Interaction Evaluation for Human-Computer Co-creativity: A Case Study

Anna Kantosalo, Jukka M. Toivanen, Hannu Toivonen
Department of Computer Science and Helsinki Institute for Information Technology HIIT
University of Helsinki, Finland
anna.kantosalo@helsinki.fi, jukka.toivanen@cs.helsinki.fi, hannu.toivonen@cs.helsinki.fi

Abstract

Interaction design has been suggested as a framework for evaluating computational creativity by Bown (2014). Yet few practical accounts on using an Interaction Design based evaluation strategy in Computational Creativity Contexts have been reported in the literature. This study paper describes the evaluation process and results of a human-computer co-creative poetry writing tool intended for children in a school context. We specifically focus on one formative evaluation case utilizing Interaction Design evaluation methods, offering a suggestion on how to conduct Interaction Design based evaluation in a computational creativity context, as well as, report the results of the evaluation itself. The evaluation process is considered from the perspective of a computational creativity researcher and we focus on challenges and benefits of the interaction design evaluation approach within a computational creativity project context.

Introduction

Evaluation is vital for guiding the development and measuring progress in computational creativity methods (Jordanous 2012). Especially formative feedback is needed to guide practical development work (Jordanous 2012). This is also true for interactive systems based on computational creativity methods, including human-computer co-creative systems – systems in which both the human and the computer take creative responsibility of the output of the program. As new human-computer co-creative systems are created we will need to address issues in their evaluation.

Bown (2014) argues for a more contextually based evaluation of creative systems within their cultural environments. We consider this to be true especially for human-computer co-creative systems as an evaluation focusing merely on the computational system's creativity is not sufficient to evaluate the success and progress of the system with regard to the user's creative process or the co-creative experience itself. Methods incorporating the user's perspective are needed for incorporating these aspects. Bown (2014) suggests learning from contextually and culturally aware evaluation methods intended for end-user evaluation established in the field of Interaction Design.

In this study paper, we first briefly discuss the similarities and differences between human-computer co-creativity evaluation and computational creativity evaluation. We then proceed to view Interaction Design in the context of computational creativity: We see how Interaction Design currently connects to computational creativity and view previous human-computer co-creation and creativity support system evaluation projects in the light of the DECIDE framework (Rogers, Sharp, and Preece 2011). Then, we move on to discuss our own case study of the Poetry Machine evaluation and illustrate how the DECIDE framework works in practice in the context of a human-computer co-creativity system evaluation. Next, we present the results of our evaluation case study and finally discuss our findings and the usefulness of this evaluation with regard to computational creativity development.

Evaluating Computational Creativity and Human-Computer Co-Creativity

Evaluation of computationally creative systems may focus on different levels of the system: According to Colton and Wiggins (2012), a distinction is often made between evaluating the "cultural value of the artefacts produced by systems, and tests which evaluate the sophistication of the behaviours exhibited by such systems". Jordanous (2012) supports a similar idea in her analysis of existing evaluation frameworks. According to Yannakakis et al. (2014), this characterization of evaluation also applies for the evaluation of co-creativity. Yannakakis et al. continue that the evaluation of the final outcomes of a co-creative process may utilize same approaches as the evaluation of the outcomes of an independent computationally creative process but the process itself is more difficult to evaluate because of the unknown nature of the human creativity process itself. In this paper, we have focused on the evaluation of the process aspects and left out the evaluation of the artefacts. However, the evaluation of artefacts can also factor into evaluating the effects and benefits of the co-creative system to its users.

Jordanous (2012) notes that computational creativity evaluation has traditionally favored expert evaluation, although the evaluation of computational creativity systems with target users has been discussed. There are still few practical examples describing the end-user-evaluation of either autonomously creative or co-creative systems. In this paper, we hope to provide the field with a practical example of how

end-user evaluation of computational creativity software involving users can be conducted in practice at early development stages.

One important difference between evaluating autonomous computational creativity systems and human-computer co-creative systems seems to be that the subjective experience of the human user of a co-creative system becomes an interesting evaluation target. Therefore, the focus of evaluating co-creative systems can not be only on evaluating the creativity of the system, but also in part on the effects the system has on the user. Yannakakis et al. (2014) conclude that the interaction between the human and the computer fosters the creativity of the tool, but the claim cannot be thoroughly evaluated with current frameworks.

