Experience Evaluations for Human-Computer Co-Creative Processes - a Proposal

Anna Kantosalo [1]

Abstract.

Human-computer co-creative software requires an evaluation perspective that better incorporates the experiences of the users. In this paper we propose an evaluation perspective based on user experience evaluation to evaluate the human-computer co-creative process. We apply this perspective into forming an evaluation plan for a human-computer co-creative poetry writing system called the Poetry Machine. We feel that the inherently contextualized and subjective user experience perspective complements the commonly assumed positions of objectively evaluating the creative outcomes and the creative algorithmic processes in the domain of computational creativity applications. We first discuss the need for such an evaluation perspective and continue to present relevant evaluation metrics. We then outline an experiment in which we propose to use the said metrics to investigate differences in a human-computer co-creation process, a human-human co-creation process, and a human-human-computer co-creation process with the poetry writing system.

1 INTRODUCTION

As computational creativity generation methods advance, we are more and more interested in how to put them to good use. Human-computer co-creativity studies how we can use computational creativity methods to promote human creativity in a co-operative creative process with a computer and *vice versa*. In this process the human becomes an integral part of the creative process itself instead of being just part of the audience. This requires adopting a new evaluation stance, which goes beyond the traditional computational creativity evaluation focuses of creative output and internal workings of the system. Evaluation of human-computer co-creativity systems must be carried out in real contexts with real users, so that we get authentic data for developing our systems further.

In this paper we propose evaluation from a human perspective focusing on the user experience of users participating in the co-creative process. User Experience is closely related to the domain of Interaction Design, which has been proposed as a useful evaluation paradigm by Bown [1], and successfully utilized in some user evaluations of computational creativity applications (see e.g. [11, 2]). The term user experience refers to a number of hedonistic aspects of interaction design summed up as the quality-in-use of interactive products or "the result of enjoyable interactions and/or anticipated interactions with a product" by Lachner et al. [12]. By including an experience oriented perspective, we can better promote collection of information that also involves the complex emotive expectations and reactions to computationally creative software that affects the creative use of such systems.

In computational creativity evaluation, a focus on user experience has been scarce so far, but Yee-King and d'Inverno [32] suggest that measuring the experiences of a human interacting with a system is more important than measuring the systems creativity. Similarly Bown [2] considers that instead of measuring an interactive systems creativity, we should focus on describing the complex network of interactions that surrounds it. We agree with these views and argue that investigating the human-computer co-creative process from the human's subjective, experiential perspective can reveal new insights useful for both the design and evaluation of such systems.

Our evaluation target is a poetry writing agent designed to collaborate with children in schools. Creative writing and poetry are both part of the national curriculum, which also emphasizes learning by doing. The Poetry Machine application is intended to help pupils create their own poetry in and out of the classroom. We are interested in what constitutes a good creative experience when working with the application. We examine this question in a comparative setting, with children writing both with a friend and/or the Poetry Machine system. We feel that the user experience oriented evaluation view will give us a more nuanced picture of the aspects influencing creative work with a creative system.

In this paper we outline our approach to investigating user experience in this comparative setting. We start by looking at what good user experience in human-computer co-creation might entail. We mix together three perspectives: user experience evaluation, computational creativity evaluation and creativity support system evaluation. We briefly discuss evaluation metrics presented in these three domains and then proceed to examining our study context. After presenting our context, we present our list of metrics, and discuss how they are based on the three perspectives and what kind of study questions we have derived from them. Finally we present our study method.

2 STUDY INSPIRATIONS

Maher [16] has suggested that similar evaluation metrics should be used for evaluating the creativity of "humans, computers, and collectively intelligent systems" regardless of their domain. Inspired by Maher's consideration, we assume the position that the human-computer co-creative user experience and human-human co-creative user experience can be studied with the same metrics. Yet we think some metrics may be more important for one case than the other. Therefore we should gather our metrics from a variety of domains: general user experience evaluation, computational creativity evaluation, and creativity support system evaluation. Finally we included

[1] Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Finland e-mail: anna.kantosalo@helsinki.fi

three case studies from the computational creativity literature to illustrate possible user experience evaluation in action.

