Preliminary Results of a Usability Study in the Domain of Technology-Based Assessment Using a Tangible Tabletop

Eric Ras
Public Research Centre Henri Tudor
av. J. F. Kennedy 29, L-1855 Luxembourg-Kirchberg, Luxembourg
eric.ras@tudor.lu

Valérie Maquil
Public Research Centre Henri Tudor
av. J. F. Kennedy 29, L-1855 Luxembourg-Kirchberg, Luxembourg
valerie.maquil@tudor.lu

ABSTRACT
The goal of our research was to investigate whether the usability respectively learnability of a tabletop device is judged significantly higher when a group uses the device compared to individuals. This research work is motivated by the fact, that useful scenarios need to be identified for tangible user interfaces in education and technology-based assessment in particular. The usage scenario was defined in the field of computer-based assessment: The subjects had to assign tangibles to images of planets on a tangible tabletop. Besides counting the time needed to solve the test item and the number of attempts, the activities on the table have been video recorded as well. The usability and learnability were measured by using the System Usability Scale (SUS). The results showed that no differences were found regarding the performance of solving the test item. Furthermore, the individuals even rated the usability significantly higher than the subjects in the collaborative setting. Reasons for this effect need to be investigated by comparing the outcomes of the video analysis with the related answers of SUS.

Keywords
technology-based assessment, tangible tabletop, usability study, SUS

1. INTRODUCTION
Assessments cover a wide range of methods that have been changing dramatically with the increasing use of new technologies. In the past, assessment practices were seen as a means for selecting students at university [1] and for monitoring educational systems (e.g., PISA and PIAAC), or they were used for diagnostic purposes.

Technology-based assessment (TBA) can facilitate learning and instruction in ways that paper and pencil cannot [2]. Technology-based assessment may avoid existing time constraints, ease the comparison of results, reduce measurement errors, etc. [3, 4]. Nevertheless, using a new technology in assessment requires first a deep understanding with regard to the actions the students perform, the group dynamics in collaborative settings, as well as the required digital literacy to use such a new technology. These aspects may impact the way how the results of a test are obtained (i.e., measured) or they may produce undesired effects (i.e., measurement errors).

During the last years, topics such as measuring solving strategies (i.e., measurement of dynamics in a test) and assessing collaborative problem solving are getting more attention. They require new technologies that support and allow collaborative activities, which can be tracked. A potential solution are tabletop-based tangible interfaces.

Tangible user interfaces (TUIs) are an approach to create new types of interaction combining physical and digital elements as part of a physical space. Exploratory, design-focused studies have suggested that TUIs provide some learning benefits, due to the additional haptic dimension, the better accessibility, and the shared space that can be used in collaborative situations [5].

The aim of this paper is to investigate whether the usability of a tabletop device is judged significantly higher when a group uses a tangible tabletop device compared to individuals using it for the same task. A drag and drop test item was solved by the adult subjects in August 2011. In addition, we were interested in analyzing solving strategies, the types of activities on the table, and the handling of the tangible objects.

The next section provides an overview of usability measurement instruments as well as the use of tangible user interfaces in technology-based assessment. Section 3 depicts the technical setup and describes the study design and related analysis procedures. Section 4 discusses the results of the usability study as well as the outcomes of hypothesis tests. The last section concludes the paper.

2. RELATED WORK
A test consists of several test items. Such an item contains a stimulus, a question, responses, a score schema, and feedback elements [6].

IT does, for example, allow rich new assessment tasks and provides powerful scoring and reporting techniques [7]. It supports the user in assessment resource development, data collection, and presentation of the results. New constructs such as cognitive and behavioral skills can be assessed; dynamics aspects (e.g., time spent to answer a question, number of clicks, choice of learning materials) can be observed and collected effectively in the near future [3].

Companion Proceedings of *IHM'11, 23ème Conférence francophone sur les Interactions Homme-Machine*, October 24-27, 2011, Sophia Antipolis, France.

In literature, we can find a number of TUI example systems, demonstrating learning situations in different kinds of application domains. A common approach is to support planning, problem solving, and simulation through TUIs. For example, the TinkerTable [8] allows construction and simulation of a warehouse to explore and solve logistical problems. This system provided apprentices with the opportunity to apply their acquired knowledge in real use scenarios. Other learning systems implement the concept of digital manipulatives [9], computationally enhanced building blocks, which allow exploration of abstract concepts. This principle is followed by SystemBlocks and FlowBlocks, two physical, modular interactive systems, which children can use to model and simulate dynamic behaviour [10].