Finally, Jordanous (2012) divides the evaluation of computational creativity systems to summative and formative evaluation. The purpose of the former is to provide a summary of a system's creativity, while the latter aims to provide constructive feedback on the system. A similar distinction is made by Hartson et al. (2003) for Interaction Design evaluation methods, with the distinction that formative evaluation is usually done iteratively during product design and summative evaluation is usually reserved for finished designs or comparisons between designs. Jordanous (2014) seems to consider formative evaluation a more important goal for current computational creativity evaluation procedures, as she regards the usefulness of evaluation results as an evaluation criteria for evaluation methods themselves. This paper focuses on the formative evaluation of an on-going project, aiming to produce results that are useful for guiding the future development of the poetry writing tool.

Interaction Design and Evaluation in Computational Creativity Contexts

The field of Interaction Design studies how to best design interactive products to facilitate human interaction and communication. As such, it seems ideal for designing human-computer co-creative tools. Interaction Design covers a multitude of design fields and approaches, such as user-centered design (Rogers, Sharp, and Preece 2011). As a methodological framework it offers iterative processes and methods for designing and evaluating interaction in specific contexts. Some Interaction Design methods have already been used in designing software based on Computational Creativity methods (Kantosalo et al. 2014).

Bown (2014) argues that the wide range of robust Interaction Design methods for observing and measuring user experience could help build a thorough empirical grounding for Computational Creativity evaluation. He continues that Interaction Design would also help to establish commonly used evaluation concepts – 'value' and 'novelty' – as constructs immediately related to the goals of the individual user. This new human-centered approach would shift the nature of the enquiry very slightly "by asking not how creative a system is, or whether it is creative by some measure, but how its creative potential is practically manifest in interactions with people."

In this section, we provide a brief review of Interaction Design evaluation in creative contexts. We cover human-computer co-creativity projects STANDUP (Waller et al. 2009), Scuddle (Carlson, Schiphorst, and Pasquier 2011), Evolver (DiPaola et al. 2013), and the Sentient Sketchbook (Yannakakis, Liapis, and Alexopoulos 2014). They all have used evaluation methods that can be seen to fall within the scope of Interaction Design. To learn more about how the creative context should be considered in Interaction Design evaluation, we include six creativity support systems that have been evaluated in the literature: the IdeaManager (Shibata and Hori 2002), a Virtual Musical Environment (VME) (Johnston, Amitani, and Edmonds 2005), the Envisionment and Discovery Collaboratory (EDC) (Warr and O'Neill 2007), the Choreographer's Notebook (Singh et al. 2011), Ugobes Pleo (Ryokai, Lee, and Breitbart 2009), and Parallel Pies (Terry et al. 2004).

We structure the review, and our subsequent description of how we evaluated the Poetry Machine, according to the DECIDE framework by Rogers et al. (2011). The DECIDE framework is a checklist with the following six items:

1. Determine the goals
2. Explore the questions
3. Choose the evaluation methods
4. Identify the practical issues
5. Decide how to deal with the ethical issues
6. Evaluate, analyze, interpret, and present the data

Each step of the framework guides the next step: Determining goals helps designers to ask relevant study questions, and questions guide the selection of methodologies. Then again, the selected methods predict some of the practical issues, which may be related to ethical questions. Finally, all previous factors are relevant to deciding how the data is best evaluated, analyzed, interpreted, and presented.

Determining Evaluation Goals

Choosing what to evaluate is often a challenge in the creative domains. Some projects attempt to measure the increase in creativity of the user, some discuss the creativity of the system, while some focus on user experiences and feedback. Carroll (2011) has noted that because creativity is difficult to define, it is often difficult to say if tests designed to measure creativity of an interactive system actually measure creativity or some other construct. Additionally, aspects of creativity may be domain specific (Carroll 2011).

It is surprising that only two of the reviewed human-computer co-creativity evaluation projects state their goals explicitly: Waller et al. (2009) investigated if their target group is capable of using the STANDUP system, and how they use it. Yannakakis et al. (2014) studied if the Sentient Sketchbook fostered the designer's creativity, specified as aspects of lateral thinking and diagrammatic reasoning. In evaluations of creativity support tools, goals have included gathering initial feedback (Johnston, Amitani, and Edmonds 2005), evaluating if the tool supports specific aspects of a creative process (Warr and O'Neill 2007; Singh et al. 2011), or what is the role of the tool in a creative process (Ryokai, Lee, and Breitbart 2009).

Exploring the Questions

Exploring the questions means the redefinition and focus of the goals to more operational questions (Rogers, Sharp, and Preece 2011). Among the Human-Computer Co-Creativity evaluation examples, only Yannakakis et al. (2014) further explain their evaluation targets as the degree and quality of use of the suggestions of a computational partner. As a type of elaboration for their implicit goals DiPaola et al. (2013) provide the set of actual questions used in their study. Among the creativity support systems, Singh et al. (2011) provide a similar list of questions asked from their users and Johnston et al. (2005) list the specific behaviors of the system they want to investigate.