2.1 Study Inspirations from User Experience Evaluation

User experience evaluation is a part of the long standing usability evaluation tradition [14] and its methodology can be viewed as an augmentation of known usability evaluation methodologies [31]. MacDonald and Atwood [14] conclude that three basic categories of methods, suited both for pragmatic, usability oriented testing, as well as hedonic, user experience oriented testing, exist: User Testing Methods, Inspection Methods, and Field Methods. The main difference to usability evaluation is that the nature of user experience evaluations is inherently subjective, since we are interested in "how the user feels about the system", including user motivations and expectations [31].

Vermeeren et al. [31] consider that there is still a lack of systematic research into evaluating and measuring User Experience. This may be the reason, why general user-experience evaluation literature and metrics are still scarce. Here Lachner et al. [12] stress the importance of defining user-experience for a specific context as the basis of evaluation: "To create a certain UX [user experience], a systematic approach and an associated description of user experience are needed to consider and measure the intended experience". To this end they report nine user experience dimensions divided into three categories based on their extensive literature review and expert interviews. The categories, named "Look", "Feel", and "Usability" include the following categories "Appealing visual design", "Communicated information structure", "visual branding", "mastery", "outcome satisfaction", "emotional attachment", "task effectiveness", "task efficiency" and "stability and performance", which have been translated into 27 Likert scale questions.

2.2 Study Inspirations from Computational Creativity

One of the first attempts for systematical, empirically grounded computational creativity evaluation criteria is Ritchie's framework, presented in [25] and elaborated in [22] and [26]. In his framework (as presented in [26]), Ritchie proposes we investigate the creativity of a program by focusing on "what is known", meaning the attributes of the creative artefact it produces instead of the internal workings of the program. In the framework [25], creative artefacts are assessed by two factors: quality and novelty. (Ritchie later [26] substituted novelty for typicality as a more suitable characteristic for computational systems.)

Pease and Colton [21] assumed a position of evaluating the impact of the creative artifact on a supposed ideal audience. This audience gives two ratings for each artifact: a well-being rating and a cognitive-effort rating. With the well-being rating users evaluate the subjective like or dislike of an artifact, and with the cognitive-effort rating they evaluate if they are willing to devote time to understanding the artifact.

Colton [4] shifted the viewpoint of evaluation from artifacts to the internal processes of the system in his Creative Tripod-framework, which requires that in order to be perceived as creative, we must be able to say that the behavior of the software is skillful, appreciative and imaginative.

The most recent efforts for setting evaluative criteria for computational creativity methods have focused on determining components of creativity from larger data sets. Such efforts include Jordanous' [10] corpus based approach on determining words related to creativity by their lexical co-occurrence as well as van der Velde et al.'s [30] efforts to determine components of creativity with human association tests. Jordanous identified in her research 14 components of creativity, including "Active Involvement and Persistence", "Dealing with Uncertainty", "Domain Competence", "General Intellect", "Generation of Results", "Independence and Freedom", "Intention and Emotional Involvement", "Originality", "Progression and Development", "Social Interaction and Communication", "Spontaneity/Subconscious Processing", "Thinking and Evaluation", "Value", "Variety, Divergence and Experimentation". Using Jordanous' work as a basis van der Velde et al. then again found 42 creativity related words through their human association studies. These 42 words were further distilled to five clusters based on human sorting: originality, emotional value, novelty/innovation, intelligence, and skill.

2.3 Study Inspirations from Creativity Support Systems

Any tool or combination of tools that can help "in the open-ended creation of new artifacts" can be considered as a creativity support system [3]. Naturally, the role and purpose of creativity support systems are varied: systems can be derived for managerial roles, communication enabling, enhancing creativity, or even to perform as creative partners [13], or more specifically for developing creative skills, creating faster, or enabling entirely new creative methods [19]. In this short literature review we have focused on general issues related to the evaluation of creativity support systems.

Evaluation problems faced by the creativity support system domain are relatively similar to the computational creativity domain: As Shneiderman [27] writes, the evaluation and design of creativity support tools is not simple, as both the requirements and the measures for success are vague. Cherry and Latulipe [3] attribute this vagueness to defining what creativity is in the first place. As an added dimension confusing user evaluations, Shneiderman notes that user behaviors are "unorthodox" when compared to traditional "productivity support" systems.

Cherry and Latulipe [3] have developed a Creativity Support Index psychometric survey tool to examine quantitatively the support of creative digital tools. Their index takes into account "six dimensions of creativity support: Exploration, Expressiveness, Immersion, Enjoyment, Results Worth Effort, and Collaboration." The Creativity Support Index is suggested as an additional evaluation metric to complement other evaluation approaches. The tool itself consists of rating scales and paired-factor comparison sections. The paired-factor comparison section is intended to ground the rating scales for the specific creative tasks users perform during the testing.

