Empirical studies on a tangible user interface for technology-based assessment: Insights and emerging challenges

Eric Ras, Valérie Maquil, Muriel Foulonneau and Thibaud Latour, Public Research Centre Henri Tudor

Abstract

The assessment of higher order thinking skills should measure knowledge and procedure as well as attitudes and dispositions. It can be considered as multidimensional regarding the cognitive processes involved in solving, for instance, a complex problem. This paper reports on two empirical studies in which a so called tangible user interface (TUI) was used for the assessment. A simple matching item, suitable for measuring recall of factual knowledge, as well as a simulation item with the potential to assess higher order thinking skills are presented. The existing knowledge about using such systems for assessment is limited. We therefore focused on evaluation of the usability, specifically the user experience (UX) and interviewed experts in order to derive a list of tensions and also advantages. A first study showed that the available constructs for evaluating usability and UX need to be adapted to reflect the collaborative problem solving setting. The second study revealed that the majority of participants rated the system positively from the usability and UX perspective. Some tensions mentioned by the experts were related to the first phase of getting used to the system: When should the system start to assess the observed solving strategies? How could the system identify single atomic contributions (i.e. single experiments with the simulation parameters)? How could we include activities outside the interactive surface of the table? The table supports the user to 'recognise the perspective of others', an important sub-skill of collaborative problem solving – how can the table track these activities? Based on the outcomes of the studies we identified eight topics. We discuss the related tensions for each and the advantages between the new technology and technology-based assessment which will impact the future development of using TUI for assessment as well as the design of assessment models and methods.

Keywords

Technology-based assessment; tangible user interface (TUI); higher order thinking skills; collaborative problem solving; usability; user experience.

Introduction

The need and pressure to develop innovative solutions sheds more light on specific skills and competencies. These so-called 21st century skills refer to skills such as complex problem solving, creativity[1], critical thinking, learning to learn and decision making etc. (Binkley et al. 2012). Furthermore, learning and working in a collaborative, connected context requires skills in communication, negotiation and for learning in the digital world (Griffin et al. 2012). Educational frameworks on a global or national scale support teachers to cope with these new challenges, however the knowledge about how to assess these skills still suffers from a low level of

[1] Center for Creative Learning. http://www.creativelearning.com/ (accessed June 22, 2012).

understanding. Bennett and Gitomer (2009) state that 'knowledge about acquisition of 21st century skills and their development is very limited. Developers of assessments do not yet know how to create practical assessment using even this partial knowledge effectively.'

Most of the research in technology-based assessment (TBA) has dealt with the improvement of assessment of traditional skills (Binkley et al. 2012). It seems that for many challenges in TBA, information and communication technologies (ICT) are the solution to assess 21st century skills but at the same time they are also the problem. On the one hand we lack scientific knowledge that tells us how assessment models and methods need to be adapted so that we can profit from the application of new ICT in assessment and, on the other, we lack practical knowledge about how to create more authentic/natural problem situations with future assessment tools.

Inexpensive technologies for sensors and processing make it possible to integrate them in systems that allow for a more natural form of interaction such as touch, speech, gestures and handwriting. Recent developments and the availability of affordable devices motivated us to develop a so-called tangible user interface (TUI) for assessment. In education, exploratory design-focused studies have suggested that TUIs provide learning benefits, due to the additional haptic dimension, the better accessibility and the shared space that can be used in collaborative situations (Marshall 2007). According to Klemmer et al. (2006) our human bodies and our interactions with physical objects have an essential impact on our understanding of the world. The physical objects and actions of TUIs allow using multiple senses of our human bodies for understanding and experiencing the world. Learning and assessment can be supported by providing new ways of representing problems, detecting the way the users are solving them and directly giving feedback. Nevertheless, no TUI has been systematically used and evaluated in the context of TBA.

Research is therefore necessary to identify tensions where TUI and assessment do not match and assessment situations where this kind of technology is in harmony with the assessment of specific 21st century skills. We developed two test items and conducted two empirical studies. The studies provide detailed insights about how users perceive the device, how they interact with it and how they collaborate to solve a problem. In addition TBA experts were interviewed to discuss tensions and advantages of using a TUI for assessment.

The following section summarises related work and states our research objectives. The two empirical studies and their results are then described. Based on the results of the studies the authors then provide a list of tensions between using a TUI and assessment and also explain where the technology is in harmony with TBA.

Related work

This section refers to 21st century skill frameworks developed recently and reflects upon higher order thinking skills (HOTS) and the use of TUI for learning and assessment in particular.

21st century skills

The OECD[2] developed several programs such as PISA (Programme for International Student Assessment) and PIAAC (Programme for the international Assessment of Adult Competences) to assess skills on a large scale.

Binkley et al. (2012) have organised the 21st century skills into four groups and have proposed a model to assess those. The groups are: ways of thinking, ways of working, tools for working and living in the world. For our work, ways of thinking is the most relevant. These skills refer to:

- creativity and innovation;
- critical thinking, problem solving, decision making;
- learning to learn, metacognition.

Projects such as ATC21S[3] are developing methods to assess skills which are part of 21st-century curriculum. ATC21S concentrates on collaborative problem-solving and learning in digital networks. 'Collaborative problem solving' is defined as the capacity to recognise the perspective of other people in a group, participate, contribute knowledge, experience and expertise in a constructive way, recognise the need for contributions and how to manage them, identify structure and procedure in resolving a problem and build knowledge and understanding as member on a collaborative setting. 'Learning through a digital network' consists of learning as a consumer and producer of information and learning in the development of social and intellectual capital (Griffin et al. 2012).

