The Fifth International Conference on Design Creativity (ICDC2018)
Bath, UK, January 31st – February 2nd 2018

EXPLORING WAYS TO SPEED UP THE APPLICATION OF METRICS TO ASSESS CO-CREATIVE DESIGN SESSIONS

E. A. Dekonick[1], J. O'Hare[1], L. Giunta[1], C. Masclet[2] and G. Cascini[3]

[1]Department of Mechanical Engineering, University of Bath, Bath, UK
[2]G-SCOP, CNRS, Université Grenoble Alpes, Grenoble, France
[3]Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy

Abstract: Analysing and comparing the results from design experiments is a crucial but often time consuming process. This problem becomes critical when a large number of experiments are to be analysed in order to evaluate the performance of a new tool or technology. In this paper we present an example of how this challenge is being addressed within the SPARK (Spatial Augmented Reality for Co-Creativity) H2020 project, which is currently developing an augmented-reality system to assist designers and non-designers to co-create more effectively. To evaluate the co-creative sessions a suite of metrics has previously been defined for the project. The scope of this paper is to radically speed up the application of the metrics by adjusting metrics and - more importantly - the way of collecting the data on co-creative sessions. The paper will be of interest to others doing quantitative analysis of creative design sessions.

Keywords: *design metrics, co-creation, mixed prototype, Spatial Augmented Reality, data collection*

1. Introduction

The SPARK (Spatial Augmented Reality for Co-Creativity) platform is an augmented reality system, currently under development within a H2020 ICT project (http://spark-project.net/), that will aid designers and non-designers co-create more effectively. This platform allows stakeholders to visualize and modify - in real time - digital representations projected on to a physical mock-up, this is referred to as a mixed prototype. By introducing spatial augmented reality into the co-creative process it is thought possible to make co-design sessions more creative, effective and efficient as the tangible nature of a SAR prototype should help to reduce the requirement for advanced design communication skills, making it easier for non-designers and designers to exchange thoughts and ideas about a proposed design.

The continued development of such a platform relies on repeated user testing where participants engage in co-design sessions using the Spatial Augmented Reality (SAR) technology. To evaluate the design sessions and understand the impact the SAR technology is having on the design process, a suite of metrics is currently in use. These metrics have considerable drawbacks in their implementation, notably,

the significant time it takes to apply them. The scope of this paper is to improve the metrics currently in use by adapting the metrics and the way in which they are applied in order to develop an approach that is equally reliable and repeatable as the current one but requiring less time to be implemented. The paper presents two approaches that were trialled to radically reduce the time needed to implement the suite of metrics.

2. Previous Work on Metrics for co-creative design sessions

In this section we describe what a co-creative design session is and review previous attempts to measure the effectiveness of co-creative design sessions before going on to introduce the suite of metrics developed within the SPARK project and the limitations of those metrics.

2.1. What is a co-creative design session?

Sanders and Stappers (2008) discuss the etymology of co-design and co-creation which are both subsets of the participatory design category. Whereas participatory design simply focuses on having the user as a partner in the design process, 'co-creation' (collaborative creation) refers to the more specific act of collaborative creativity. Furthermore, 'co-design' (collaborative design) is the application of co-creation throughout the design process. Sanders and Stappers (2008) ultimately define co-design as "… collective creativity as it is applied across the whole span of a design process". In addition, they specify that co-design can be used to describe the collaboration of both designers and non-designers in the design development process. In line with these definitions, we define a co-creative design session as: a pre-arranged session that involves designers and people not trained in design working together in the design development process.

2.2. Literature review on assessing the effectiveness of co-creative design sessions

Several authors have also begun to study the effectiveness of co-creative design sessions. Perttula, et al. (2006) mention that while it appears that the total number of ideas generated increases when co-design is implemented, the quality and diversity of these ideas decreases. Franke and Piller (2004) and Gultekin-Atasoy et al. (2014) have looked more towards studying the design sessions themselves to understand the benefits co-design may provide, without attempting to intervene in the design process. Beyond these examples, there has been limited previous work describing how to assess the effectiveness specifically of co-creative design sessions. However, several authors have considered how to measure creativity within conventional creative design sessions. Such measures can be broadly categorised as being either process- or outcome-based (Shah et al, 2003). In essence, process-based approaches observe the creative cognitive processes or creative thinking. Process-based approaches tend to be subjective - as no commonly agreed upon techniques to conduct them exists - and time consuming (Verhaegen et al, 2011). Also, it is difficult to establish firm causal links between the occurrence of a cognitive process and the effectiveness of an idea generation tool/technique (Shah et al, 2003).