Choosing Methods

There is a wide range of Interaction Design evaluation methodologies, including formal vs. informal testing methods, thinking aloud vs. observation, and summative vs. formative testing (Lewis 2006). It is common for designers to combine different methods to gather rich data (Rogers, Sharp, and Preece 2011). Mixed-methods approach combining quantitative and qualitative data is also the evaluation recommendation of the NSF Workshop on Creativity Support Tools (Carroll 2011).

The selection of Interaction Design methods is affected by multiple factors: Firstly, the purpose of the evaluation, context of use, and type of data to be gathered matter (Rogers, Sharp, and Preece 2011). Secondly, practitioners must consider the reliability, thoroughness and validity of methods (Hartson, Andre, and Williges 2003). Finally, a number of case based issues contribute to the selection, such as cost efficiency and the target group.

All of the studied projects described the methods used in the study but not necessarily the rationale behind their selection. Notably three of the projects, including the Sentient Sketchbook, the Idea manager, and Choreographer's Notebook, used remote methods, including the collection of usage logs to determine the quantity of use or usage patterns. Shibata and Hori (2002) explained they needed longitudinal remote data collection because creativity is dependent on the context and environment of the users and thus impossible to study in a laboratory setting. Nearly all laboratory studies seem to have strived to simulate creative situations for the users, with the exception of STANDUP and Evolver.

Methods have also been applied to creative contexts in different ways. For example, the tasks used in the evaluation are very differentiated, some evaluations having more explorative tasks with only a general goal (e.g. Pleo and Parallel Pies), while others used more specific tasks with scripted roles for the participants (e.g. EDC).

Identifying Issues

Regardless of the chosen methods, all methods require representative participants, representative tasks and representative environments in which participants are observed (Lewis 2006). These dimensions define most of the practical issues related to any Interaction Design evaluation, and were not absent from the example projects either. For example, finding suitable users was difficult for the STANDUP project (Waller et al. 2009).

The creative context also proposes some additional issues to evaluation: Experiences from creativity support tool evaluation show that errors in the interfaces may sometimes provide additional opportunities for the users, and that spending significant times at a task may indicate immersion, not poor quality of interaction (Carroll 2011). Therefore, some metrics loaned from Interaction Design may not suit the evaluation in the creative setting (Carroll 2011).

The novelty and value of artefacts produced by creative systems become highly dependent on user and context in creativity contexts, as suggested by Bown (2014): For instance, Shibata and Hori (2002) studied a creativity support tool intended to catalyze idea generation. They had their users to evaluate the novelty and practicality of the ideas for themselves, instead of trying to assign objective values to the produced ideas.

Ethics

As with all human studies, ethical issues require specific care with Interaction Design evaluations involving users. Very few specific ethical issues were reported in the example studies, and in general they were unrelated to creativity: Waller et al. (2009) report issues related to child participants and Warr and O'Neill (2007) note the use of consent forms and stress to users that they are evaluating the software, not the users.

Analysis and Presentation

The chosen methods define the type of data collected to a great extent but researchers still have to choose how to analyze and present the data, as well as account for its validity, generalizability and scope (Rogers, Sharp, and Preece 2011). Many of the sample cases focus on the creative process and key interactions related to it in their analysis: Yannakakis et al. (2014) analyzed use patterns from log files and identified important process milestones from them with the help of the user provided qualitative data. Singh et al. (2011) also analyzed logs noting key changes in the creative processes by presenting rationale for the use. Warr and O'Neill (2007) recognized different sub-activities and key interactions in the idea generation process of their users based on video logs. Ryokai et al. (2009) illustrated the process through a detailed example and Carlson et al. (2011) focused on process related user quotes. As a semi-process oriented reporting approach Waller et al. (2009) focused on analyzing interaction paths and Terry et al. (2004) analyzed how well the interaction model enhanced the workflow of their users.

Feedback plays a great part in most of the evaluation projects; Waller et al. (2009), Johnston et al. (2005), Shibata and Hori (2002), Warr and O'Neill (2007), and Terry et al. (2004) report new ideas for improvement. Most projects also used user quotes to illustrate key findings or feedback; only Yannakakis et al. (2014), Warr and O'Neill (2007), and Terry et al. (2004) do not use user quotes at all.