JISC[4], for example, funds a huge number of technology-based assessment projects in the UK. The Learning Literacies for the Digital Age project (LLiDA) reviewed the kinds of capabilities valued, taught, assessed and defined in competence frameworks. Their 'framework of frameworks' puts emphases on components of digital and learning literacy, which covers also skills in communication, collaboration and problem solving (Leighton 2011). A comprehensive report on assessment in the digital age highlights outcomes of case studies conducted in the UK. The authors conclude that assessment should capture the process in addition to the learning outcome, that it is important to explore the potential of technology to enable the more efficient use of a practitioner's time and effort and to provide a digital environment in which all learning and assessment related activities take place where evidence is consolidated (JISC e-Learning team 2010).

Higher order thinking skills (HOTS)

The term 'thinking' is difficult to define and depends on the view about thinking and the purpose it should serve. In assessing HOTS, Schraw and Robinson (2010) state that thinking and human cognition in general is a goal-oriented activity: 'Thinking is related to gather and evaluate information that is relevant to this goal; it requires constructing a meaning and conceptual representations that support the analysis of events around us; and it requires that we engage in strategic decision making and

[2] Organisation for Economic Co-operation and Development. http://www.oecd.org/ (accessed October 22, 2012).

[3] Assessment and Teaching of 21st Century Skills. http://atc21s.org/ (accessed October 22, 2012).

[4] Joint Information Systems Committee. http://www.jisc.ac.uk (accessed October 22, 2012).

judgments that enhance our ability to self-regulate.' According to Schraw and Robinson, HOTS can be refined to *reasoning* skills, *problem solving* and *critical thinking*, *argumentation* skills and *metacognition* (i.e. thinking about thinking).

The assessment of thinking skills is related to three different outcomes: knowledge, procedure, as well as, attitudes and dispositions (Schraw and Robinson 2011). The first focuses on the assessment of facts, concepts or mental models. Procedural outcomes refer to specific thinking skills and strategies, self-regulatory processes or as stated by Schraw and Robinson, content-based procedural skills. So-called learning traces, which log the actions performed by a learner, are a means to assess procedural outcomes. Analysis of notes written, think aloud protocols or eye tracking are just a few examples for assessing real-time problem solving skills.

Computer-based simulations can assess complex thinking skills that cannot be measured by more traditional assessment methods (Lane 2011). An added benefit of simulations is strategies and processes can be measured.

In the next section tangible user interface will be introduced as a specific type of natural user interface.

Tangible user interfaces (TUIs) for learning and assessment

While current web-based e-assessment frameworks exploit multimedia capabilities of graphical user interfaces to support a large range of different questions types, their possibilities for supporting and measuring team-oriented, creative and communication skills in a collaborative context are extremely limited. Even if some platforms (for example, TAO (Ras et al. 2010)[5] which is currently used in PISA and PIAAC) are designed to accommodate multi-test taker networked assessments together with log collection and analysis capacities, they still lack delivery modes that can capture a wide range of 21st century skills. Hence, we need to explore new technologies offering a better support for collaborative activities, enhanced understanding through more natural activities and realistic problems and finally that allow us to gather activity traces.

Using a new technology in assessment requires a deep understanding of the actions the users perform and the group dynamics in collaborative settings as well as the digital literacy to use this new technology. These aspects may impact the way in which the results of assessments are obtained or they may produce undesired effects.

TUI is a technology to create new types of interactions, combining physical artefacts with digital visualisations in a common interactive space. The idea of TUI is to make computer bits tangible and to allow users to grasp and manipulate them with their hands (Ishii 2008). This concept has a number of advantages for social and contextual interactions, such as collaboration (Hornecker and Buur 2006).

In the literature, we can find a number of examples, demonstrating learning using TUI, without an emphasis on assessment. For example, the Chromatorium (Rogers et al. 2002) is an environment where children can discover and experiment with the mixing of colours. In a similar type of setup, students learned about the behaviour of light (Price et al. 2009). Through physically manipulating a torch and blocks on a table surface, the students explore how projected light beams react and understand

[5] http://www.tao.lu (accessed October 22, 2012).

the concepts of reflection, absorption, transmission and refraction. Another interactive tabletop called TinkerLamp provides a simulation environment for learning purposes where students design a warehouse (Jermann et al. 2009).

Another learning system implements the concept of 'digital manipulatives', computationally-enhanced building blocks which allow the exploration of abstract concepts (Resnick et al. 1998). This principle is followed by SystemBlocks and FlowBlocks, two physical, modular, interactive systems which children can use to model and simulate dynamic behaviour (Zuckerman and Arida 2005).

The approach of concept mapping for self-regulated learning is also documented (Oppl et al. 2010). A tangible tabletop allows users to reflect and evaluate their learning tasks through externalising and representing their knowledge on concept maps represented through physical and digital elements. Users can, for instance, place physical tokens, assign names, make connections and use tokens as containers.

Only a few systems go beyond the learning aspects and implement possibilities for assessment. The Learning Cube (Terrenghi et al. 2006) acts as a tangible learning platform where multiple-choice tests can be taken. By turning the cube, the user selects the right answer for a given question and then shakes it to select it. Another example provides new possibilities for assessing spatial and constructional ability. The Cognitive Cubes (Sharlin et al. 2002) are a set of construction cubes that are provided for building a 3D shape shown on a screen. The change of each shape is recorded and scored for assessment.

Although a number of publications can be found that describe different kinds of strengths of TUI for learning environments, there is still a lack of knowledge on using TUI as a tool for assessment and, specifically, on how to assess the perceived user experience in an assessment context. To the knowledge of the authors, there is no prior work addressing the development and measurement of higher order thinking skills in a mixed physical-digital environment, such as TUIs.