Outcome-based approaches have gained in popularity in recent years as they address some of these concerns by focusing on the (external, observable) ideas generated rather than the (internal, unobservable) process by which they are generated. With this type of approach, an idea generation tool/technique can then be considered to be effective if its use leads to "good" ideas. The metrics developed for outcome-based approaches therefore attempt to relate aspects of the generated design ideas to the effectiveness of the applied idea generation method (Verhaegen et al, 2011). In the seminal work by Shah et al (2003), four metrics were proposed to measure the effectiveness of an ideation method, covering the quality, quantity, variety and novelty of the ideas generated. These metrics have been used in many subsequent studies, with several authors proposing refinements (e.g. Nelson, 2009).

2.3. The original co-creative session performance metrics: definition, application and problems

The co-creative session performance metrics, initially proposed in O'Hare et al. (2016), relate to the quality of outputs and effectiveness of co-creative sessions. They are intended to be used to assess a range of different types of co-creative sessions from idea-generation sessions to review-focused sessions. Whilst the metrics have been developed as part of the SPARK project, which has a focus on

the use of augmented-reality technology in product and packaging design co-creation sessions, we believe that they can be more generally applied to any type of (co-)creative design session.
The first four of these metrics, shown in Table 1, were based on established metrics for creativity (Shah et al, 2003) but adapted for application within the SPARK project, whereas the last two are new. The proposed suite of metrics, along with a description of each of the metrics and how to calculate them, is presented in Table 1. An overview of the metric implementation process is shown in Figure 1.

Table 1. Original co-creative session metrics (O'Hare et al., 2016)

Theme	Metric title	Description and definition
Quality of outputs	Quantity	New ideas generated during the session *No. of sub-solutions in 'After' morphological chart - No. of sub-solutions in 'Before' morphological chart*
	Quality	New ideas generated that are taken forward for further development *No. of new ideas - No. of rejected ideas*
	Variety	How varied the ideas generated during the session are *(No. of original feature rows with new sub-solutions ÷ No. of original feature rows) + (No. of new feature rows ÷ No. of original feature rows)*
	Novelty	How novel the ideas generated during the session are *Mean average novelty score from three experts for each idea generated*
Effectiveness of co-creative sessions	Filtering effectiveness	The success of the filtering activities during the session *No. of ideas rejected ÷ (No. of ideas rejected - Desired No. of ideas to keep)*
	Task progress	How the participants in the session progress through tasks *Weighted scoring for tasks completed or generated*

1.2.1. Problems with the original way of applying the metrics

The main problems that caused the implementation of the metrics to take too long were described by Mombeshora et al. (2017), and can be summarised as follows. It was known in advance that most of the sessions would be conducted in either Italian or Spanish, languages not spoken by the research team applying the metrics. It was envisaged that the pre- and post-session interviews would provide sufficient context about the project and the aims of the session to allow the research team to complete the analysis with the aid of the video recordings of the session. In practice, this was not possible as the interviews did not provide sufficient detail and context. As a result, it was necessary to wait for the transcription and translation of the session recordings - which took several weeks to be completed.
Once the transcripts were received, the researcher was able to draft the 'Ideas Chart' (a brief summary of each idea generated in the session) by reviewing the session transcripts and video recordings. This Ideas Chart was validated with one of the creative practitioners involved in the session, often several weeks after the session had taken place. There were two problems with this approach. First, it was very challenging and time consuming for the researcher to draft the Ideas Chart because it was not very clear from the transcripts *when* ideas had emerged, even with the aid of the video recordings. Secondly, the significant time elapsed between finishing the session and reviewing the Ideas Chart with one of the practitioners (several weeks) meant that the practitioners might well have forgotten some ideas from the session, or might describe them in a different way having had time to reflect on the session.
There were also problems in creating the Morphological Charts for the session, which were required for the Variety and Filtering Effectiveness metrics. The intended approach was that the researcher would create a 'before' morphological, chart representing the start of the session, by reviewing the information obtained from the pre-session interviews as well as the design representations that were brought into the session by the design team (if any). An 'after' Morphological Chart would then be constructed based on the ideas still under consideration at the end of the session. The difference between the two charts would show the evolution that occurred over the course of the session. However, the lack of depth in understanding of the project and lack of depth pre- and post-session interview meant that this proved impossible. Instead a collaborative analysis with one of the designers from the session was necessary. During this session the Morphological Chart was generated and the scoring took place for the Novelty

