decomposed than to use the remote controller for selecting nutrients.
Overall, 32 students (78%) succeeded to perform all assigned exercises, 6 students performed only one exercise while 3 students failed to perform any exercise.
The total execution time for the 11 students performing all assigned exercises varied between 705 sec. (with 22 errors) and 1972 sec. (with 10 errors). The total mean time on task was 1207.8 sec. i.e. 20.1 min (SD=8.75).

## 5 Heuristic evaluation results

The heuristic evaluation was done by assessing the usability of the ARTP against ergonomic principles. Then usability problems were grouped according to their severity (estimated impact) in three categories: severe, moderate and minor, as shown in Table 4.
Two usability experts tested the scenario by performing each task in order. The evaluation followed the process of consolidating the list o usability problems according to the procedure described by Law & Hvannberg [12].
Individual usability problems were recorded and documented. Many of them were found in several tasks. Then each expert filtered the usability problem set by retaining a set of unique usability problems. The localization of each of them was updated in order to note all tasks which are affected. Finally, the lists of filtered usability problems were merged in order to produce a set of unique usability problems.

**Table 4.** Usability problems per task and severity

| Tasks | Severe | Moderate | Minor |
|---|---|---|---|
| All tasks | 1 | 8 | 1 |
| Demo program | 2 | 1 | |
| Exercise 2 | 1 | 2 | 2 |
| Exercises 1 and 3 | | 1 | |
| Total | 4 | 12 | 3 |

After consolidation, a total number of 19 usability problems were retained from which 10 apply for all tasks, 5 for the second exercise, 3 for the demo program and 1 for the exercises 1 and 3.
Most of the usability problems (12 problems of 19) are moderate, 4 are severe and 3 are minor. 10 of 19 usability problems are general since they affect all the tasks. Most of the specific usability problems (5 problems of 9) are affecting the second exercise. Each usability problem was recorded following a template as illustrated in Table 5.

**Table 5.** Example of usability problem description

| | |
|---|---|
| Id | UP5 |
| Task | All |
| Context of use | Users are adjusting the screen position for many reasons: to better fit the virtual image, to fit with their height or to allow ball manipulation |
| Usability problem | The superposition is wrong although it should be accurate regardless the position of the “see-through” screen |
| Criterion | User guidance - prompting |
| Severity | Moderate |
| Suggestions for designers | Revise the visualization program. Elaborate on a set-up specification in order to harmonize the height dimensions for the table / chair and “see-through” screen. |

The context of use enables a rapid identification / localization of the problem as well as the replication of the evaluation. An analysis of the problem description occurring in a given context makes it possible to identify the causes and to structure the usability problems accordingly. This is very useful for designers which can establish priorities in order to fix most of the usability problems in a

short period of time and with less development effort.
Based on this analysis, the usability problems were grouped into categories, as illustrated in Table 6.

**Table 6.** Categories of usability problems

| Category / ergonomic criteria | Prompting | Legibility | Feedback | Compatibility | Minimal actions | Info density | Total |
|---|---|---|---|---|---|---|---|
| Selection | 2 | | 2 | 2 | 1 | | 7 |
| Visualization | 3 | | | | | 1 | 4 |
| Superposition | 2 | 3 | | | | | 5 |
| Glasses | | | | 1 | | | 1 |
| Other | 2 | | | | | | 2 |
| Total | 9 | 3 | 2 | 3 | 1 | 1 | 19 |