Evaluation of usability and user experience

Tangible user interfaces require new means to evaluate their user experience and usability in an assessment context. User experience refers to the 'way in which people interact with a product over time : what they do and why' (Bevan 2009).User experience is measured to improve both pragmatic and hedonic goals of the user. Conventional usability scales focus on effectiveness, efficiency related to a specific working task and ease of use and satisfaction. Bevan states that the goal of measuring usability is to design systems in order to improve efficiency, effectiveness, user satisfaction and comfort with the system (Bevan 2009). Venkatesh et al. (2003) define performance expectancy as the degree to which an individual believes that using the system will help them to attain gains in a job or specific task. Effort expectancy is defined as the degree of ease associated with the use of the system (Davis 1986).

The simple usability scale (SUS) (Fleck et al. 2009), the technology acceptance model (Bennett 2002) and the unified theory of acceptance and use of technology

(UTAUT) (Davis 1986) do not explicitly distinguish between physical and digital characteristics of TUI. They have been developed to compare software from the usability perspective.

The following research objectives have been formulated to drive this research work.

Research objectives

- Develop a TUI for assessment and investigate its usefulness for assessing collaborative problem solving as well as tensions.
- Conduct case studies in order to understand the interaction between user and TUI on the one hand and communication within the team through the TUI on the other.
- Develop a measurement instrument tailored to evaluate usability and user experience in an assessment context with a TUI.

Case studies

Two empirical TUI studies have been conducted. The first study investigates the use of a matching item to assess the recall of factual knowledge. The second uses a performance item using simulation and shows potential to measure HOTS in the future. Both case studies fit the criteria of collaborative problem solving: working together to solve a common challenge which involves the contribution and exchange of ideas, knowledge or resources to achieve the goal.
The purpose of both studies was explorative, which meant to observe human-to-human behaviour and interactions between each individual and the device. The goal was not to develop a comprehensive assessment model and method. This will be a future goal after having observed the system in practice first.

Controlled experiment: Matching item

The task was to assign the correct name to the corresponding image of a planet (according to QTI[6] standard: 'associate' item). When a user places a card, holding the name, onto a planet, the user receives immediate feedback in the form of a red (false answer) or green (correct answer) circle that surrounds the planet. The system counts the number of attempts (i.e. the wrong pairing of a card and a planet's image was counted as one failed attempt) and the time needed to solve the task.

According to the revised taxonomy of Bloom, such a task can be classified under the level 'remember' and the cognitive process 'recognise' (Anderson and Krathwohl 2001). Facts are recalled from long-term memory and associated with the projected images of the planets. Hence the task consists in matching verbal to pictorial representations of the solar system objects.

Research questions and study design

The following research questions were addressed:

- Do groups perform significantly better compared to individuals with regard to the number of attempts and time needed (in seconds) to solve the planet task? (RQ1)
- Do groups rate the usability significantly higher than individuals? (RQ2)
- Which characteristics of the physical space are used by groups to solve a problem collaboratively? (RQ3)

[6] IMS Question and Test Interoperability Specification. http://www.imsglobal.org/question (accessed October 22, 2012).

Answering these questions would help future development of, for example, logging mechanisms to detect solving strategies of groups or simply to understand better from a user experience point of view how to develop such systems (see also next study for a more detailed exploration of user experience).

In order to answer these questions a between-subject, controlled experiment was conducted with a control group of eleven individuals solving the test items alone, as well as 24 participants divided into eight groups of equal size (i.e. experimental group). The participants were randomly selected from the research department and assigned to both groups with an average age of 32 years.

For the experiment, we set up a tangible tabletop system, based on the optical tracking framework reacTIVision[7]. The worktop has an interactive area of 75 x 100 cm. On the table, we projected an image of the solar system, showing the sun, the eight planets and Pluto. The position of the planets did not reflect the distance to the sun. In addition we created nine cards each with 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 solving activity was captured by a video camera and at the end we distributed a questionnaire with questions about background knowledge of astronomy and the system usability scale (SUS) (Brooke 1996). In addition we took notes on the performance of the group and how the users placed themselves around the table. The set up for the experiment is shown in figure 1.

Results

RQ1: An interesting outcome was that the performance on solving the test item was almost the same except the group setting needed a slightly lower number of attempts (M = 5.88, SD = 4.96) than the individuals (M = 6.50, SD = 4.79). This was probably due to the fact that almost every group was involved in discussion before they dropped the cards on the table.

RQ2: A non-parametric test (P =0.008) revealed that the control group (individuals) even rated the usability significantly higher (M = 90.0, SD = 5.40) than the subjects in the collaborative setting (M = 81.7, SD = 7.47).

The statistical dispersion of the group setting regarding usability was higher, therefore it makes sense to investigate the more extreme usability ratings and their reasons. First, comments were given by the subjects immediately after the experiment related to the table itself such as different, position dependent, illumination of the table; the height of the table was problematic for smaller people; others complained about incorrect detections of the markers (comment: the number of attempts has been corrected accordingly). Also, the system usability scale does not cover usability from a group perspective and so the question is whether it is valid to use such instruments to measure usability in a collaborative setting. SUS is also not tailored to natural user interface and does not reflect the assessment context. For the next study an adapted usability scale was developed to address this issue.

[7] A toolkit for tangible multi-touch surfaces. http://reactivision.sourceforge.net/ (accessed October 22, 2012).