and Quality metrics. The ‘check ideas’ stage shown in Figure 1 was therefore a considerable activity, taking an equivalent amount time as the recorded design session itself.

3. Methodology

This paper reports on two attempts to speed up the implementation of these metrics and the checks that were conducted to see whether their use resulted in similar scores in comparison to the originally proposed method by Mombeshora (2017), which we refer to as the ‘transcript based’ approach. In order to achieve this, the scores from the sessions previously conducted (‘A’ sessions) were compared to two new rounds of analysis using two completely different approaches to speed up the process. The first new approach, referred to as the ‘participants’ perceptions’ approach, made use of self-reported scores about the performance of the session provided by the participants. The second new approach, referred to as the ‘live capture’ approach, made use of an observer in the session to apply the metrics. The new sessions (‘B’ and ‘C’ sessions), conducted with the same design consultancies as the reference sessions (A), were then also coded using the original, transcript based metric application process. The assumption was that the transcript based approach was the most robust and reliable, based on objective data, and the scores coming out the two new approaches were then compared against the transcript based approach in terms of the scores generated and the time taken to apply them.

3.1. Using participants’ perceptions to measure the outcomes from co-creative sessions

Two complete co-creative design sessions (‘B’ sessions) were conducted with two design consultancies: Stimulo and Artefice. They aimed to eliminate the need to generate an Idea Chart or any Morphological Charts - which had proven so problematic and time consuming to create. The approach consisted of three major elements. First, a live observer to record ideas as they were generated during the session. Secondly, once the session was completed the participants were asked to fill in a questionnaire. This was followed by the third element, an interview with structured questions.

An observer would be embedded in the session, observing and collecting ideas as the session ran. The ideas collected by the observer live were then be used in the post-interview session. The post-session interview could be conducted immediately after the design session without having to wait for additional data processing, meaning that the novelty of the individual ideas could be scored immediately.

The questionnaire was given to each of the designers/facilitators that had participated in the co-creative session. The questions asked the respondents to rate subjectively (on a Likert scale) the quality, quantity and variety of the ideas generated during the session compared to a normal co-creative session. These data were used to evaluate the Quality, Quantity and Variety metrics. Note that a ‘normal’ co-creative session had previously been completed with each of the design consultancies (‘A’ sessions) and analysed using the transcript based approach. This meant it was possible to use these earlier sessions as the benchmark. The transcript based approach was also applied to the ‘B’ sessions, in order to make a comparison of the approaches.

3.2. Using a live capture approach to measure the outcomes from co-creative sessions

Another two complete co-creative design sessions (‘C’ sessions) were conducted with the same two design consultancies. The approach in this instance consisted of three major elements: a pre-session interview, a live observer to create the Morphological Chart on-the-fly, and a post-session interview to validate the ideas captured and the Morphological Chart generated.