From the point o view of the violated ergonomic principle, most of the usability problems are related to prompting, feedback and legibility (general principle: user guidance). Other two problems are related to the information density and minimal actions (general principle: work load).
A percentage of 36.8% (7 out of 19) of usability problems are related to the difficulties experienced by users during the selection of an organ or a nutrient. Selection is done with the pointing device and the problems are related to prompting (organ name not displayed or cursor blocked when leaving the selection area), lack of feedback (pointer ball not recognized), compatibility (oral cavity not shown because the real object is too big and lies outside the selection area) and minimal actions (selecting the last item in the menu could be done easier).
These problems were also described in a less formal way, in the negative aspects mentioned by students ("*Sometimes the ball is blocked resulting in a wrong selection of an organ*", "*I couldn't select well during the first exercise*", "*Is difficult to move the cursor*").
Most difficult was to select small organs ("*It is difficult to find some organs*", "*I could not select de oral cavity*"). This was also due to the fact that a torso is shared by 2 students. While for one of them the oral cavity is at hand, for the other is difficult to select it because the torso has to be moved until the organ is in the selection area of the camera. It was also difficult to select the small organs placed much closed to each other like duodenum and pancreas. It was difficult for the students to maintain the position of the cursor on the organ. In fact, several selection problems were provoked by the size of the real object. Many students complained that the real object is too big ("*We had to move the torso*", "*The torso is too big and not everything could be observed on the screen*").
Visualization and superposition lack of accuracy represent the next categories of usability problems identified by the heuristic evaluation. These problems are mainly associated with prompting and legibility. Students also complained about this ("*The screen is not synchronized with the torso*", "*The projected image was not well superimposed over the torso*", "*Some times the projected image was not clear*").
The 3D wireless stereo glasses were another source of discomfort. From the one hand, for some students was difficult to accommodate the glasses and headphones, especially if they were already wearing their own glasses. Most of the students also complained about the eye pains after user testing ("*Glasses are too small*", "*Uncomfortable glasses*", "*I felt a pain in my eyes during and after testing*"). This was mainly due to the interference between the infrared emitters which provoked shuttering of glasses.

## 6 Conclusion

Several usability problems exist that were identified and documented by heuristic evaluation and user testing results. User testing provides with quantitative measures of effectiveness and efficiency, such as rate of completion and time on task. While these quantitative measures are good to generally assess the usability of the learning scenario, they are not descriptive enough to reveal individual usability problems.
The novelty of our approach consists in the integration of results of user testing and heu-

ristic evaluation. In this respect, we analyzed the usability problems and grouped them into categories which are similar to the structuring of negative aspects mentioned by students. This way is possible to enrich the description of each type of usability problem.

Heuristic evaluation performed by experts provides with a complete description of individual usability problems, as well as with suggestions for fixing it. The severity enables a prioritization for designers while the ergonomic criterion violated by the design helps in understanding the problem and its effect on user's interaction with the system.

Comparing the quantitative and qualitative measures collected via user testing and heuristic evaluation is increasing confidence in the evaluation results. It was clear that most of the usability problems were related to the selection of organs / nutrients, visualization and superposition. From an ergonomic point of view, these problems have to be fixed in order to ensure appropriate user guidance.

The description of usability problems is more reliable and useful when is complemented with qualitative measures such as negative aspects mentioned by students after testing. The excerpts from students' comments are helping designers to better understand how users perceive the ease of use of specific AR-based interaction techniques.

Many students complained about eye pains provoked by the wireless stereo glasses. Therefore it was strongly recommended to replace them with wired stereo glasses and to include this requirement into the technical specification of the desktop AR platform. Also, calibration of technical devices should be improved and automated as much as possible.

Another requirement which should be included in a technical specification of an AR platform is related to the workplace. Many students found it difficult to manipulate the real object and complained about the lack of compatibility between the real object and the platform. Therefore it was recommended that the height of the see-through screen be adjusted with respect to the height of the table, in order to allow the hands to freely manipulate real objects.

**Acknowledgement**

This work was supported by the ARiSE research project, funded by the EC under FP6-027039.

**References**

[1] R. A. Azuma, "Survey of Augmented Reality," *PRESENCE: Teleoperators and Virtual Environments*, Vol. 6, No. 4, pp. 355-385, 1997.

[2] A. Bastien and D. L. Scapin, "Ergonomic criteria for the evaluation of human-computer interfaces," *Technical report No. 156*, INRIA, Roquencourt, France, 1993.

[3] C. Bach and D. L. Scapin, "Adaptation of Ergonomic Criteria to Human-Virtual Environments Interactions," In *Proceedings of Interact'03. IOS Press*, pp. 880-883, 2003.

[4] C. Bach and D. Scapin, "Obstacles and perspectives for Evaluating mixed Reality Systems Usability," In *Mixer workshop, Proceedings of IUI-CADUI Conference 2004*, pp. 72-79. ACM Press, 2004.