Evaluation of the Poetry Machine

The Poetry Machine (Kantosalo et al. 2014) aims to solve the problem of the empty paper for its users, school children studying poetry or just exercising writing. The user selects a theme (in the tested version one out of 8 options), and the Poetry Machine provides a draft poem consisting of poetry fragments. The editing interface simulates a set of fridge magnets. The user can edit the draft by dragging words and rows around, removing them, or entering new ones. The user can also ask for further assistance from the computer, by using a feature called the "robot". By dragging words or rows on the robot, the robot provides the user with similar fragments or rhyming words.

The Poetry Machine has been developed at the University of Helsinki, based on the poetry generation methods developed earlier in the group (Toivanen et al. 2012). However, the version evaluated in this paper does not utilize the full functionality of these methods. Instead we decided to use simple fragment based approaches to provide pieces of poetry and rhyme candidates that can be expanded to full poems by users of the system. The Finnish poetry fragments and rhyme dictionaries are automatically extracted from a corpus containing children's literature from Project Gutenberg. This simplistic setting makes it easier to assess the effectiveness of the current interface of the system and also provides a basic setting for further iterative testing.

Planning the Evaluation

In the next paragraphs we describe the evaluation process of the Poetry Machine through the DECIDE framework.

Determining Evaluation Goals We selected three goals for the evaluation of the Poetry Machine: (1) discovery of usability problems, (2) evaluation of its usefulness, and (3) evaluation of its enjoyability. The first goal is a typical Interaction Design evaluation goal, yielding concrete remarks on how to improve the interface. In this case, eliminating usability problems is a vital step before conducting additional, comparative testing on the contents of the co-creation. The second goal, usefulness, is defined here as the system's ability to make creative writing easier for children. Finally the last goal, enjoyability, is related to the ISO-9241-11 (ISO/IEC 2010) satisfaction parameter, but combined with fun, which with child users correlates with usability (Sim, MacFarlane, and Read 2006).

Exploring the Questions In the question exploration phase, each goal was elaborated with a set of sub-questions, which could be more easily approached with specific Interaction Design evaluation methods. Our primary study questions were:

1. Usability
   (a) Are children able to use the program?
   (b) Is the interface graphically pleasing to children?
2. Usefulness
   (a) What features of the program are the most useful for children?
   (b) Does the program make creative writing easier for children?
3. Enjoyability
   (a) Do children exhibit negative signs, such as signs of boredom or frustration, when using the program?
   (b) Do children exhibit positive signs, such as smiling, or willingness to continue the activity for a longer period of time?
   (c) What activities do children name when asked about the most fun/boring features in the program?

Most of the questions can be further divided into sub-sub-questions, such as "Do children use all of the features or only few?".

We intentionally excluded questions, such as "Does the tool promote learning or creativity?". These questions were considered outside the scope of the first evaluation, but more experiments are planned for evaluating the pedagogical potential of the tool, and alternatives for promoting creativity.

Choosing Methods In order to gather a wide range of feedback, we decided to use a mixed-methods approach with two methods: Peer Tutoring and a small group session we call Group Testing. We chose the paired Peer Tutoring composition proposed by Edwards and Benedyk (2007) in which two users work as a pair – the first participant first learns the use of the tool and then teaches it to his or her partner. In the Group Testing we simulated a small group teaching scenario with one teacher teaching a group of five pupils on how to write a poem with the Poetry Machine. By using the methods in a school environment, we attempted to imitate some culturally and contextually aware conditions.

Peer Tutoring was selected as it has been previously used with young children in usability tests organized at school. It offers a natural context for using the tool with a friend, diminishing biases resulting from an unbalanced adult-child relationship between the users and the researchers administering the test (Höysniemi, Hämäläinen, and Turkki 2003). It is also good for eliciting comments from children (Edwards and Benedyk 2007), as well as for fostering creativity, experimentation and problem solving-skills within the test situation (Höysniemi, Hämäläinen, and Turkki 2003). Group Testing allowed us to observe the use of the system in a more authentic, teacher driven learning situation.

Observation of behavioral signs is considered more trustworthy than self reports in the case of children (Hanna, Risden, and Alexander 1997), and it is used in both methods to provide both quantitative and qualitative data. To collect more qualitative data, both methods were coupled with an appropriate background questionnaire and a post task debriefing. With Peer Tutoring we used a paired interview. For the Group Testing we developed a group-based, game-like, feedback gathering method called Feedback Game (Kantosalo and Riihiaho 2014).

Each of our six Peer Tutoring sessions started with *tutor introduction*: The researchers presented themselves to the tutor pupil, and the facilitating researcher helped him or her to fill a background questionnaire. During the next step, *tutor training*, the tutor was encouraged to explore the tool and

write a poem with it. Next, during *tutee introduction*, the tutee was introduced to the test setting and filled the background questionnaire, while the tutor read a book. This was followed by the actual *peer tutoring* phase, during which the tutor guided the tutee in writing a poem with the tool. Finally the tutor and the tutee were interviewed in what we call the *pair interview* phase.