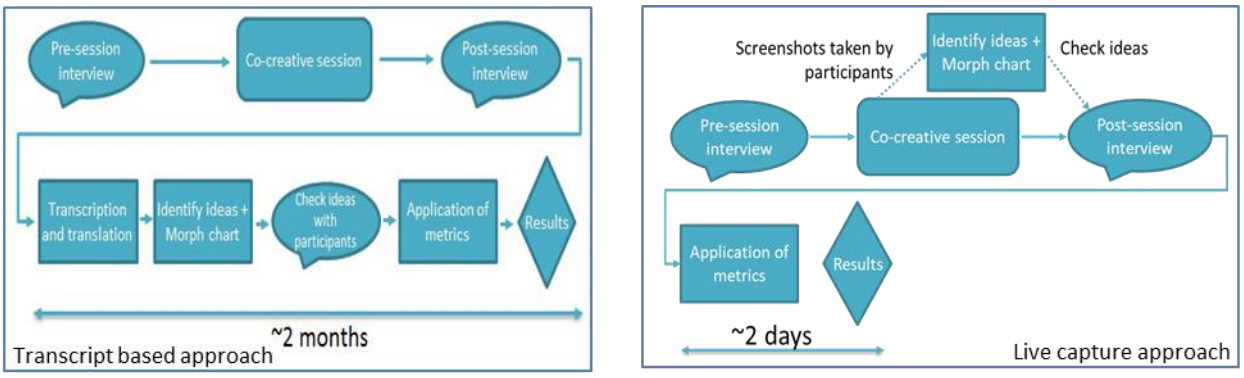

Figure 1. Implementation process for the transcript based approach and live capture approach

Figure 1 provides a side-by-side comparison of the transcript based and live capture approaches. One of the main novelties of the live capture approach was that the ideas were captured by the participants themselves, which they did by taking a screenshot of the SAR prototype whenever they felt they had generated a new idea. All the screenshots captured by the participants during the session were presented back to them during the post-session interview. This gave them an opportunity to discard any screenshots that were taken by mistake and add ideas that had not been captured during the session. Similarly, the Morphological Chart created by the observer was presented to the participants and its validity checked. It was not possible to apply the metrics to these case studies the using transcript based approach as no session transcripts were created (due to time and cost limitations).

4. Results and Discussion

The following sections present the results of applying the metrics using the 'participants' perceptions' approach and the 'live capture' approach.

4.1. Participants' perceptions measures

The two design sessions ('B' sessions) conducted were markedly different to each other, making a direct comparison of them inappropriate. Stimulo worked with a simulated consumer and focused on materials, colours and finishes as they attempted to develop the final aesthetics of a product. Stimulo encountered no issues with the use of the SPARK platform and quickly integrated the technology into their co-creative activities. In contrast, Artefice's co-creative session was focused on designing and developing packaging and how to convey important information through that packaging. Artefice had considerable difficulties in the use of the platform, due in part to the large set of images (200+ images) they uploaded for use in the SAR system.

The results are presented in Table 2 for Stimulo and Table 3 for Artefice. The first column shows the result from the 'normal' co-creative sessions ('A' sessions) that were completed with the same design consultancies but without the aid of the SAR system. The second column shows the scores from the 'B' sessions, this time with the aid of the SAR system. Two designers/facilitators were present in each of the sessions, with the average of the two 'self-reported' scores shown. The third column shows the 'perceived performance factor'. This factor was calculated from the self-reported scores as follows. The 5-point Likert scales used to capture the self-reported scores gave the following guidelines to help the participants score the performance of the session:

- Score of one - session had resulted in half the performance of a normal session.
- Score of three - session was very similar to the performance of a normal session.
- Score of five - session resulted in twice the performance of a normal session.

Scores were recorded using this scale for the quantity of ideas, variety of ideas and task progress metrics. The five-point Likert scale was used as a compromise between granularity of response options and time taken to complete the survey (as respondents tend to take more time to select a response when there are more options). The descriptors of 'half the performance', 'similar performance' and 'twice the performance' were provided to ensure that participants evaluated the session in a similar manner.

The perceived performance factor was calculated by converting the average score for each metric to a decimal factor, using linear interpolation. For example, the average self-reported score for quantity by the Stimulo designers was 4.5, which gave a multiplication factor of 1.75.
The fourth column shows the 'perceived performance score' - which was calculated by multiplying the 'normal' session score (from the 'A' sessions) by the perceived performance factor. The final column shows the 'actual score' calculated using the transcript based approach (Mombeshora et al, 2017). 'NA' indicates in the table where it was not possible to calculate a score.

