SPARK

D4.2

RESULTS OF THE EXPERIMENTS BENCHMARKING THE PLATFORM

Approval Status

	Name and Surname	Role in the project	Partner
Author(s)	Fatma Ben-Guefreche Jean-François Boujut Cédric Masclet Maud Poulin Guy Prudhomme Niccolò Becattini Nicolas Carbone Jamie O'Hare Lorenzo Giunta	Researcher Researcher Researcher Researcher Researcher Researcher Researcher Researcher Researcher	GINP GINP GINP GINP GINP POLIMI POLIMI UBATH UBATH
Reviewed by	Jean-François Boujut Elies Dekoninck Niccolò Becattini Gaetano Cascini	WP4 Leader Researcher Assistant PC Project Coordinator	GINP UBATH POLIMI POLIMI
Approved by	Gaetano Cascini	Project Coordinator	POLIMI

History of Changes

Version	Date	Description of Changes	By
0.1	26/01/18	Collection of contributions by all partners involved in the preparation of the deliverable	Fatma Ben-Guefreche Jean-François Boujut Cédric Masclet Maud Poulin Guy Prudhomme Niccolò Becattini Nicolas Carbone Jamie O'Hare Lorenzo Giunta
1.0	12/02/18	Creation of complete draft	Jamie O'Hare
1.1	16/02/18	Review and suggested revisions	Elies Dekoninck
1.2	20/02/18	Revision of section 4.4	Niccolò Becattini
1.3	21/02/18	Revision of section 4.3	Maud Poulin
2.0	23/02/18	Draft V2.0 for final revision	Jean-François Boujut
2.1	24/02/18	Proof-read for submission to PC	Jamie O'Hare
3.0	26/02/18	Final review and editing	Gaetano Cascini

Document Details

Dissemination Level	Public
Due Date	28.02.2018
Issue Date	26.02.2018
Contract Number	H2020-ICT/2015-688417
Electronic file location	Codendi, Spark web site
File name	D4.2_WP4_Results of the experiments benchmarking the platform

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report describes the activity of Tasks 4.3, 4.4 and 4.5 and mainly focuses on the presentation of the results of the platform test in controlled environments with design teams and end-users[1]. This study aims at responding to Objective 3 of the SPARK project: "Study and analyse how and to what extent the SAR technology can stimulate and enhance design creativity through a comparison against a pre-defined metrics in real operational environment".

The report presents analyses of the two sets of experiments carried out in autumn 2017, both in Grenoble and Milan, involving design teams from Stimulo and Artefice, design agencies partners of the SPARK Consortium. During these experiments, we had the opportunity to test three conditions, as defined in D4.1. One 'standard' condition, one with an ICT solution and another one with the Spatial Augmented Reality (SAR) platform developed by the SPARK consortium. We performed four types of analysis:

- A performance analysis regarding co-creation and idea generation
- A gesture-based analysis of end-users/designers interactions
- A speech-based analysis of design moves
- And a usability survey at the end of each session and a transversal project analysis

The co-creative sessions performance analysis provided interesting preliminary results on the platform use. Stimulo SAR sessions resulted in improved idea generation (quantity, variety, quality and novelty) and task progress compared to sessions with standard design representations but differences were not significant in some cases and were not better than AR sessions. Artefice SAR sessions resulted in improved idea generation in terms of quantity and quality of ideas compared to sessions with standard design representations but was worse or the same in terms of variety and novelty of ideas, task progress and filtering effectiveness. The SAR technology, at its current level of development and with the user interface available in Autumn 2017, has shown good potential from the results of the co-creative performance metrics but has not consistently outperformed the other conditions across all metrics and both companies.

Gesture analysis allowed to analyse the influence of the technology on the intensity and type of interactions the participants had during the co-design sessions. This is an indication of the usability of the platform and the engagement of the participants in the design task. Our results show that the Spatial Augmented Reality condition does not favour the end-users' interactions, being they Stimulo or Artefice. However, we observe a comparable percentage of interaction, which is encouraging for the SPARK project as it tends to demonstrate the interest of the platform in diverse co-design situations. Additionally, we noticed a relatively stable pace of interaction in all the sessions regardless

[1] The term 'end-users' in the case of the two experiments of WP4 were intended to be representative of potential consumers of the product being worked.

of the condition. This tends to suggest that there is no clear effect of the conditions (including the technology) on the rhythm of the sessions, suggesting that SAR technology at this stage of its development does not foster more interactions. The number of virtual artefacts referred to by participants was lower in the SAR condition than in the AR and Standard conditions. The profile of the graphs suggests that there is an effect of the technology on the intensity of virtual artefacts in the co-design sessions.