Both Group Testing sessions started with an *introduction* phase, during which the participating children filled in the same background questionnaire as the Peer Tutoring participants. This was followed by *instruction by the teacher*, during which the teacher shortly composed a poem at the front of the classroom explaining the use of the tool. We then moved on to the *poem writing* phase, during which each child composed a poem, the teacher instructing them when necessary. Feedback from the children was then gathered in the *the Feedback Game* phase. In the game children answered questions like "Was it fun to use the poetry tool?" on a five step Likert scale turned into a gameboard. Each question was followed by a round of arguments. Finally a separate *teacher interview* was conducted to learn how the teacher perceived the effects of the tool on his class.

Identifying Issues As a sensitive user group children require specific care in selecting and applying test methods. Both, the Peer Tutoring test and the Group Testing were conducted on site, in a small classroom at a local Finnish school. To gather enough material to make for possible test session failures, we decided to work with a fairly large group of children. We recruited a class of 9-10-year-old pupils. Their teacher selected 22 participants (12 for Peer Tutoring, 10 for Group Testing) according to criteria provided by us. The sample is large, but narrow, which is somewhat typical for Interaction Design evaluation with children (see e.g. the sample sizes in (Sim, MacFarlane, and Read 2006) or (Höysniemi, Hämäläinen, and Turkki 2003)). Further testing with more varied users is planned.

To ensure unintrusive data collection we videotaped each session, and the researcher acting as the main facilitator in charge of interviewing and helping was accompanied by one or two additional observers, who were present at all times. Additionally we performed automatic data collection of the artefacts produced by the children, including recording which words in each poem were computer generated.

To promote creative thinking, we decided to use a very generic test task — the general goal of "writing a poem". In Peer Tutoring, this proved very difficult for some of the tutors, who were unfamiliar with poetry and required thus more guidance, such as suggesting a topic in one case. The tutees seemed to respond to the task more positively, possibly due to peer presence. We were also worried the tutors might try to push tutees to a specific creative direction during testing and discouraged this by allowing only the tutee direct access to the mouse and keyboard during the peer tutoring. We were happy to see the tutors seldom did anything to affect the creative content of their tutee's poem. The same open task worked well with the Group Testing participants.

Ethics As the participants of the study were all underaged, we requested a permission from the guardians of each pupil with a letter sent to them through the school. Additionally, we emphasized the volunteer nature of the study at the beginning of each session, explained the secrecy of all raw material, and noted we were there to recruit the pupils' help, not to evaluate them. During two of the Peer Tutoring sessions we held longer pauses to allow the tutor pupils to take a recess or have lunch before continuing with their tutee.

Analysis and Presentation All sessions were analyzed from videotaped material. All peer tutoring session videos were analyzed by two researchers; the facilitator and one observer. Each Group Testing session video was analyzed by the facilitator. Additionally field notes were used to note important factors during testing. The facilitator counted instances of use for each feature from the Peer Tutoring videos, as well as positive and negative gestures. Both facilitator and observer additionally observed the tape for interesting comments, actions and usability problems. The problem listings obtained were combined and duplicates were merged into single problems. Each problem was rated by frequency and assigned a severity rating. It was not possible to conduct an equally robust analysis of the Group Testing sessions, because of limitations in taping each participant individually. More general observations were made instead. The pair interviews and Feedback Game sessions were transcribed and the transcripts analyzed for common elements and improvement ideas.

Evaluation Results

The analysis revealed a number of interesting issues related to the evaluation goals and user ideas for improving the tool. Additionally we were able to find some interesting elements related to the use patterns and creative processes of the users.

Usability We found 82 unique usability problems through the Peer Tutoring tests. The problems ranged from practical interface problems, such as how to move words, and aesthetic problems, such as the appearance of buttons on screen, to more conceptual problems including for example misunderstanding of what publishing a poem means. A solution for each problem was suggested based on the problem's manifestation during testing and improvements are being carried out to allow further testing.