Table 2. Results of the metric application to the Stimulo sessions using the participants' perceptions approach

	Normal 'A' session (Y)	**Average self-reported score (out of 5) for 'B' session**	**Perceived performance factor (Z)**	**Perceived performance score (YxZ)**	**Actual score for 'B' session**
Quality	7	NA	NA	NA	4
Quantity	13	4.5	1.75	**22.75**	**12**
Variety	1.98	4.5	1.75	**3.45**	**1.69**
Novelty	5.73	NA	NA	NA	7.77
Filtering Effectiveness	0.65	NA	NA	NA	NA
Task Progress	24	4.5	1.75	42	NA

Table 3. Results of the metric application to the Artefice sessions using the participants' perceptions approach

	Normal 'A' session (Y)	**Average self-reported score (out of 5) for 'B' session**	**Perceived performance factor (Z)**	**Perceived performance score (YxZ)**	**Actual score for 'B' session**
Quality	4	NA	NA	NA	2
Quantity	5	3.75	1.38	**6.89**	**15**
Variety	7.5	3.25	1.13	**8.44**	**3.45**
Novelty	NA	NA	NA	NA	3.33
Filtering Effectiveness	1	NA	NA	NA	NA
Task Progress	NA	3.33	1.17	NA	NA

The final two columns allow us to perform a basic check of the validity of a couple of the self-reported scores. For instance, the Stimulo designers' perception was that the B session (using the SAR system) had resulted in nearly twice as many ideas as a normal session. In fact, the actual quantity of ideas (12) was significantly less than the average generated in the two normal sessions (21.5). This overestimation of the performance from the self-reported scores is seen elsewhere, with all the results that we were able to compare (the shaded boxes) incorrect by a factor of two or three.
It thus appears that the participant's perception approach to the application of the metrics is an unsatisfactory replacement for the original, transcript based approach. Whilst it was possible to apply the metrics with considerably less time and effort, the initial results are significantly different to those obtained using the transcript based approach - which is assumed to be the more accurate one. The inaccuracies may be due to variance in the nature of each co-creative session or it may be that the participants overestimated the session performance because they were subconsciously biased in favour of the SAR system. Whatever the cause, it was decided not to pursue the participants' perceptions approach any further.
Despite this set back, a number of interesting observations were made. First, the use of a live observer sat in the room (who was fluent in the language being used in the session) helped to understand the events of the design session and keep track of the ideas generated in real time. This enabled the researcher to conduct the post-session interview immediately after the design session and to seek

clarifications from the participants about the ideas generated. This significantly reduced the time and effort required to complete the Idea Chart.
Another interesting observation was that the participants regularly used the screenshot function present in the SPARK platform without being instructed or prompted to do so. Every time the participants generated an idea they were interested in they took a screenshot. This action was a natural behaviour that occurred in both the Stimulo and Artefice sessions and did not appear to influence the progression of the design session.

4.2. Live capture approach

In July 2017, a further two co-creative sessions ('C' sessions) supported by the SPARK platform were completed, one with Stimulo and one with Artefice. The 'live capture' approach to applying the metrics was used. The reporting of the Variety metric was altered at this point, splitting it into two components. The 'Variety Coverage' component was defined as the number of pre-existing rows of the Morphological Chart for which new ideas were generated. The 'Variety New Rows' component was defined as the number of new rows to the Morphological Chart for which new ideas were generated. The results are shown in Table 4.

Table 4. Results of the metric application to the Stimulo sessions using the live capture approach

	Normal 'A' session	**Actual score for 'B' session**	**Average of sessions A and B**	**Scores for 'C' session**	**Difference (average A and B vs C)**
Quality	7	4	**5.5**	**6**	9%
Quantity	13	12	**12.5**	**10**	20%
Variety Coverage	5	5	5	6	20%
Variety New Rows	3	2	2.5	3	20%
Novelty	5.73	7.77	**6.75**	**7.3**	8%
Filtering Effectiveness	0.65	NA	0.65	0.44	32%
Task Progress	24	NA	24	8	67%

The results from the Artefice sessions are presented in Table 5.

Table 5. Results of the metric application to the Artefice sessions using the live capture approach

	Normal 'A' session	**Actual score for 'B' session**	**Average of sessions A and B**	**Scores for 'C' session**	**Difference (average A and B vs C)**
Quality	4	2	**3**	**3**	0%
Quantity	5	15	**10**	**6**	40%
Variety Coverage	2	7	4.5	5	11%
Variety New Rows	0	14	NA	0	NA
Novelty	NA	3.33	**3.33**	**3.79**	14%
Filtering Effectiveness	1	NA	1	0.67	33%
Task Progress	NA	NA	NA	12	NA

Again the data sets could not be completed entirely due to the iterative changes that were made between the A, B and C sessions at both design consultancies, and - in some cases - it was not possible to go back and get a complete set of scores. The fairest comparison can be made in places where the rows are complete and the scoring method stayed most similar. These are figures in bold text. The percentage differences for those are now more reasonable. There are the percentage differences between the sessions which show variation between the sessions but the scores are now no longer significantly misaligned, as they were for the 'participants' perceptions' approach.