Concerning speech analysis, the degree of involvement is measured by the shifts between speakers and the category of speakers during verbal interactions. The creativity in terms of fluency represents the quantity of new ideas generated during a co-creative session. We attempted to verify if Spatial Augmented Reality (SAR), as the shared design representation (mixed prototype), releases some cognitive load from working memory, enabling a more fluent generation of ideas (Fluency increases). As for creativity in terms of quality of ideas, it depends on the matching between the exploration of the problem and the solution space (within the design space). Here we attempted to verify if Spatial Augmented Reality (SAR) facilitates co-designers to efficiently explore design alternatives (both to identify requirements and propose design changes). Creativity in terms of convergent thinking corresponds to the capability to make selections among various alternatives in design (whole design proposals or part of them). In that case we attempted to verify if Spatial Augmented Reality (SAR) facilitates co-designers to select contents to be used for the configuration of product interface and the composition of the packaging.

Usability study and longitudinal study: Overall system usability was rated significantly better for the SAR system than the 'standard' system but was slightly worse than the AR system. Most important aspects of usability were 'Collaboration', 'Exploration' and 'Immersion'. SAR system scored poorly on 'Immersion' aspect, probably due to the technical difficulties encountered during the SAR sessions. Designers were reasonably satisfied with the usability of the SAR system but there is still room for improvement. The Follow-up Survey indicates that designers appreciated the capability of the SAR technology to enhance collaboration with the end-users and the ability to make quick, real-time modifications on the prototype. SAR technology was also useful for quickly eliminating poor ideas. Several technical challenges were identified that require further attention, including poor system reliability and differences in colour rendering between the end-users view and the designer's view on the tablet PC user interface. Designers can see the potential value of SAR technology to support co-creative design sessions but some significant technical challenges and limitations of the current prototype technology need to be addressed.

This report shows a significant step forward in terms of usability study and performance analysis of the SPARK SAR platform. The methodology applied is complex and robust. If the results appear less than conclusive they must be considered to reflect the state-of-the art of the technology being developed. WP4 provides important input for the improvement of the user experience of the platform, as well as precious learnings for WP5 experiments.

1. INTRODUCTION

This deliverable reports on the results of experiments that are aimed at building our understanding of how, and to what extent, the SPARK Spatial Augmented Reality (SAR) technology can stimulate and enhance creativity within co-creative design sessions. Furthermore, the experiments are intended to provide some initial evidence of the effectiveness of the SPARK platform in comparison with competing solutions and standard practices in realistic operational environment (SPARK Objective 3). Finally, the experiments should provide some useful feedback and insights concerning the real-life use of the technology that can inform the final technology development activities, which are due for completion by M31 (31st July 2018).

Initial planning of the experimental activities and analysis procedures was provided in D4.1 'Definition of the experimental protocol for a creative design process and case studies'. Within this report, Section 2 provides an overview of the experimental protocol and experimental groups. Section 3 provides specific details of the methodology as implemented. This includes details of the various analysis techniques applied, which cover:

- the performance of the sessions in terms of their creative outputs;
- the gestures of the session participants;
- the spoken interactions of the session participants; and,
- the improvements to the efficiency of the overall design process.

Section 4 provides the results of the experimental sessions based on the application of the aforementioned analysis techniques and also - through a qualitative review - the lessons learnt from each session. Section 5 provides conclusions on the results presented.

2. OVERVIEW OF THE EXPERIMENTAL PROTOCOL AND EXPERIMENTAL GROUPS

In order to compare the results of the experiments and help to identify clear differences, in D4.1 we have proposed to set up three different conditions for the experiment:

- Condition 1: with standard design representations;
- Condition 2: with state of the art ICT technology;
- Condition 3: with SAR technology (the SPARK technology).