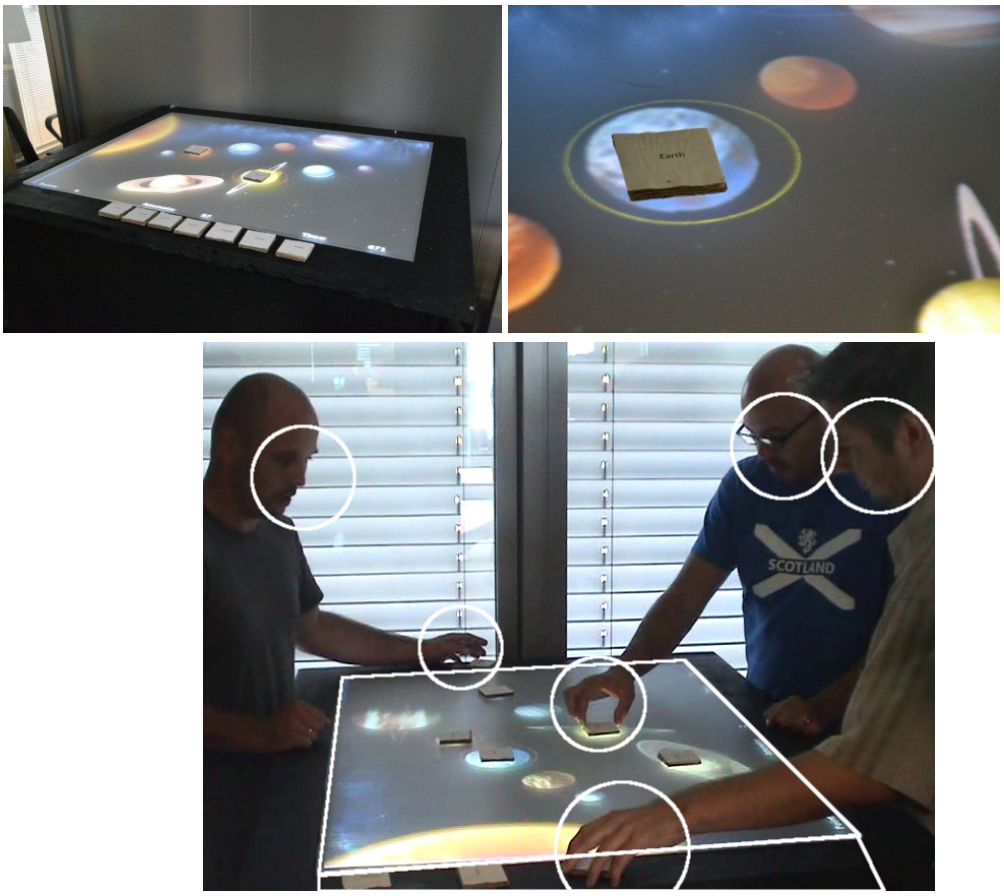

Figure 1: The tabletop provides a shared space where gazes, body postures, gestures, physical actions, interaction area and storage places meet

RQ3: A video analysis of the group situation provided a range of insights concerning the aspects of the physical space that supported collaborative problem solving on a TUI (Maquil and Ras 2012). The video data was analysed using the CLM framework (Fleck et al. 2009) and the different mechanisms of collaborative learning (i.e. making and accepting suggestions, negotiating, joint attention and awareness, narrations) have been extracted. The goal of the analysis was to extract different types of behaviour from this simple collaborative problem solving task and to identify the physical aspects which support this behaviour. For each mechanism of collaborative learning, a set of key scenes have been extracted and analysed based on *where* the interaction took place, *what* was physically manipulated and *who* was either active or passive subject of the interaction.

The analysis revealed that three major characteristics of the physical space were intensively used during the collaborative problem solving activity. The *physical interaction objects* allowed users to perform a variety of actions, reaching beyond direct control of the system. For instance, participants held a few of them in their hands to concentrate the discussion on the assignment of this subset of planets. In other situations, participants placed the cards with some noise on the table, in order to increase the awareness of other participants.

A second physical characteristic was the *shareability of the space*, which was provided by the simultaneous visibility of actions on the projection, the space between participants and the storing place of the objects. The space was used

extensively for a variety of actions, reaching from pointing actions, over physical manipulations, to observations.

Finally, we identified the use of *non-responsive spaces*, i.e. spaces where no input to the system is provided and that were used to demonstrate suggestions, exchange ideas and set a common focus. For instance, it was common practice to first hold a card above a planet for some seconds before dropping it which allowed the other participants to intervene or negotiate if not in agreement.

Case study: Simulation item

For the second prototype, we implemented a simulation task. The aim for the users was to explore and understand the relation of external parameters to the production of electricity in a windmill. In contrast to the previous application, this task can assess complex problem solving and reasoning skills, i.e. HOTS, but again, the goal was to observe the interaction between the users and the system, not to develop a perfect assessment model and method.

Using physical objects, users can change parameters such as wind speed, number of blades and height of the windmill (figure 2). Each of the objects represents one of the parameters; the value is increased or decreased by rotation. The output variables (i.e. rotor speed and energy produced) are represented through additional physical objects. Both output parameters and the look of the windmill (height or number of blades) changed real-time according to the manipulations of the input variables. Because all the tangibles can be moved freely on the table and exchanged, each participant gets a vote and hence collaboration and motivation was expected to be improved.