[5] J. Gabbard, D. Hix, E. Swan, M. Livingston, T. Herer, S. Julier, Y. Baillot and D. Brown, "A Cost-Effective Usability Evaluation Progression for Novel Interactive Systems," In *Proceedings of Hawaii International Conference on Systems Sciences, Track 9*, pp. 90276c, IEEE, 2004.

[6] K. Hornbaek, "Current practice in measuring usability: Challenges to usability studies and research," *Int. J. Human Computer Studies*, No. 64, pp. 79-102, 2006.

[7] E. T. Hvannberg, E. L. C. Law and M. C. Larusdotir, "Heuristic Evaluation: Comparing ways of finding an reporting usability problems," *Interacting with Computers*, No. 19, pp. 255-240, 2007.

[8] ISO/DIS 9241-11:1994 Information Technology – Ergonomic requirements for office work with visual display terminal (VDTs) - Guidance on usability.

[9] B. Johnson and L. Christensen, *Educational Research. Quantitative, Qualitative and Mixed Approaches. Third edition*, Sage Publications, 2008.
[10] E. L. C. Law and E. T. Hvannberg, "Complementarities and convergence of heuristic evaluation and usability test: A case study of UNIVERSAL brokerage platform". *Proc. Of NordiCHI Conference 2002*, ACM, pp. 71-79, 2002.
[11] E. L. C. Law, "Evaluating the downstream utility of user tests and examining the developer effect: A case study," *International Journal of Human-Computer Interaction*, Vol. 21, No. 2, pp. 147-172, 2006.
[12] E. L. C. Law, M. C. Lárusdóttir and M. Norgaard (Eds), *Downstream Utility 2007: The Good, the Bad, and the Utterly Useless Usability Evaluation Feedback*, Toulouse, France, IRIT Press – Toulouse, 2007.
[13] E. Law and E. T. Hvannberg, "Consolidating usability problems with novice evaluators," *Proceedings of NordiCHI 2008*, ACM Press, pp. 495-498, 2008.
[14] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," *Proc. ACM CHI'90*, Seattle, WA, 1-5 April 1990, pp. 249-256.
[15] M. Scriven, *Evaluation thesaurus, 4th ed*, Newbury Park, CA: Sage Publications, 1991.
[16] D. Sjoberg, T. Dyba and M. Jorgensen, "The future of empirical methods in software engineering research," *In Proceedings of FOSE 2007*, IEEE Computer Society, pp. 358-378, 2007.
[17] M. Theofanos and W. Quesenbery, "Towards the Design of Effective Formative Test Reports," In *Journal of Usability Studies*, Issue 1, Vol. 1. pp. 27-45, 2005.
[18] J. Wind, K. Riege and M. Bogen, „Spinnstube: A Seated Augmented Reality Display System," In *Virtual Environments, Proceedings of IPT-EGVE – EG/ACM Symposium*, pp. 17-23, Eurographics, 2007.

**Dragoş Daniel IORDACHE** received a Master degree in educational counselling from the University of Bucharest in 2006 and is currently a PhD student in educational sciences. Currently he is a psychologist at ICI Bucureşti. Dragoş Daniel Iordache is a member of the Romanian HCI group (RoCHI – SIGCHI Romania) and he served as conference reviewer for the last two editions of the national conference. His research interests include: usability and pedagogical evaluation of educational systems, usability guidelines, user testing and heuristic evaluation. He is author/co-author of 3 journal papers and 10 conference papers.

**Costin PRIBEANU** received the PhD degree in Economic Informatics from the Academy of Economic Studies in 1997. Currently he is a senior researcher at ICI Bucureşti. Costin Pribeanu is a Senior ACM member since 2009 and the Vice-Chair for conferences of the Romanian HCI group (RoCHI – SIGCHI Romania) since 2009. His research interests include: task analysis, user interface design, task-based design of user interfaces, design patterns, usability evaluation, and usability guidelines. He is author / co-author of 4 books, 5 edited books 6 book chapters, over 30 journal papers and over 50 conference papers. He is currently member of the Springer HCI Book series editorial board and co-editor in chief for the Romanian Journal of Human Computer Interaction. He also served as program committee member / reviewer for several major HCI conferences.