A working paper by Resnick et al. [24] on design principles for creativity support systems highlights 12 guidelines for designing creativity support systems that could be summed up as follows: Support exploration through self-revealing interaction, cater for a variety of skill levels in a variety of projects, support many paths and many styles, support collaboration, support open interchange, make it simple, choose 'black-boxes' carefully, design tools to be enjoyable by yourself as well as others, use a multimethod approach to balance user requirements, use an iterative design approach, design for designers, and evaluate your system. In our opinion, the first four guidelines are most important for creativity, when the the rest are somewhat general the list ending with suggestions typical for any interaction design project.

Compton and Mateas [5] define a more specific, yet domain independent type of creativity support system: Casual Creators are simple tools to help facilitate searching for interesting creative artefacts in specified creative domains. They argue that the use of casual creators is firstly motivated by fun and a sense of ownership over the found artefacts, and designers should ensure the quality of the user experience by making sure there are no broken artefacts in the allowed search space and that there is a large enough number of surprising artefacts in the search space, yet it should be low enough to promote the sense of ownership and creativity when discovering one by intentionally searching the search space. In addition the system must offer a meaningful creative experience in the first few minutes to receive an audience. Fun and a sense of ownership are already intersting experiental factors that can be measured. Compton and Mateas add three more user experience descriptors intended to amount to a flow state: playfulness, powerfulness and pleasure.

2.4 Study Inspirations from Case Studies

The casual creator definition of Compton and Mateas has been used by Davis et al. [6] in their case study investigating the use of a casual creator system called the ”Drawing Apprentice” designed to simulate a human collaborator. Davis et al. tested their system in a comparative setting where the computational methods within the system were temporarily replaced by a professional artist. The focus of their study was on ”evaluating the quality of the collaboration and how influential their collaboration was in defining artistic goals” in an open-ended drawing task. The qualitative study listed the most critical evaluation metrics users used to determine if an agent's contributions 'make sense' as follows: spatial awareness, visual similarity and perceptual logic. Spatial awareness entails exhibiting a sense of what has been previously drawn and not ”messing it up”. Visual similarity entails contributing the process with visually similar material to that provided by the user. Perceptual logic means that the agent must adapt its new input to the existing drawing, especially when drawing spatially near it.

Jacob and Magerko [9] also conducted a comparative evaluation of three different degrees of interactivity in a co-creative dancing agent. In their study, a user was asked to interact with a dancing co-creative agent in three conditions: an agent mirroring their moves, a pre-authored version of the agent, and a fully interactive version of the agent. After each condition the users were asked to evaluate their experience with a few known usability evaluation instruments, including the creativity support index. Unfortunately, the small sample size of their pilot study does not yet show clear indications of which of the instruments work best.

Bown [2] investigated musical improvisation with three musicians and a computationally creative musical system. He used a focus group in which the musicians took turns in playing with the algorithm, observing others play with it and discussing their experiences. The intention of the study was to examine how musicians approached and discussed performing with computational systems focusing on questions of effectiveness, autonomy, novelty and surprise. In his experiment, Bown found that participants who played with the system had a stronger engagement than observers and the musicians did not compare the system with human standards.

3 STUDY CONTEXT

Our case study evaluates an interactive poetry writing system called the Poetry Machine. The most important contextual factors in our study are the system itself, including its intended use as well as the related interactions, our intended user group, and the specific use cases we are evaluating. The poetry writing application is designed for use at the Finnish comprehensive school and is developed by a multidisciplinary team of computer scientists and pedagogical researchers. More thorough evaluations of the educational use of the tool are being conducted separately from this study as a part of the on-going development effort.

The main idea behind the Poetry Machine application is to help pupils overcome creative droughts: When starting a new poem, the pupil chooses a theme to witch the application then generates a short poem excerpt. The pupil can modify the excerpt by moving words, adding words, or removing them entirely. The pupil has also a few tools for asking more materials from the computer - a magic wand that can be dragged over a word for replacement suggestions, a rhyming tool that gives different types of rhymes for a word dropped on it, and a robot button, which will generate more lines to the poem being modified.