Figure 2: Simulation item 'windmill' (left); three users collaborating (right)

The simulation task was implemented on the same tangible tabletop system that was used for the matching item. New, round, physical objects have been created and labeled with text and a small image. The task was described as follows: '*The task is to find out which factors have an influence on the amount of electric power produced by the windmill and to answer the multiple-choice questions. Please access the simulation environment by clicking "start" and explore and try to understand the environment, the parameters and the effect the parameters have. When you are done we will ask you to answer a small questionnaire about this simulation.'*

Research questions and study design
The following research questions were investigated:

- How could we measure usability and user experience in an assessment context? Do the users significantly rate the usability and user experience better than average? (RQ4)
- What physical and digital aspects are particularly useful or critical for effectively and efficiently using the TUI? (RQ5)
- How do the subjects proceed to solve the complex problem solving item? What are typical solving patterns? (RQ6)

In a case study, we asked ten groups of three people each to collaboratively explore and understand the relation of the parameters of the windmill. The participants were randomly selected from the research department and divided into groups by taking into account their language skills, so that discussions could take place in their preferred language (French or English). Information on the task was projected as a text onto the table. Participants were given explanations about the purpose of the evaluation, then asked to read the text and finally asked to start the simulation environment. The experimentation was recorded with two cameras and at the end participants were asked to fill out two questionnaires. While one questionnaire aimed to assess the knowledge of the participants gained during the exploration phase, the second one consisted of questions on the usability and user experience of the system. Further, a researcher directly observed the groups and took notes of the solving strategies and arising usability issues. A group interview with three TBA experts was made after they participated in the experimentation in the same group.

Results
RQ4: For this study a new questionnaire was developed to access usability and user experience. The goal was to adapt and combine existing well accepted constructs to the assessment context and tangible interfaces in general. The main constructs measured for usability were performance expectancy, pragmatic quality of the physical and visual objects (windmill and tachometers) and the effort expectancy. The user experience was assessed with six items. Table 1 lists the questions and their origin, i.e. from which measurement construct the item was reused or even adapted.

Each item has a 7-point Likert scale with the neutral value of 4. After keying in the data, the polarity of some items was reversed (marked with ‘recoded’ in table 1). A reliability test of the five different constructs was done. Table 2 shows the descriptive results as well as the reliability for each scale.

Fisseni states that a Cronbach’s alpha <0.80 should be considered as small, 0.80-0.90 as medium and >0.90 as a high reliability (Fisseni 1997). The results in the table show that four out of five constructs obtained low to highly reliable scales. The pragmatic quality of physical objects has the lowest reliability. But Cronbach’s alpha depends on the number of items that belong to a scale: a higher number of items that positively correlate amongst each other produce higher reliability levels. The pragmatic quality of physical object constructs has only four items. Nevertheless, a more detailed analysis of the constructs is necessary to adapt the questionnaire before its next use.

Table 1: Usability and user experience questionnaire

Performance expectancy	**Origin***
1. The system enables me to identify the structure and procedure in exploring the simulation (recoded)	1. TAM**
2. Using the physical objects enables me to change parameters (recoded)	2. TAM**
3. The visual feedback enables the group to understand the effect of the changing parameters (recoded)	3. TAM**
4. The simulation environment enables the group to explore all different parameters settings (recoded)	4. TAM**
5. The 'table' setting enables me to collaborate with the other participants (recoded)	5. New
6. The simulation environment enables the group to answer the questions correctly (recoded)	6. New
Pragmatic quality: Are the physical objects (table and tangibles)	**Origin***
1. artificial or natural?	1. AttrakDiff
2. complicated or simple?	2. AttrakDiff
3. impractical or practical?	3. AttrakDiff
4. cumbersome of straightforward?	4. AttrakDiff
Pragmatic quality: Are the visual objects (windmill and tachom.)	**Origin***
1. artificial or natural?	1. AttrakDiff
2. complicated or simple?	2. AttrakDiff
3. impractical or practical?	3. AttrakDiff
4. cumbersome or straightforward?	4. AttrakDiff
5. unpredictable or predictable?	5. AttrakDiff
6. confusing or clearly structured?	6. AttrakDiff
Effort expectancy	**Origin***
1. I found the system easy to use (recoded)	1. TAM/SUS
2. I found the simulation environment unnecessarily complex to use	2. SUS**
3. I think that I would need the support of a skilled person to be able to use the system (recoded)	3. SUS
4. I would imagine that most people would learn to use the system very quickly (recoded)	4. TAM
5. The amount of mental effort spent to solve the task was high for me (thinking, deciding, calculating, remembering, looking, searching)	5. New
6. Using the system was (frustrating … fulfilling)	6. AttrakDiff
7. Using the system was (tense … relaxed)	7. AttrakDiff
User experience: The system is	**Origin***
1. conventional or inventive	1. AttrakDiff
2. unimaginative or creative	2. AttrakDiff
3. conservative or innovative	3. AttrakDiff
4. dull or captivating	4. AttrakDiff
5. undemanding or challenging	5. AttrakDiff
6. ordinary or novel	6. AttrakDiff

*TAM (Davis 1986); SUS (Brooke 1996); UTAUT (Venkatesh et al. 2003);AttrakDiff (AttrakDiff 2012)

** adaptation to TBA context

Table 2: Descriptive statistics and reliability

	N	Range	Min	Max	Mean	Std. Dev.	Cron. Alpha
Performance expectancy	28	2.83	4.17	7.00	5.84	.756	.703
Pragmatic quality of physical objects	30	4.25	2.50	6.75	5.48	.832	.618
Pragmatic quality of visual objects	29	3.17	3.83	7.00	5.69	.826	.722
Effort expectancy	30	3.43	3.57	6.29	5.91	.822	.749
User experience	28	4.17	2.83	7.00	5.58	.960	.860

By looking at the descriptive statistics of the constructs, we observe that they all pass the neutral value of 4 – users have rated the system positively; a one-sample t-test confirmed that the participants significantly rate the system better than neutral. Looking at the minimum values obtained, we can see that some users still rate the system much lower than neutral. Therefore, the video analysis could help to identify the problems related to usability and user experience which lead to these low values.