An Augmented Reality (AR) version of the SPARK technology was selected as the 'state of the art' technology to be used in Condition 2. The main reasons for this choice were that: there is currently significant interest in using AR technology in design, and, it would enable a clear examination of the benefits of SAR technology over AR technology, as both conditions would be using the same basic SPARK technology in terms of the user interface used by the designer and the backend Information System.

Those three conditions were replicated in two different locations, with one industrial partner participating at each venue. GINP hosted tests with Stimulo, whilst POLIMI hosted the tests with Artefice. Table 1 summarises the experimental conditions for the two partners.

The 'SAR condition' involved the use of the SPARK Spatial Augmented Reality technology. One designer used the SPARK user interface, installed on a tablet PC (8" screen). The second designer and the end-users viewed the SAR prototype.

Table 1: Group and project distribution schema

	Artefice	**Stimulo**
Condition 1 – CG1 (Standard)	Team A1 / Project A	Team S / Project S1
Condition 2 – CG 2 (AR)	Team A2 / Project A	Team S / Project S2
Condition 3 – TG (SAR)	Team A3 / Project A	Team S / Project S3

For the 'AR condition', the visualisation of the augmented prototype was done through a tablet PC (8" screen) - see Figure 1. When viewed with the naked eye, the physical prototype is white with black triangles, which act as markers for the optical tracking system. One designer used the SPARK user interface, installed on a tablet PC (identical to the SAR condition). The second designer and the end-users shared one tablet PC through which they could view the digital overlay on the prototype.

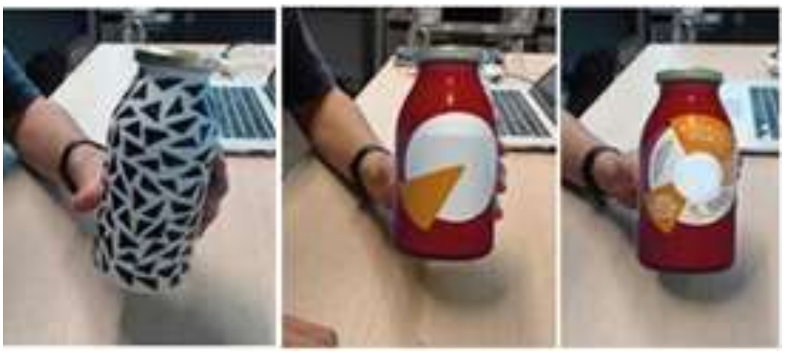


Figure 1: View of the Augmented Reality prototype i) without augmentation, showing the marker pattern ii) with augmentation - design in progress, iii) with augmentation - completed design proposal.

For the ‘standard’ condition, the designers were asked to use conventional materials and tools to prepare the design representations for their sessions. For the Stimulo ‘standard’ session, the initial designer proposals were displayed on a large television screen using presentation software. After this, physical prototypes (featuring neutral colours) were presented along with Pantone colour swatches were used to discuss alternative colour schemes - see Figure 2 (left). For the Artefice ‘standard’ session, the designers elected to use a collage method, which involved pre-preparing a variety of logos and graphic elements as stickers that could be applied to the cardboard sleeve of the soup packaging and re-positioned as required. Further elements were added by hand drawing directly on to the cardboard sleeve - see Figure 2 (right).

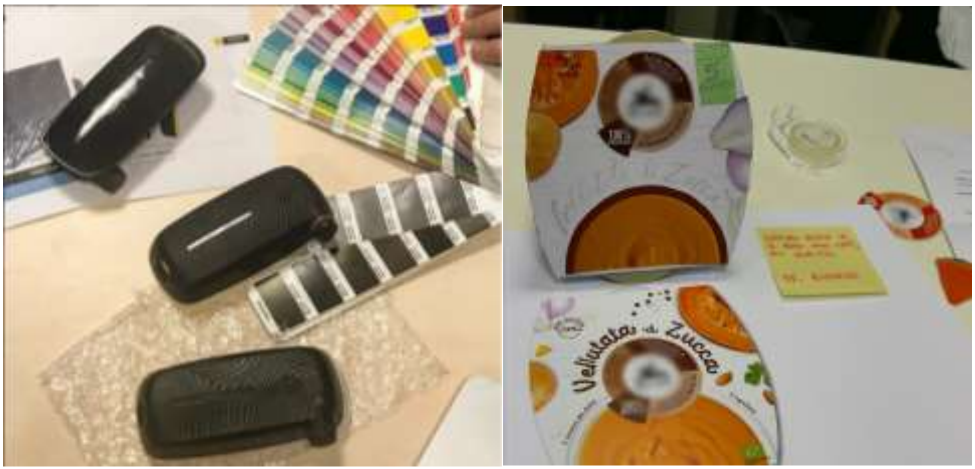


Figure 2: Left - Physical models and Pantone colour matching system used in the Stimulo 'standard' session. Right - collage system used in the Artefice 'standard' session.