We intend to study the application with children, who are 9 - 12-years of age. Children as users are somewhat different from adults: Firstly their cognitive skills, including concentration [18] and problem solving are not yet fully developed [7]. Secondly, also their physical skills, including motor and perception skills are still developing [7]. Finally the relationship between children participating in the test and the adult researchers administering the tests affects overall test success [20] and may cause bias in testing [23]. Therefore testing with children requires special considerations so that it can be done in a comfortable, secure atmosphere in which children feel free to express their feelings about the application to be tested.

Children are usually divided to different groups based on their age as children of similar age tend to share similar capabilities and abilities with respect to user testing. A review of the cognitive, perceptional and motor skills of children in different age brackets can be found for example in Hourcade's article [7]. However, it is important to note that individual differences between children are large and the ability of specific individuals to participate in testing can be boosted by task specific factors, such as good domain information, social support and good instructions [7].

We are interested especially in the experiences of a child creating a poem together with the Poetry Machine in a co-creative process. However, we are also interested to see, how this situation differs from two different scenarios: Two children writing a poem together without computational aid, as well as two children collaborating with one computer.

4 OUR METRICS

Based on the literature, we selected seven metrics, which we feel are most suitable for our context whilst giving a good overall picture of the quality of use. The metrics are: Fun, Enjoyment, Expressiveness, Outcome Satisfaction, Ease of writing (task efficiency), Collaboration and Ownership. The metrics and relevant background literature is shown in table 1.

The first metric, Fun, is derived from the Fun-metric suggested by Compton and Mateas in their casual creator work [5]. It was selected as it is a very clear experiential factor and coincidentally a lot of research exists on using fun as an experience parameter in children's software: Besides a useful descriptor for user experience [28], it is considered important for keeping children's attention in educational software [15] and their willingness to use a specific product altogether [23, 29]. Added to the criticism towards productivity,

Table 1. Selected metrics and the most influential related concepts in literature

Metric	Literature
Fun	Fun [5]
Enjoyment	Enjoyment [3], Pleasure [5]
Expressiveness	Intention and emotional involvement [10], Expressiveness [3]
Outcome Satisfaction	Outcome satisfaction [12], Quality [26], Well-Being rating [21], Value [10], Results worth the effort [3]
Ease of writing (task efficiency)	Task efficiency [12]
Collaboration	Social interaction and communication [10], Collaboration* [3], Support collaboration* [24]
Ownership	Sense of Ownership [5]

speed and efficiency oriented views of evaluating software intended for children [7] and Shneiderman's [27] considerations on the unorthodox use of creativity support systems, we must conclude that Fun is indeed one of the most important experience factors in this research.

The next metric, Enjoyment, builds on the Fun metric. It also draws from the work of Compton and Mateas, this time their pleasure parameter [5]. This metric however, is more similar to what Cherry and Latulipe [3] call Enjoyment. In the work of Cherry and Latulipe, Enjoyment is a metric that clearly spans a larger time frame. As a measure of enjoyment, they ask their users, if they would like to use the software in the future on a regular basis. This requires the users to evaluate how the enjoyment, or fun, they have had with the software is enough to support future use.

The Expressiveness metric relies mostly on the similarly named metric from the work of Cherry and Latulipe [3]. Under Expressiveness, Cherry and Latulipe ask their users, if they were able to be creative and express themselves in the use of the software. These questions are related to Jordanous' metrics on Intention and emotional involvement, which deal with self expression and emotional fulfillment in the creative task.

The outcome satisfaction metric is the metric most directly related to the user experience evaluation literature presented here. But as it implicitly requires the users to cast an evaluation on the final outcome of the creative process, it is also clearly related to the Quality, Well-Being, and Value ratings presented in computational creativity literature. A similar metric is also present in the work of Cherry and Latulipe [3].

Ease of writing is a specification of the Task Efficiency dimension suggested by the study of Lachner et al. [12]. However, we want to avoid specific efficiency metrics, as efficiency may not be a primary goal for creative users. Here we focus on the subjective efficiency experienced by the users, which is likely to reflect on the overall user experience.

The Collaboration metric is evident in a large part of the creativity support literature mentioned here. However, Jordanous' views on Social interaction and communication [10] fit this descriptor best as it fits mutual influence, sharing and feedback with different agents. This is different from the Collaboration [3] and Support collaboration [24] dimensions in creativity support literature, which usually specifically mention another user with whom the user is collaborating. Therefore we have noted this difference with a star at the end of the metrics in table 1.