Different measurement instruments exist in order to measure usability. The model of *Unified Theory of Acceptance and Use of Technology* (UTAUT) [11], which is a measurement instrument based on the former measurement instrument of the *Technology Acceptance Model* (TAM) [12, 13], is often used. UTAUT contains constructs for measuring performance expectancy, effort expectancy, attitude towards using technology, intention, anxiety, self-efficacy, facilitating conditions, and social influence. Another usability measurement instrument is called *System Usability Scale* (SUS) [14], which consists of a reliable, low cost usability scale. SUS allows comparing the usability between systems. The motivation to use SUS for this research work was to get feedback on the usability of the system fast and to have a reference point in order to compare the usability with future releases and variants of the tangible tabletop. In the initial paper of Brooks, he stated that "SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own." Nevertheless, during the last years several studies investigated the factor structure of SUS. Lewis and Sauro [15], for example, revised the results of an analysis of Bangor et al. who stated that there is only one factor [16]. Lewis and Sauro collected 324 completed SUS questionnaires and conducted a factor analysis that converted to a two factor solution: 8 items represent the factor usability and the two remaining items refer to the factor learnability with Cronbachs Alpha coefficients of 0.91 respectively 0.70.

3. CASE STUDY DESIGN AND SETUP

For the experiment, we set up a tangible tabletop system, based on the optical tracking framework "reactivision". The worktop was sized 95x120cm, with an interactive area of 75x100cm. On the table, we projected an image of the solar system, showing the sun and each of the nine planets. We further created 9 cards, each showing the name of one of the planets. A camera and projector had been placed underneath the table to track the positions of the physical objects and project feedback onto the semi- translucent tabletop surface.

The implemented task was to assign the correct name of a planet to the correct image of a planet (according to QTI standard: associate item). The cognitive nature of the task was on the level of remembering, i.e., locating long-term memory that is consistent with the presented images on the table. Since the focus of the experiment was on usability and not learning, we consider the target group not as essentially important. The subjects were male and female colleagues at our institute with an average age of 30 years. Knowledge about planets was considered to be general knowledge. Hence, no prerequisites were required to take part in this test.

When a user places a card onto a planet, the user gets an immediate feedback in form of a red (false answer) or green (correct answer) circle that is shown around the planet. During the solving of the task, the system counts the number of attempts (i.e., wrong pairing of a card and a planet's image was counted as one attempt), and the time needed to solve the task. As soon as all the planets are correctly assigned, the system shows a scoring window that displays the results. A between-subject controlled experiment was conducted with a control group of 11 individuals solving the test items alone as well as 24 subjects divided into 8 groups of equal size (i.e., experimental group). The subjects were randomly selected from the research department and assigned to both groups.

Prior to each test, the users were explained the concept of a TUI and how it detects physical objects. We described their task and which kind of feedback they can expect from the system. They could place themselves around the table as they preferred.

The experimentation was video recorded and at the end we distributed a questionnaire. The questionnaire consists of three questions asking the background knowledge on astronomy and the System Usability Scale (SUS) with ten questions. Further, we took notes on the performance of the group and how the users placed themselves around the table.

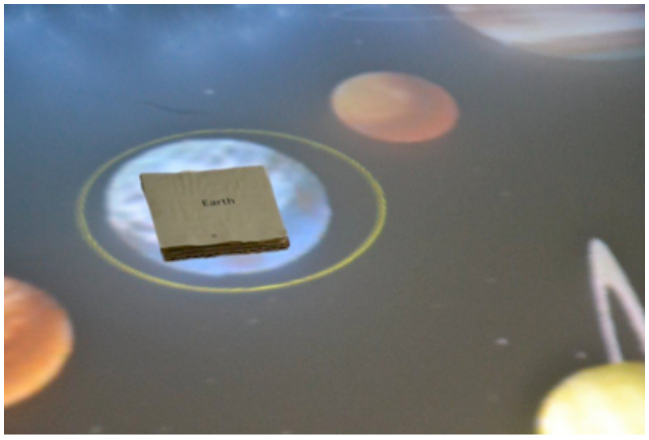

Figure 1: Tangible tabletop device

Figure 2:Mapping the card (tangibles) with the images of the planets

4. DATA ANALYSIS AND RESUTLS

4.1 SUS and Hypotheses

SUS was used to measure usability(U) and learnability(L):

- I think that I would like to use this system frequently (U)
- I found the system unnecessarily complex (U)
- I thought the system was easy to use (U)
- I think that I would need the support of a technical person to be able to use this system (L)
- I found the various functions in this system were well integrated (U)
- I thought there was too much inconsistency in this system (U)
- I would imagine that most people would learn to use this system very quickly (U)
- I found the system very cumbersome to use (U)
- I felt very confident using the system (U)
- I needed to learn a lot of things before I could get going with this system (L)

Each item has a 5-point Likert scale (i.e., "1" corresponds to "strongly disagree",..., "5" corresponds to "strongly agree"; range of a item =[1,5]). According to Brooke each item contributes to the SUS scale with a range from 0 to 4 [14]. For positively-worded items (1, 3, 5, 7 and 9), the score contribution is the scale position minus 1. For negatively-worded items (2, 4, 6, 8 and 10), it is 5 minus the scale position. To get the overall SUS score, the sum of the item score contributions were multiplied by 2.5. Thus, SUS scores range from 0 to 100 in 2.5-point increments.