3. METHODOLOGY

3.1. CASE STUDIES AND PARTICIPANTS

The experiments featured designers from Stimulo and Artefice, with three sessions organised with each industrial partner (i.e. six sessions in total, shown in table 1 above). For the Artefice sessions it was possible to have different teams of designers with equivalent expertise to work on the same product and same initial brief for each of the three conditions. For the sessions with Stimulo, only two designers were available to participate in the experiments. With this limitation, using the same product and brief for each condition was not desirable as it would have risked the designers (consciously or sub-consciously) carrying over ideas from one session to another, or becoming bored of the task due to the repetition. It was therefore necessary to vary the case study product for each of the conditions, although efforts were made to ensure the session briefs were as similar as possible in scope, task and design stage.

Table 2: Summary of the products, scope and participants for each condition.

		SAR condition	**AR condition**	**Standard condition**
Stimulo	Product description	Hand held device for assessment of human exposure to electromagnetic fields	Smart fitness product to monitor performance when using gym equipment	Hand held device for communicating your location in an emergency
	Session brief	Define the colours, materials and finish of the main housing. Define the location and pattern of LED status lights and speaker. Location of logo.	Define the colours, materials and finish of the main housing. Location of logo.	Define the colours, materials and finish of the main housing for specific environments. Define the location and pattern of LED status lights.

	End-users	Female, age 18-30 Male, age 18-30 Male, age 45-60	Female, age 18-30 Male, age 18-30 Male, age 45-60	Female, age 18-30 Male, age 18-30 Female, age 45-60
	Designers	Creative Director, 14 years of experience, male Designer and Business Developer, 15 years of experience, male		
Artefice	Product description	Fresh soup - single serving in plastic bowl with film lid and cardboard sleeve		
	Session brief	Further develop three pre-prepared alternative designs for the cardboard sleeve graphics and layout by combining graphical elements (colours, logos, text, images etc) in order to propose a complete packaging design.		
	End-users	Female, age 18-30 Male, age 18-30	Male, age 18-30 Male, age 18-30	Female, age 30-45 Female, age 30-45
	Designers	Digital Creative Director, 16 years of experience, female Art Director, 18 years of experience, female	Senior Art Director, 19 years of experience, male Graphic Designer, 10 years of experience, male	Art Director, 10 years of experience, female Junior Art Director, 1 year of experience, female

The teams of designers for each session were selected to ensure reasonable consistency across the conditions in terms of experience and skills. The ‘end-users’ in this case were intended to be representative of potential consumers of the product being worked and so they were selected to match the target demographic of the case study product for the session they would participate in i.e. for the Stimulo emergency location beacon session the end-users had to match the description of '18-60 years old adults who enjoy hiking in the mountains'.

3.2. CO-CREATIVE SESSION PERFORMANCE METRICS

The methodology used for the application of the co-creative session performance metrics is presented in the following sub-section. Two supplementary data gathering activities were completed to support the analysis of the co-creative session performance metrics. The first of these was a usability survey (described in Section 3.2.2), which was intended to help understand the designers’ experience of the technology they used during the session. The second supplementary activity was a follow-up survey (described in Section 3.2.3), which was intended to help understand the designers’ overall experience of the session.

3.2.1. Methodology for the application of the co-creative performance metrics

The co-creative performance metrics have been developed iteratively since the start of the project, with the final version presented in D4.1. Here we provide a brief summary of the metrics that were used to evaluate co-creative performance.

Table 3: The co-creative performance metrics to be applied in WP4.