The Ownership measure is derived from the ideas of Compton and Mateas [5]. They note that it is important to support a sense of ownership, and tie the user experience tightly into the user feeling control over finding meaningful artifacts in the search space. We feel that it is an attribute which may be especially important in investigating co-creative scenarios.

The list of metrics is quite short, since children are only able to concentrate for a limited time. This limits the number of questions and the number of question themes we are able to address in a questionnaire or an interview. Instead of covering a wide range of topics, we decided that we would narrow the list and within that list focus on some specific aspects, namely Expressiveness, Ease of writing, Collaboration and Creative Contribution. We have derived ten Likert scale statements to be used as part of questionnaires from these metrics, focusing on the said aspects. The full list of statements can be seen in table 2.

As we see from the statements, many of them require that users do some sort of implicit quality evaluation of artifacts or other outcomes of the process. Therefore it could be argued that quality and novelty are implicit metrics throughout the study, although we do not mention them specifically as metrics. It will be interesting to see, if our users base their judgment of the items in these metrics, when they have a chance to freely elaborate on their thoughts.

Natuarlly, there were other metrics we would have liked to include in our study: The Exploration metric, evident in many works, including Jordanous [10], Cherry and Latulipe [3] and Resnick et al. [24] was a very strong candidate. However, it became difficult to invent Likert statements that would fit both writing with the system as well as writing with a friend. Similarly, the Immersion dimension, present again in both Jordanous' list as well as in the work of Cherry and Latulipe, and related to the flow state mentioned by Compton and Mateas [5], is also very important, but difficult to fit into our limited schedule and the limited cognitive effort of our users. Yet, we will keep these metrics in mind when analyzing our data, as users may explain their answers and reveal important data related to these metrics as well.

Notably, we excluded some works altogether - the work of van der Velde et al. [30] did not reveal any additional dimensions we would have been interested in evaluating specifically. Also the metrics used in the case studies proved too specific for our case.

5 METHODOLOGICAL APPROACH

We are interested in the experiences of pupils working with the Poetry Machine system to create new poetry. We are especially interested in what aspects of the computational partner promote a good co-creative experience and whether the pupils see the role of the program as an active collaborator or just a tool. The Collaboration and Ownership metrics selected for our study stress this point of view. We feel that to investigate user experience thoroughly, we will need to look at co-creative poetry writing in a broader setting, involving both human-computer, human-human, and human-human-computer partnerships. Therefore we have planned a comparative evaluation with the following three conditions:

Table 2. Likert statements derived from the metrics

Metric	Likert statement
Fun	Writing the poem was fun
Enjoyment	I would like to write poems in this way in the future
Expressiveness	I was able to be creative
	I was able to express my own thoughts well
Outcome satisfaction	I was happy with the poem I wrote
Ease of writing	It was easy to start writing the poem
	It was easy to write the poem
Collaboration	I got good ideas from the other writers
	Other writers were helpful for my writing
Ownership	I feel the final poem is mine

- A: 1 Pupil and the Poetry Machine
- B: 2 Pupils
- C: 2 Pupils and the Poetry Machine

Our experiment is a same-participant design, meaning, we will form pairs of pupils and each pair will participate in all three conditions. There is an intrinsic value in paired test setups used with children: children tend to be more relaxed when communicating with adults as a peer is present in the test space. This can be achieved either by conducting the testing with pairs of pupils (see e.g. [8]) or just by conducting several tests simultaneously (for example [29], and [28]). The order of the conditions will be randomized so that we will have altogether six different treatments in which the order of the conditions varies. We will perform two tests for each treatment and run 12 tests in total. The participants will be recruited at schools and we will conduct the tests at the school in order to offer children a natural environment.

All test conditions will have a variant of the same task: "Write a poem to congratulate a friend who likes cats/dogs/mice." Having the same task will ease comparisons between the conditions. In the test condition A each pair is split up to write a poem with the Poetry Machine according to the task description. In the task condition B each pair will be given two permanent pens of different color and a piece of paper to write their poem on. We chose pen and paper over possible word processing programs, as in our experience children may not share an equivalent knowledge of all word processors and they may get carried away by layout details and other visual aspects of the program. The different color permanent pens accompanied by a video of the test situation will produce a log comparable to the log produced by the Poetry Machine application. In the final test condition C each pair will work together on the same Poetry Machine application instance writing a poem together.