RQ5: To identify usability issues, we used direct observation by one observer, who is expert in TUI design and combined this data with the results of the questionnaires. The collected insights allowed us to identify advantages and challenges of TUIs dealing with the physical objects themselves, as well as the digital projection.

TUI aspects found to be critical from the perspective of usability were:

- While most of the groups immediately rotated the objects after placing them on the table, three groups first used their finger to move the needle of the tachometer. In a discussion the participants explained that this was due to the previous screen, where the button was activated by using touch.
- All of the groups experienced problems to understand that the values of the output tangibles cannot be modified through turning. Users explained that this was due to the same round shape as the input tangibles, so they assumed the need to be manipulated the same way.
- In several cases the system's precision lead to doubts about the influence of some parameters. As there was some jitter in the detection of the orientation of the objects, the value of the wind speed was always slightly changing, creating a permanent modification of the energy production - participants were often unsure if this was caused by some action from their side or not.

TUI aspects found to be advantageous from the perspective of usability and user experience were:

- All groups understood how to manipulate the parameters after a short trial phase without external support. This is confirmed through the questionnaire, where users rated the system as quick to learn (M = 6.03, SD = 1.21).
- Users found that the physical objects allowed them to change parameters very quickly (M = 6.07, SD = 0.98).
- The physical objects and the visual objects (windmill and tachometer) were also highly rated as simple in their pragmatic quality (physical objects: M = 6.17, SD = 0.87; visual objects: M = 6.27, SD = 0.78).

- Users found that the amount of mental effort spend to solve the task was particularly low (M = 6.13, SD = 0.73).
- From a user experience point of view, users highly rated the system as being 'relaxed' (M = 6.17, SD = 0.87) and 'captivating' (M = 6.10, SD = 0.96).

RQ6: To answer this question, we used direct observations and combined them with an additional analysis of the video data.

The results revealed that all groups proceeded to solve the task in a similar way. After having read the introductory text, the users of all groups verbally agreed and one pressed 'start'.

All groups followed up with a phase where they explored how the tool worked and explored what could be modified in the simulation. This phase was done in two different ways. Half of the groups did this first exploration simultaneously and without talking (figure 3). Each of the participants individually grasped a few tangibles and moved and rotated them in parallel until they understood the tool and the task. After that, these groups changed their way of working and started to try out individual modifications one by one. One participant suggested something to try, which was then realized, observed and commented on by the whole group.

Figure 3: Working simultaneously during the exploration of the problem space (left); parameters with no influence were placed on the border of the table (right)

The other half of the groups did not first explore the tool and task simultaneously, but immediately performed actions stepwise, by leaving time for observing and commenting what happened. In this case, the first actions also dealt with understanding the problem space, whereas the later actions aimed to fully comprehend the type of relation between parameters.

During this process of understanding the relations between the parameters, several patterns emerged that could be observed in most of the groups. To identify the relation between two parameters, participants usually manipulated one parameter with one hand, while gazing at another parameter or the windmill. Furthermore, during the discussion phases, pointing gestures were intensively used to express and coordinate themselves, pointing to tangibles and visualizations in the common, shared space. Finally, some groups made use of the borders for facilitating the understanding. They put the parameters that had no effect on the side, in order to concentrate their work on the remaining ones (figure 3).

The solving patterns for the various groups are summarized in table 3.

Table 3: Solving patterns

Solving pattern	**Number of groups**
Verbally agree before pressing the start button	10 groups
Rotate one parameter while observing another one	10 groups
Work simultaneously during first exploration phase *or* Immediately work in small steps that are planned, observed and commented	6 groups 4 groups
Use pointing gestures during discussions	10 groups
After first exploration, perform step-wise different actions that are planned, observed and discussed by the whole group	10 groups
Use the borders of the table to place parameters that have no influence	4 groups

Discussion

To identify harmony and tensions of using TUIs in computer-based assessment, we grouped the different results from the empirical studies and analysed their advantages and disadvantages from a TBA point of view. These insights were then discussed in a group interview with experts on TBA after they had participated in the case study. Next, we list the most important tensions and also the reasons why TUI are in harmony with the requirements of TBA.

A) Sharing different spaces

The simultaneous visibility of actions on the projection, the space between the participants and different storing places support the user to intensively communicate using gestures, talk and by physically manipulating objects. The large size of the space increases the possibility of more than one person getting actively involved – this ensures collaboration. Nevertheless, there is a challenge for ensuring that every user has the same opportunities to interact, since the view on the different spaces depends on the position at the table. The users could even feel responsible for a space on the table or specific tangibles.

B) Relevance of non-responsive spaces

Spaces where the TUI does not interact with the user are essential to solve a task. These spaces were used for offline interactions, i.e. to understand the problem or even prepare and discuss a potential solution in a group, to make individual suggestions, or to sort out irrelevant tangibles. However, these activities are part of the solving strategy and therefore need to be tracked by the system. The technological means to track and interpret these activities are more difficult to realise.