5. Conclusions

The aim of this research activity was to reduce the time and effort required to apply a suite of co-creative session performance metrics. The original, transcript based approach was too time consuming as it required full transcripts for each session and an interview with the participants several weeks after the session. A second approach was tested that relied on the self-reported perceptions of the session performance from the designers in the session. The comparison with results calculated using the transcript based approach suggested that the self-reported scores were significantly different to the actual results of the session. The lessons learnt from this led to a third 'live capture' approach, which involved: the participants themselves capturing their ideas by taking screenshots; having an observer sat in the session noting ideas and generating the Morphological Chart on-the-fly; and a post-session interview, in which the ideas and Morphological Chart were validated with the participants.

Whilst the ability to make direct comparisons between the three approaches was limited by the changes in the data capture methods, the types of co-creative session being analysed, and the limited number of sessions analysed, it appears that the third approach provides results in a similar range to the transcript based approach whilst enabling a reduction in the time and effort required for analysis from around two months to two days - see Figure 1. The 'live capture' approach will be taken forward for further use within the forthcoming testing of the SPARK project.

The reliability of capturing the Morphological Chart on-the-fly has just recently been tested with a first inter-rater check and found not to be as reliable as we would like. This is being addressed through the development of coding guidelines for this specific research task and more thorough inter-rater checks. Further work also involves checking the metrics and the 'live capture' approach in a set of controlled experiments where very similar creative outcomes would be expected and the same scores would be expected from different raters.

Beyond their application within the SPARK project, we suggest that the co-creative performance metrics and the metric application methodology are sufficiently generic that, with further refinement, they could be applied by other researchers investigating co-creative session performance across domains including engineering design, architectural design, software engineering etc.

Acknowledgement

The work reported in this paper was completed as part of the SPARK project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.688417. This paper reflects only the authors' views and the European Commission is not responsible for any use that may be made of the information it contains.

References

Franke, N. and Piller, F. (2004). Value Creation by Toolkits for User Innovation and Design: The Case of the Watch Market. *Journal of Product Innovation Management, 21*(6), pp.401-415.

Gultekin-Atasoy, P., Lu, Y., Bekker, T., Eggen, B. and Brombacher, A., 2014, August. Evaluating value design workshop in collaborative design sessions. In *Proceedings of the NordDesign 2014 conference*, Helsinki, Finland.

Mombeshora, M., Dekoninck, E., O'Hare, J., Boujut, J. F., & Cascini, G. (2017). Applying multiple metrics in the performance measurement of design sessions in industry: a co-design case study. In *Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 2: Design Processes, Design Organisation and Management.* Vancouver, Canada.

Nelson B. A., Wilson A. O., Rosen D., Yen J., 2009. Refined metrics for measuring ideation effectiveness, *Design Studies, 30*(6), pp. 737-743.

O'Hare, J., Mombeshora, M., Varvatis, C., Ben-Guefrache, F., Masclet, C., Prudhomme, G., Martens, P. and Becattini, N. (2016). Results from the Experimental activities and presentation of the research metrics framework. [Online] SPARK Project. Available at: http://spark-project.net/wp-deliverables.

Perttula, M., Krause, C. and Sipilä, P. (2006). Does idea exchange promote productivity in design idea generation?. *CoDesign, 2*(3), pp.125-138.

Shah, J., Smith, S. and Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies, 24*(2), pp.111-134.

Sanders, E. and Stappers, P. (2008). Co-creation and the new landscapes of design. *CoDesign, 4*(1), pp.5-18.

Verhaegen, P.A., Peeters, J., Vandevenne, D., Dewulf, S., Duflou, J.R., (2011). Effectiveness of the PAnDA ideation tool. *Procedia Engineering, 9*, pp. 63–76.