Enjoyability The enjoyability of the tool was evaluated based on gestures recorded from the Peer Tutoring videos and user comments. All of the six girls, who participated in the Peer Tutoring tests, seemed to show more negative gestures than positive when composing a poem. Four of the six boys however showed more positive signs. This could be taken as an indication of a generally negative reception for the prototype, however there is some ambiguity in interpreting gestures of children: Hanna et al. (1997) consider frowning a negative sign, but during testing this seemed rather to be a sign of concentration, which according to Read et al. (2002) should be considered as a positive sign. Also, as Carroll points out (2011), these signs may have to be interpreted differently due to the creative context. If we interpret these possible signs of concentration as positive, only one pupil displayed more negative gestures during testing. Most of the

negative comments heard during testing had to do with the ambiguity of the task: some children were unsure of what poems are and how to write one. Other negative comments heard during the Peer Tutoring indicated usability problems, and in one case disapproval of the concept itself. Less negative comments were heard during the Group Testing, where children received more clear instructions from their teacher.

The interview and Feedback Game results support a more positive response to the tool: All Peer Tutoring participants gave great scores for the prototype (4 or 5 stars out of 5), 5 out of 12 stating reasons related to the perceived fun of the tool. Additionally two pupils would recommend the tool to their peers based on fun. All Feedback Game participants agreed the tool was fun, and four of them specifically indicated they were willing to participate in a similar test because writing poems during the test was so fun. Enjoyability is also supported by anecdotal evidence provided by the teacher after the testing, during a later visit to the school, and the general reception children gave to the tool. This includes one child mentioning after a test that she had actually stayed after school as she was so enthusiastic to try the tool out.

Usefulness The tool was found useful by both the pupils and their teacher: The pupils clearly responded positively to writing poems with the tool. 12 out of 22 pupils indicated that poem writing with it was fun. Six pupils out of 22 also considered that writing poems with the tool was easier than writing otherwise. One pupil specifically mentioned that existing words given by the computer helped his writing process. The teacher highlighted motivation issues: He considered that the pupils were faster to get to work and more engaged with the program than in a typical lesson. He specifically mentioned that one of the pupils, who usually had difficulties with coming up with ideas for creative writing worked very autonomously throughout the session. The teacher also reported later that one of his pupils had been inspired by the tool to start poem writing as a hobby.

All pupils were able to write a poem during testing, however two of them seemed to reproduce one written before the test session. Also, some of the tutors required some ideation help for writing their poem and the facilitator suggested a theme for them, helping the process along with some open questions.

No formal evaluation of the educational value of the tool was made and children were not asked to specifically evaluate the learning potential of the tool, but many of the children considered the tool useful for learning: Seven pupils wanted to recommend the tool to others as they saw it as useful for learning. Three pupils considered autonomously that they had themselves learned to write poems with the tool. The teacher was also able to see the tool as a useful part for future lessons.

Use Patterns and Creative Process To gain a better understanding of the use of the tool, we recorded how many times each feature was used by the children during testing. While some of the users were writing with no apparent pattern, the data showed two clear strategies utilized by some of the pupils. The first strategy was to use one of the row-boxes, originally intended to note the row structure in the final poem, as a storage-unit. A pupil using this storage-strategy would shift words within the interface from the operational area to the storage-unit and back according to his or her poem idea. The final poem would consist in a large part of words suggested by the computer. Four participants in the Peer Tutoring test were seen using this strategy. The second strategy, robot-induced-ideation, was seen specifically in one of the pupils. He would primarily engage with the robot, looking always first through its suggestions and only then added a word either invented by the robot or himself.

By looking at the usage data recorded during the use, Peer Tutoring participants wrote shorter poems than the Group Testing participants. The average length of Peer Tutoring participants' poems was 11.6 words (median 11, minimum 6 and maximum 23), while the Group Testing participants wrote 25.4 word poems on average (median 19, minimum 12 and maximum 55). On average, 28% of the final words in the poems written by Peer Tutoring participants were provided by the computer (either in the initial draft, or suggested by the robot tool), while 34% of the words used by Group Testing participants originated from the computer. In both test setups two pupils decided not to use any of the suggestions provided by the computer, while in the Group Testing one participant relied entirely on words suggested by the computer, acting as a sort of an editor. However, the logs do not record all of the effects of the tool to the writing of the children – for example one child said during a Peer Tutoring session that "something came to my mind from this" and pointed to one of the robot's suggestions.

We did not attempt to evaluate the quality of the poems and the possible effect of Poetry Machine on them. A larger sample would be needed, as well as a comparative set of poems, either from the same age group or from earlier poems written by these pupils.

User Ideas The user ideas collected during testing are summarized in table 1. Peer Tutoring and Group Testing produced different kinds of ideas. On average one Peer Tutoring session produced one idea, whereas each Group Testing session managed to produce two. The ideas gathered during Group Testing are also more immediately related to the conceptual level of the system, while the Peer Tutoring ideas also address more specific interaction ideas. We discuss the main ideas below.