Regarding usability and learnability, we followed the suggestions of Lewis and Sauro [15], and calculate usability (sum of the items 1, 2, 3, 5, 6, 7, 8, and 9) and learnability (sum of the items 4 and 10) in addition to the total score of SUS. To make the usability and learnability scores comparable with the overall SUS value (ranging from 0 to 100), the summed score was multiplied by 3.125 and 12.5, respectively.

Besides providing the results of the descriptive statistics, the following hypotheses were tested:

$H_{1.1.}$ "Usability" – The average usability *u* of the experimental group (collaborative problem solving) is higher than the average usability *u* of the control group (individual problem solving).

$\mu(u_{coll}) > \mu(u_{ind})$ *where* $\mu(.) = mean$

$H_{0.2}$ $\mu(u_{coll}) \leq \mu(u_{ind})$

$H_{1.2.}$ "Learnability" – The average learnability *l* of the experimental group (collaborative problem solving) is higher than the average learnability *l* of the control group (individual problem solving).

$\mu(l_{coll}) > \mu(l_{ind})$ *where* $\mu(.) = mean$

$H_{0.2}$ $\mu(l_{coll}) \leq \mu(l_{ind})$

A one-sided independent samples t-test was applied. A significance level of $\alpha = 0.05$ (error type I) was chosen.

4.2 Results

After data collection no missing values were encountered. One data set of the individual group was detected as complete outlier and hence omitted from analysis – the subject had difficulties to understand the questionnaire. The following two tables depict the results for the individuals solving the task alone as well as the individuals who solved the task in a group of three. A Shapiro-Wilk test revealed that the data distribution were not significantly different from a normal distribution, with two exceptions: attempts (individual with P=0.86) and the SUS total (group with P=0.07). Therefore, non-parametric tests were conducted for the hypothesis tests.

Both tables show only minor differences regarding the usability and learnability. Nevertheless, the control group (M = 90.0, SD = 5.40) rated the SUS total score about 10% higher than the control group (M = 81.7, SD = 7.47). Usability as defined by Lewis and Sauro was also rated 10% higher by the control group (M = 88.4, SD = 6.43) compared to the experimental group (M = 78.9.7, SD = 8.05). Learnability was only rated 5% higher.

The previous two tables show the performance of the two groups in terms of time needed and the attempts (wrong pairing of cards and planet's images). The interesting outcome is that the results only differ on a very small scale.

The following table summarizes the results of the hypothesis tests. The independent samples t-test revealed a significant difference in the usability between the experimental group and the control group, t (32) = -3.32, p = .002, however, not in the direction we expected: The control group (individual setting) has judged usability higher than the experimental group (collaborative setting). The same counts for the SUS total score. No significant difference has been found for learnability.

5. DISCUSSION AND CONCLUSION

An interesting outcome was that the performance on solving the test items was almost the same; only the group setting needed a slightly lower number of attempts. This was probably due to the fact that almost every group was discussing before they dropped the cards on the table. The unexpected result of having a lower usability judgment of the group compared to the individuals needs to be investigated in more detail. The statistical dispersion of the group setting regarding usability was higher, so it makes sense to investigate the more extreme usability ratings and the reasons for this by analyzing the video data. First comments were given by the subjects immediately after the experiment, which were related to the table itself such as different illumination of the table depending on the position around the table; the height of the table was problematic for smaller persons; some other complained about single incorrect detections of the markers (remark: the number of attempts has been corrected accordingly).