We will videotape all test conditions and observe the tests in situ for authentic communication related to the creative task at hand. We will promote thinking aloud where possible to record the users impressions as they occur. In addition to observations, we will administer a short questionnaire based on the Likert scale statements provided, after each condition. After all three conditions, each pupil will also answer similar questions which ask them to compare the three conditions. After this each pair will be interviewed on their experiences in a discussion covering the same themes. This will give the children a chance to elaborate on their views on the questions provided. We will use befitting discussion aids involving tactile rating scales and physically interactive objects to enable children to answer questions related to these abstract themes in a meaningful way.

The main differences between our study and the three case studies are that instead of investigating open-ended exploration, as Davis et al. [6] did in their experiment, we provide our users with a clear task, which stays relatively similar from one condition to another. The purpose of the task is to provide a more natural context for the users to work in. Unlike the other studies, we also deliberately compare collaboration with a human with collaboration with a computer. By investigating both dimensions and their combination, our users can provide us with more detailed descriptions of their relative experiences.

Our analysis will focus on key differences and similarities in the creative experiences of the users in conditions A, B, and C. We will be interested to see, what kind of language users use to describe their collaborators in the different conditions. We are especially interested in the role of the computer as a creator in the experiment: Does it take the same role in the two different conditions it participates in? Are the collaborative responsibilities of the human collaborator similar to the computational one, and how? By analyzing the results from a role based view, we hope to add to the understanding of different roles taken by creative software in the creative process, as previously described by Lubart [13], Maher [17] and Nakakoji [19].

6 DISCUSSION AND CONCLUSIONS

We have argued for a more contextual evaluation of human-computer co-creative systems focusing on the user experiences of the human participating in the co-creative process. We have presented a set of factors from previous approaches to evaluating user experience, creativity of computational creativity systems, and creativity support systems. We have shown how to select suitable metrics based on the specific contextual needs of our research. Finally we have suggested an experimental approach into studying user experiences of children creating together with a co-creative computational partner.

Much work is still needed to establish the user experience oriented evaluation of the co-creative process. The on-going work here can be seen as a starting point, upon which to build a much larger understanding of the specific evaluative needs of different contexts and creativity domains with regard to experience evaluation. We expect that the experiment we present here, once carried out, can demonstrate some preliminary benefits of the approach, as well as the usefulness of the suggested metrics for evaluating the co-creative experience in the creative writing domain.

Yet further work will need to be carried out to establish the relevance of experience evaluations in the field of computational creativity at large: We need to see a variety of different domains, experiment with different criteria, and formalize investigative tools, such

as questionnaires. But most of all, we need to see where and how the experience evaluation results can be used to design more fluent human-computer co-creative processes.

ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland under grant 276897 (CLiC).

REFERENCES

[1] Oliver Bown, 'Empirically grounding the evaluation of creative systems: Incorporating interaction design', in *The Proceedings of the Fifth International Conference on Computational Creativity*, pp. 112–119, (June 10-13 2014).

[2] Oliver Bown, 'Player responses to a live algorithm: Conceptualising computational creativity without recourse to human comparisons?', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 126–133, (June 29 - July 2 2015).

[3] Erin Cherry and Celine Latulipe, 'Quantifying the creativity support of digital tools through the creativity support index', *ACM Trans.Comput.-Hum.Interact.*, **21**(4), 21:1–21:25, (jun 2014).

[4] Simon Colton, 'Creativity versus the perception of creativity in computational systems.', in *AAAI spring symposium: creative intelligent systems*, pp. 14–20, (2008).

[5] Kate Compton and Michael Mateas, 'Casual creators', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 228–235, (June 29 - July 2 2015).

[6] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko, 'Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent', in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, pp. 196–207, New York, NY, USA, (2016). ACM.

[7] Juan P. Hourcade, 'Interaction design and children', *Found.Trends Hum.-Comput.Interact.*, **1**(4), 277–392, (apr 2008).

[8] Johanna Höysniemi, Perttu Hämälinen, and Laura Turkki, 'Using peer tutoring in evaluating the usability of a physically interactive computer game with children', *Interacting with Computers*, **15**(2), 203–225, (4 2003).

[9] Mikhail Jacob and Brian Magerko, 'Interaction-based authoring for scalable co-creative agents', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 236–243, (June 29 - July 2 2015).

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

[11] Anna Kantosalo, Jukka Toivanen, and Hannu Toivonen, 'Interaction evaluation for human-computer co-creativity: A case study', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 276–283, (June 29 - July 2, 2015 2015).