1	Inputting multiple words together should be easy
2	Users should be able to remove all words easily
3	Proposed words should be more familiar
4	Proposed words should be more tightly linked to words pointed out by the user
5	Proposed words could be displayed under the word to be replaced
6	A quick way to add punctuation is needed
7	Drafts should have more familiar words
8	Proposed words should be more related to the topic
9	Proposed words should have better rhymes
10	Drafts should have more rhymes

Table 1: Ideas collected from users during testing

Using the Results in Developing the Poetry Machine

The usability evaluation results are already used to enhance the interface in order to support test situations in which we focus more on the content of the interactions instead of their fluidity. The initial results will guide our research into the pedagogical potential of the tool, and we will further focus in the development of the tool as a motivating agent.

The use patterns collected show potential principles on the base of which further interaction in the tool can be build to support human-computer co-creativity. For example the storage-strategy should be investigated further as an interaction paradigm in the system. The relationship between the robot-induced-ideation and the quantity of computer provided words in the system should be investigated further in the tests, and means for promoting it could include a more active computational participant.

The feedback provides many possibilities for further development of the computational creativity methods used in the system. Especially the ideas give concrete suggestions as to how the system should be developed further.

(1) Instead of just providing simple fragments without any cohesion between them, methods for adding more coherence between the proposed fragments should be investigated. Here the computer could propose fragments that are well suited to the fragments already proposed and also written by the user. Methods of textual coherence based on vector space models of words and linguistic fragments (Mikolov et al. 2013) or corpus word statistics could be used here to enhance the results.

(2) The quality of the rhymes have room for improvement. Methods for improving the quality of rhymes are many, including metrics based on word length etc. Also many different kinds of rhymes like syllabic rhymes, half rhymes, assonances, consonances, and alliteration could be used to add more variation.

(3) Words suggested by the system could be more familiar to the users. However, the users were not unanimously supporting the use of only familiar words. During group testing, one pupil noted that "there were these words you use more seldomly, so there were a couple I could select for my poem". Therefore, tying the words better to the context, proposing synonyms and antonyms for the words pointed out by the user, and using a mix of more and less typical words a good balance between vocabulary enhancing and supporting words could be attained.

In the future, the system could also be used for teaching metrical systems prevalent in traditional poetry. The computer could, for instance, propose that the user could write a sonnet and then track the number of syllables on each line of the poem. If the number of syllables on some line did not fit the metrical structure of a sonnet the computer could propose changing, for instance, one word on the line to satisfy the metrical constraints.

Conclusions

We have shown how to conduct an interaction design method based evaluation on a human-computer co-creativity tool called Poetry Machine. The evaluation conducted in this case study has similarities to other evaluation cases of human-computer co-creative tools and creativity support tools. Especially interesting is the varied set of evaluation goals that can be supported through Interaction Design methodologies. In creative contexts however, the selection of methodology seems to be especially important: Mixed methods should be used to gain a varied set of data. Also specific care has to be taken to create a test situation that allows the flow of creativity by either using remote study methods, methods that have been found to suit creative contexts, or setting up the evaluation in a creative environment. Tuning methods for creative contexts also requires selecting suitable tasks for the users to do within the test situation.

A very interesting aspect to Interaction Design evaluation planning and practice within the creative context are the issues faced during testing. It seems that some traditionally used interaction design evaluation measures, such as time, or facial gestures are not useful within a creative context, as some negative signs, such as frowning, may actually indicate positive aspects, such as concentration or immersion instead. Most of the issues related to human-computer co-creativity testing with interaction design evaluation methods still seem to be concerned with typical interaction design evaluation problems, such as selecting suitable users.

The analyzed sample cases revealed that typically the analysis of human-computer co-creativity evaluation results is similar to that of Interaction Design evaluation. For example, quotes are frequently used to illustrate key issues. Interestingly many projects have also focused on how the creative process of the user is supported by the interface. A large part of the cases also provided feedback and improvement ideas.

We have illustrated here how such formative evaluation results can be applied to practical computational creativity development work by providing a list of gathered user ideas and presenting concrete ideas on how to use them for further development. However, a simple listing of the ideas is not enough – to defend design decisions and to tune solutions to actual user needs, we need to look at the qualitative data as a whole.

Based on the projects studied for this paper, it seems interaction design evaluation methods have already taken a place within human-computer co-creativity evaluation and the philosophical foundations of this work are also being laid in the computational creativity community. Through our case study, we have demonstrated in a formalized manner, how to plan and conduct Interaction Design method based evaluation for a human-computer co-creativity tool and how the results can be applied in practice. With this we have shown how interaction design evaluation practices offer an interesting, complementary evaluation approach to human-computer co-creation tools, providing results that can be put to practical development use.