C) Versatility of physical interaction objects

Physical objects allow manipulating parameters in a natural and simple way. They even enable more efficient interactions, such as turning a parameter without looking at it in order to be able to monitor another value. Beyond that, users can hold them in their hands, use them to make gestures and suggestions, or accompany the

interaction with the object with noise to increase awareness in the group (putting it loudly on the table). Nevertheless adding physicality to the objects can also lead to wrong expectations to interact with them (for example, a round object does not have to mean that its rotation will impact something); the user holding objects can impact the solving strategy or harm other users to contribute. Furthermore, even if physical objects can be manipulated precisely by a user, this does not necessarily mean that the system has the same precision for recognizing these manipulations. Hence, there is a challenge for adapting the online and offline interactions to the versatility of interaction objects.

D) User experience versus usability
Some features may increase user experience, such as providing different modalities to the user to interact with the system. Offering touch and the possibility to manipulate physical objects can decrease the understandability and learnability of the system, if interactions with touch and physical interactions are used in an inconsistent way.

E) Identifying experimental units
The parameters can be changed quickly which is an advantage to understand the simulation. However, the challenge for assessing problem solving skills then becomes to recognize these small experimental 'units' and track them as atomic contributions of the participants.

F) Complexity of the task
Our item was rather simple according to the number of parameters and the scenario chosen (windmill simulation). This was confirmed by the fact that the users rated the 'mental effort required' as low. We therefore decided to keep the task description more general (see above) and let the users decide how to proceed and when to stop. The questionnaire was given to the users when they finished their experiments with the TUI. The experts criticised the fact that the users were not allowed to use the TUI while answering the questions as this would allow better predictions of the solving strategy and its scoring. Nevertheless, a more complex scenario would require more self-regulation skills from the users.

G) When does the solving activity really start?
The studies showed that the participants first explored how the system works and how the parameters could be manipulated. More than half of the users started to interact with the table individually without observing the activities of the other team members. This resulted in sometimes chaotic sequences of changing parameters followed by users suggesting hypotheses of how parameters impact others without any deeper reflection. For assessment, the question is to detect when the learning phase has finished and when a system should start tracking the actual solving strategy.

H) Recognize perspectives of others
According to the experts, the TUI supports the users in recognizing and understanding the perspectives of others – an important skill of collaborative problem solving. Hypotheses can be easily explained and tested using the different modalities. A perspective of an individual for example, a misunderstanding) can be recognized by the others through realising it on the table. As mentioned before, the

simulation item requires that the users actively synchronize their actions of manipulating and observing the environment (i.e. to avoid parallel modifications that cannot be traced). This helps them to work collaboratively on one single experimental unit.

Conclusion

We agree with Binkley et al. who argue that technology-based assessment 'has the potential to support educational innovation and development of 21st century skills, such as complex problem solving, communication, team work, creativity and innovation' (Binkley et al. 2012).

Both examples provided insights on how the TUI is used and experienced while solving a simple and a complex item. Several spaces are used besides the interactive surface of the table to solve the items. The physicality of the objects enables the participants to interact with the system and with each other and, hence, positively impacts the usability and user experience. Further, users can recognize the perspective of other users, they can conduct small experiments collaboratively and, therefore, build knowledge and understanding in a team. However, adding physicality to an interface also brings up new challenges. Users cannot face exactly the same conditions around the table and interactions may be very unstructured and distributed across different spaces and modalities making them difficult to trace for the system. Furthermore, manipulation possibilities may be new for the users, requiring a learning phase at the beginning.

For the future, the table will allow the use of learning traces and hence support the tracking of real-time problem solving items. This is especially interesting to get indicators for self-regulation behaviours. The self-regulatory component of HOTS is considered essential to develop higher order cognitive abilities (Leighton 2011). Further, other researchers have found evidence that all human cognition, from simple *recall* to *evaluation* (Anderson and Krathwohl 2001) and problem solving refers to both declarative knowledge, as well as procedural knowledge. The classification of lower or higher order thinking is based on the 'organization and cohesion of networks of declarative and procedural knowledge structures' (Leighton 2011). This evidence must be considered for the design of TUIs for assessment purposes.

The simulation item used in our studies were rated as simple (83% of the maximum score was achieved on average). Currently, a TUI application with a more complex environment for traffic simulation is being developed. A larger number of parameters and their complex interdependencies will require that the users systematically experiment to answer the questions. This application will provide ideal settings for assessing HOTS including self-regulating behaviour in a collaborative, realistic setting.

Acknowledgments

Many thanks go to Katja Weinerth and Bob Reuter of the University of Luxembourg who designed the windmill task. We are grateful to the college students of the Lycée technique d'Esch who developed a first prototype for the simulation item and Warda Atlaoui for building the table and the matching item. Further, we very much appreciate the voluntarily participation of our colleagues in our study. Thanks to Eric Tobias for commenting and revising earlier versions of this article.

References

Anderson, L.W., and D.R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Complete ed. New York: Longman.

Attrakdiff. *A tool for measuring hedonic and pragmatic quality*. www.attrakdiff.de (accessed February 19, 2013).

Bennett, R.E. 2002. Inexorable and inevitable: The continuing story of technology and assessment. *Technology, Learning, and Assessment* 1, 1.

Bevan, N. 2009. What is the difference between the purpose of usability and user experience evaluation methods? In *Workshop User eXperience Evaluation Methods in Product Development (UXEM'09 Workshop at 12th Conference on Human-Computer Interaction, INTERACT 2009)*. Uppsala, Sweden.

Binkley, M., O. Erstad, J. Herman, S. Raizen, M. Ripley, M. Miller-Ricci and M. Rumble. 2012. Defining twenty-first century skills. In *Assessment and teaching of 21st century skills*, ed. P. Griffin, B. Mcgaw and E. Care. 17-66. Dordrecht: Springer.