[12] Florian Lachner, Philipp Naegelein, Robert Kowalski, Martin Spann, and Andreas Butz, 'Quantified ux: Towards a common organizational understanding of user experience', in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 56:1–56:10, New York, NY, USA, (2016). ACM.

[13] Todd Lubart, 'How can computers be partners in the creative process: classification and commentary on the special issue', *International Journal of Human-Computer Studies*, **63**(4), 365–369, (2005).

[14] Craig M. MacDonald and Michael E. Atwood, 'Changing perspectives on evaluation in hci: Past, present, and future', in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pp. 1969–1978, New York, NY, USA, (2013). ACM.

[15] Stuart MacFarlane, Gavin Sim, and Matthew Horton, 'Assessing usability and fun in educational software', in *Proceedings of the 2005 conference on Interaction design and children*, IDC '05, pp. 103–109, New York, NY, USA, (2005). ACM.

[16] Mary L. Maher, 'Evaluating creativity in humans, computers, and collectively intelligent systems', in *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, pp. 22–28, Lancaster, UK, UK, (2010). Desire Network.

[17] Mary L. Maher, 'Computational and collective creativity: Whos being creative?', in *Proceedings of the third International Conference on Computational Creativity*, pp. 67–71, (May 30 - June 1 2012).

[18] Panos Markopoulos and Mathilde Bekker, 'On the assessment of usability testing methods for children', *Interacting with Computers*, **15**(2), 227–243, (2003).

[19] Kumiyo Nakakoji, 'Meanings of tools, support, and uses for creative design processes', in *International Design Research Symposium 06, Seoul, Korea*, pp. 156–165, (2006).

[20] Mona Patel and Christine A. Paulsen, 'Strategies for recruiting children for usability tests', in *meeting of the Usability Professionals Association: FL (June 2002). Available: http://www. air. org/usability/publications/christinepaulsen/recruitingchildren. pdf (Retrieved at December 7 2005)*, (2002).

[21] Alison Pease and Simon Colton, 'Computational creativity theory: Inspirations behind the face and the idea models', in *Proceedings of the Second International Conference on Computational Creativity*, (2011).

[22] Alison Pease, Daniel Winterstein, and Simon Colton, 'Evaluating machine creativity', in *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*, pp. 129–137, (2001).

[23] Janet C. Read and Stuart MacFarlane, 'Using the fun toolkit and other survey methods to gather opinions in child computer interaction', in *Proceedings of the 2006 Conference on Interaction Design and Children*, IDC '06, pp. 81–88, New York, NY, USA, (2006). ACM.

[24] Mitchel Resnick, Brad Myers, Kumiyo Nakakoji, Ben Shneiderman, Randy Pausch, Ted Selker, and Mike Eisenberg, 'Design principles for tools to support creative thinking', (2005).

[25] Graeme Ritchie, 'Assessing creativity', in *Proceedings of the AISB symposium on artificial intelligence and creativity in arts and science, York, England*, pp. 3–11. Citeseer, (2001).

[26] Graeme Ritchie, 'Some empirical criteria for attributing creativity to a computer program', *Minds and Machines*, **17**(1), 67–99, (2007).

[27] Ben Shneiderman, 'Creativity support tools: accelerating discovery and innovation', *Commun.ACM*, **50**(12), 20–32, (dec 2007).

[28] Gavin Sim, Brendan Cassidy, and Janet C. Read, 'Understanding the fidelity effect when evaluating games with children', in *Proceedings of the 12th International Conference on Interaction Design and Children*, IDC '13, pp. 193–200, New York, NY, USA, (2013). ACM.

[29] Gavin Sim, Stuart MacFarlane, and Janet Read, 'All work and no play: Measuring fun, usability, and learning in software for children', *Computers & Education*, **46**(3), 235–248, (4 2006).

[30] Frank van der Velde, Roger A. Wolf, Martin Schmettow, and Deniece S. Nazareth, 'A semantic map for evaluating creativity', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 94–101, (June 29 - July 2, 2015 2015).

[31] Arnold P. O. S. Vermeeren, Effie L. Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila, 'User experience evaluation methods: Current state and development needs', in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, NordiCHI '10, pp. 521–530, New York, NY, USA, (2010). ACM.

[32] Matthew Yee-King and Mark d'Inverno, 'Experience driven design of creative systems', in *Proceedings of the Seventh International Conference on Computational Creativity*, pp. 85 – 92, (June 27 - July 1 2016).