Future work will first concentrate on a deep analysis of the video data: a) How do the characteristics of the physical objects (e.g., table, tangibles) and the spaces (e.g., the table surface itself, above the table surface, between the individuals) support problem solving for both individual and collaborative settings; b) What is the effect of a tangible tabletop on the solving strategies and what are common solving strategies on a tangible table top? Second,

Table 1. Results of SUS usability test (*individual* setting – control group)

	N	Min.	Max	Mean	Std. Dev.	Skewness	Kurtosis
SUS total score	10	83.0	100	90.0	5.40	.579	1.33
Usability score (Item 1,2,3,5,6,7,8,9)	10	81.2	100	88.4	6.43	.549	1.33
Learnability score (Item 4, 10)	10	87.5	100	96.2	6.04	-1.03	1.33

Table 2. Results of SUS usability test (*group* setting – experimental group)

	N	Min.	Max	Mean	Std. Dev.	Skewness	Kurtosis
SUS total score	24	70.0	93	81.7	7.47	-.170	-1.32
Usability score (Item 1,2,3,5,6,7,8,9)	24	65.6	90.6	78.9	8.05	-.002	-1.40
Learnability score (Item 4, 10)	24	75.0	100	92.7	8.96	-.839	-.485

Table 3. Results regarding test performance (*individual* setting – control group)

	N	Min.	Max	Mean	Std. Dev.	Skewness	Kurtosis
Time needed (sec)	10	41	152	79.8	29.6	1.57	4.18
Attempts	10	0	15	6.50	4.79	.334	-.486

Table 4. Results regarding test performance (*group* setting – experimental group)

	N	Min.	Max	Mean	Std. Dev.	Skewness	Kurtosis
Time needed (sec)	24	21	155	79.7	42.1	.506	-.737
Attempts	24	0	13	5.88	4.96	.053	-1.62

Table 5. One-tailed independent sample t-test

	t	df	p-value	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Upper	Lower
SUS total score *	-	-	.008	-	-	-	-
Usability **	-3.32	32	.002	-9.53	2.87	-15.3	-3.68
Learnability **	-1.14	32	.262	-3.54	3.10	-9.86	2.78

* non-parametric test; ** parametric test

another interesting issue is to investigate whether the type and timing of feedback impacts collaborative problem solving. How is feedback given by humans perceived and used compared to feedback given by the system? Do time and type of feedback have an impact on the solving performance?

SUS does not cover usability from a group perspective and the question can be posed whether the validity of such instruments is given to measure usability in a collaborative setting. Future work should address this issue by extending or adapting existing measurement instruments.

6. ACKNOWLEDGMENTS

This evaluation could not have been done without the help of Ourda Atlaoui who was strongly involved in the development of the tangible UI.

7. REFERENCES

[1] Gipps, C.: Socio-cultural aspects of assessment. Review of Educational Research 24 (1999) 355-392

[2] Bennett, R.E.: Inexorable and inevitable: The continuing story of technology and assessment. Technology, Learning, and Assessment 1 (2002)

[3] Csapó, B., Ainley, J., Bennett, R., Latour, T., Law, N.: Technological issues for computer- based assessment of the 21st century skills. Draft White Paper 3. The University of Melbourne, CISCO, INTEL, MICROSOFT, Melbourne (2010)

[4] Grundwald Associates LLC Report: An open source platform for internet-based assessment. Grunwald Associates LLC, Bethesda, MD (2010)

[5] Marshall, P.: Do tangible interfaces enhance learning? : 1st international conference on Tangible and Embedded Interaction (TEI 2007). ACM, Baton Rouge, LA, USA (2007) 163

[6] Sclater, N.: Conceptualising item banks (2. Chapter). In: Sclater, N. (ed.): Item Banks Infrastructure Study (IBIS). HEFCE (2004)

[7] Scalise, K., Gifford, B.: Computer-based assessment in E-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. Journal of Technology, Learning, and Assessment 4 (2006), 3-44

[8] Zufferey, G., Jermann, P., Lucchi, A., Dillenbourg, P.: TinkerSheets: using paper forms to control and visualize tangible simulations. Third International Conference on Embedded and Tangible Interaction (TEI 2009), Cambridge, UK (2009), 377-384

[9] Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., Silvermann, B.: Digital manipulatives: new toys to think with. Computer Human Interaction (CHI 1998). ACM, Los Angeles, CA USA (1998), 281-287

[10] Zuckerman, O., Resnick, M., Saeed, A.: Extending tangible interfaces for education: digital montessori-inspired manipulatives. Computer-Human Interaction 2005 (CHI 2005), Portland, Oregon, USA (2005)

[11] Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 27 (2003), 425-478

[12] Davis, F.D.: A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results (PhD Thesis). Sloan School of Management, Vol. PhD. Massachusetts Institute of Technology (1986)

[13] Davis, F.D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13 (1989), 319-339

[14] Brooke, J.: SUS - A Quick and Dirty Usability Scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.): Usability Evaluation in Industry. Taylor & Francis, London (1996)

[15] Lewis, J.R., Sauro, J.: The Factor Structure of the System Usability Scale. In: Kurosu, M. (ed.): Human Centered Design, Proceedings, Vol. 5619 (2009), 94-103

[16] Bangor, A., Kortum, P.T., Miller, J.T.: An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction 24 (2008), 574-594