Brooke, J. 1996. Sus - a quick and dirty usability scale. In *Usability evaluation in industry*, ed.. P.W. Jordan, B. Thomas, B.A. Weerdmeester and I.L. Mcclelland London: Taylor & Francis.

Davis, F.D. 1986. A technology acceptance model for empirically testing new end-user information systems: Theory and results. PhD thesis, Massachusetts Institute of Technology.

Fisseni, H. 1997. *Lehrbuch der psychologischen diagnostik*. 2. Aufl. ed. Göttingen: Hogrefe.

Fleck, R., Y. Rogers, N. Yuill, P. Marshall, A. Carr, J. Rick and V. Bonnett. 2009. Actions speak loudly with words: Unpacking collaboration around the table. Paper presented at the ACM International Conference on Interactive Tabletops and Surfaces, November 23-25, in Banff, Canada.

Griffin, P., E. Care and B. Mcgaw. 2012. The changing role of education and schools. In *Assessment and teaching of 21st century skills*, ed. P. Griffin, B. Mcgaw, and E. Care. 1-15. Dordrecht: Springer.

Hornecker, E. and J. Buur. 2006. Getting a grip on tangible interaction: A framework on physical space and social interaction. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems 2006 (CHI 2006), in Montreal, Quebec, Canada.

Ishii, H. 2008. Tangible bits: Beyond pixels. Paper presented at the 2nd international conference on Tangible and Embedded Interaction, in Bonn, Germany.

Jermann, P., G. Zufferey, B. Schneider, A. Lucci, S. Lépine and P. Dillenbourg. 2009. Physical space and division of labor around a tabletop tangible simulation. Paper presented at the 9th international Conference on Computer supported collaborative learning, in Rhodes, Greece.

JISC E-Learning Team. 2010. Effective assessment in a digital age - a guide to technology-enhanced assessment and feedback. In *Technology enhanced Assessment*, ed. JISC. Bristol: JISC.

Klemmer, S.R., B. Hartmann and L. Takayama. 2006. How bodies matter: Five themes for interaction design. Paper presented at the 6th Conference on Designing Interactive Systems (DIS 2006), June 26-28, in University Park, PA, USA.

Lane, S. 2011. Issues in the design and scoring of performance assessments that assess complex thinking skills. In *Assessment of higher order thinking skills*, ed. G. Schraw and D.R. Robinson. Charlotte, North Carolina: IAP-Information Age Publishing Inc.

Leighton, J.P. 2011. A cognitive model for the assessment of higher order thinking. In *Assessment of higher order thinking skills*, ed. G. Schraw and D.R. Robinson. Charlotte, North Carolina: IAP-Information Age Publishing Inc.

Maquil, V. and E. Ras. 2012. Collaborative problem solving with objects: Physical aspects of a tangible tabletop in technology-based assessment. Paper presented at the 10th International Conference on the Design of Cooperative Systems - From research to practice: Results and open challenges (COOP 2012), 30 May - 1 June, in Marseille, France.

Marshall, P. 2007. Do tangible interfaces enhance learning? Paper presented at the Proceedings of the 1st international conference on Tangible and embedded interaction, in Baton Rouge, Louisiana, USA.

Oppl, S., C.M. Steiner and D. Albert. 2010. Supporting self-regulated learning with tabletop concept mapping. In *Interdisciplinary approaches to technology-enhanced learning*, ed. M. Mühlhäuser, W. Sesink, A. Kaminski and J. Steimle. Waxmann Verlag.

Price, S., T. Pontual Falcão, J.G. Sheridan and G. Roussos. 2009. The effect of representation location on interaction in a tangible learning environment. Paper presented at the 3rd International Conference on Tangible and Embedded Interaction, in Cambridge, UK.

Ras, E., J. Swietlik, P. Plichart and T. Latour. 2010. Tao - a versatile and open platform for technology-based assessment. In *Sustaining TEL: from innovation to learning and practice, 5th European Conference on Technology Enhanced Learning (EC-TEL 2010)*, ed. M..Wolpers, P.A. Kirschner, M. Scheffel, S.Lindstaedt and V. Dimitrova, 644-49. Barcelona, Spain: Springer.

Resnick, M., F. Martin, R. Berg, R. Borovoy, V. Colella, K. Kramer and B. Silverman. 1998. Digital manipulatives: New toys to think with. Paper presentat at the Proceedings of the SIGCHI conference on Human factors in computing systems, in Los Angeles, California, USA.

Rogers, Y., M. Scaife, S. Gabrielli, H. Smith and E. Harris. 2002. A conceptual framework for mixed reality environments: Designing novel learning activities for young children. *Presence: Teleoper. Virtual Environ.* 11, 6: 677-86.

Schraw, G. and D.R. Robinson. 2011. *Assessment of higher order thinking skills*. Charlotte, North Carolina: IAP-Information Age Publishing Inc.

Sharlin, E., Y. Itoh, B. Watson, Y. Kitamura, S. Sutphen and L. Liu. 2002. Cognitive cubes: A tangible user interface for cognitive assessment. Paper presented at the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, in Minneapolis, Minnesota, USA.

Terrenghi, L., M. Kranz, P. Holleis and A. Schmidt. 2006. A cube to learn: A tangible user interface for the design of a learning appliance. *Personal and Ubiquitous Computing* 10, 2: 153-8.

Venkatesh, V., M.G. Morris, G.B. Davis and F.D. Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly* 27, 3: 425-78.

Zuckerman, O., and S. Arida. 2005. Extending tangible interfaces for education: Digital montessori-inspired manipulatives. Paper presented at the SIGCHI conference on Human Factors in Computing Systems, in Portland, USA.