Acknowledgments

This work has been supported by the Academy of Finland (decision 276897, CLiC) and by the European Commission (FET grant 611733, ConCreTe). We wish to thank the pupils

and teachers who participated in this research, and K. Tiuraniemi and M. Hynninen for participating in the data collection.

References

Bown, O. 2014. Empirically grounding the evaluation of creative systems: Incorporating interaction design. In *Proceedings of the Fifth International Conference on Computational Creativity*, 112–119.

Carlson, K.; Schiphorst, T.; and Pasquier, P. 2011. Scuddle: Generating movement catalysts for computer-aided choreography. In *Proceedings of the Second International Conference on Computational Creativity*, 123–128.

Carroll, E. A. 2011. Convergence of self-report and physiological responses for evaluating creativity support tools. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*, 455–456. ACM.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: the final frontier? In *ECAI 2012 : 20th European Conference on Artificial Intelligence*, 21–26.

DiPaola, S.; McCaig, G.; Carlson, K.; Salevati, S.; and Sorenson, N. 2013. Adaptation of an autonomous creative evolutionary system for real-world design application based on creative cognition. In *Proceedings of the Fourth International Conference on Computational Creativity*, 40–47.

Edwards, H., and Benedyk, R. 2007. A comparison of usability evaluation methods for child participants in a school setting. In *Proceedings of the 6th International Conference on Interaction Design and Children*, 9–16. ACM.

Hanna, L.; Risden, K.; and Alexander, K. 1997. Guidelines for usability testing with children. *interactions* 4(5):9–14.

Hartson, H. R.; Andre, T. S.; and Williges, R. C. 2003. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction* 15(1):373–410.

Höysniemi, J.; Hämäläinen, P.; and Turkki, L. 2003. Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers* 15(2):203–225.

ISO/IEC. 2010. Iso 9241-210 ergonomics of human-system interaction – part 210: Human-centered design for interactive systems.

Johnston, A.; Amitani, S.; and Edmonds, E. 2005. Amplifying reflective thinking in musical performance. In *Proceedings of the 5th Conference on Creativity & Cognition*, 166–175. ACM.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Jordanous, A. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*, 129–136.

Kantosalo, A., and Riihiaho, S. 2014. Let's play the feedback game. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 943–946. ACM.

Kantosalo, A.; Toivanen, J. M.; Xiao, P.; and Toivonen, H. 2014. From isolation to involvement: Adapting machine creativity software to support human-computer co-creation. In *Proceedings of the Fifth International Conference on Computational Creativity*, 1–8.

Lewis, J. R. 2006. Sample sizes for usability tests: Mostly math, not magic. *Interactions* 13(6):29–33.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Read, J.; MacFarlane, S.; and Casey, C. 2002. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children*, volume 2, 1–23. Shaker Publishing Eindhoven.

Rogers, Y.; Sharp, H.; and Preece, J. 2011. *Interaction Design: Beyond Human Computer Interaction.* Wiley, 3rd edition.

Ryokai, K.; Lee, M. J.; and Breitbart, J. M. 2009. Children's storytelling and programming with robotic characters. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition*, 19–28. ACM.

Shibata, H., and Hori, K. 2002. A system to support long-term creative thinking in daily life and its evaluation. In *Proceedings of the 4th Conference on Creativity & Cognition*, 142–149. ACM.

Sim, G.; MacFarlane, S.; and Read, J. 2006. All work and no play: Measuring fun, usability, and learning in software for children. *Computers & Education* 46(3):235–248.

Singh, V.; Latulipe, C.; Carroll, E.; and Lottridge, D. 2011. The choreographer's notebook: A video annotation system for dancers and choreographers. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*, 197–206. ACM.

Terry, M.; Mynatt, E. D.; Nakakoji, K.; and Yamamoto, Y. 2004. Variation in element and action: Supporting simultaneous development of alternative solutions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 711–718. ACM.

Toivanen, J. M.; Toivonen, H.; Valitutti, A.; and Gross, O. 2012. Corpus-based generation of content and form in poetry. In *International Conference on Computational Creativity*, 175–179.

Waller, A.; Black, R.; O'Mara, D. A.; Pain, H.; Ritchie, G.; and Manurung, R. 2009. Evaluating the standup pun generating software with children with cerebral palsy. *ACM Transactions on Accessible Computing* 1(3):16:1–16:27.

Warr, A., and O'Neill, E. 2007. Tool support for creativity using externalizations. In *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition*, 127–136. ACM.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.