Metric title	Metric definition
Quantity of ideas	Quantity of ideas generated during the session, counted as the number of screenshots taken by participants, verified in the post-session interview.
Variety of ideas	Variety (coverage) - Number of original feature rows that contain a new idea counted on the morphological chart created by a native-speaker observer in the session.
	Variety (new rows) - Number of new feature rows added counted on the morphological chart created by a native-speaker observer in the session.
Quality of ideas	Number of new ideas generated that are taken forward at the end of the session for further development. Determined by participants' consensus in the post-session interview.
Novelty of ideas	Novelty score from 1 to 10 for each of the ideas captured as a screenshot during the session. Determined by participants' consensus in the post-session interview.
Task Progress	Task Progress = 3pts x (Number of high importance tasks resolved or new tasks created) + 2pts x (Number of medium importance tasks resolved or created) + 1pt x (Number of low importance tasks resolved or created). Captured from pre- and post-interview with the session leader.
Filtering Effectiveness	Filtering Effectiveness = Number of ideas rejected ÷ (Number of ideas considered - Desired number of ideas to retain)

The hypothesis concerning these metrics is that: 'Co-creative design sessions completed with SAR-based design representations (TG) will result in improved idea generation (quality, variety, quality and novelty), task progress and filtering effectiveness compared to similar sessions completed with standard design representations (CG1) or AR-based design representations (CG2)'.

Application of the metrics involved a number of data gathering activities. First, a pre-session interview was conducted with the lead designer, just before the start of the session. The designer was asked about their objectives for the session, what ideas had previously been generated in the project, any open tasks from previous sessions (for the Task Progress metric), and how many ideas they would like to end up with by the end of the session (for the Filtering Effectiveness metric).

At the start of the co-creative session itself, all participants were asked to request a screenshot/picture to be taken of the prototype whenever they felt that they had generated 'a new idea'. To avoid too much disruption to the session, the screenshots/pictures were taken by the designers, either using the screenshot feature built into the SPARK platform user interface, or by taking a photograph with their cameraphone during the 'standard' sessions.

A researcher was sat in the room during each of the sessions to take live notes about the ideas that were being discussed. These notes were captured in the form of a table, based on a 'Morphological Chart', where the rows describe the feature or function that is being discussed, and a description of the potential embodiment options being discussed are captured in the columns. An example of this type of table is provided in Table 4.

Table 4: Example of the 'Morphological Chart'-type table used to for the Variety metric assessment

	Option 1	**Option 2**	**Option 3**	**Option 4**
Tomato graphic position	On right hand side	On left hand side		
Text: Polpa di pomodoro	In Italian	In English		
Italian map graphic	Small	Large		
Text: Picked and preserved in Italy	Picked and preserved in Italy	Picked and preserved within 10 hours		
Text: 100% Italian tomatoes	100% Italian tomatoes	On-plant ripened tomatoes	100% made in Italy	Grown and made in Italy
Text position: 100% Italian tomatoes	Around circle	In red quadrant		
Number of quadrants with text	4	3		
Basil leaves size	Small	Large		
Background	Transparent	White	Red	

After the session, a joint interview with both of the designers from the session was completed. In this session, the designers were presented with the screenshots/pictures of the ideas that had been captured during the session. They were asked to confirm that all the ideas had been captured and that none of the ideas were duplicates or captured by mistake (for the Quantity metric). The designers were then asked to rate each of the ideas in terms of their novelty on a scale from one (low novelty) to 10 (high novelty) (for the Novelty metric) and then asked to decide if each idea would be taken forward in the project for further development (for the Quality and Filtering Effectiveness metrics).

Next, they were then presented with the Morphological Chart that had been captured by the researcher and asked to confirm the accuracy of the chart. They were also asked to identify which of the rows, if any, described new features/functions of the product that had not previously been considered within the project (for the Variety metric).

Finally, the list of open tasks from pre-session interview was revisited to check which tasks had been completed and what new tasks, if any, had been generated during the session.

3.2.2. Methodology for the usability testing

In D4.1 it was proposed that an assessment of the usability performance of the platform be conducted using the NASA Task Load Index (NASA-TLX). Prior to the experiments, a variant of the NASA-TLX known as the Creativity Support Index (CSI) was selected. The CSI tool has been developed based on the NASA TLX survey method, but with a greater emphasis on evaluating creativity support tools (Cherry & Latulipe, 2014). It has previously been used to evaluate new methods for design interaction (Zaman et al, 2015) and creativity supports tools in areas such as architecture and design (Cherry & Latulipe, 2014), visualisation and simulation, video editing (Tang, Lee & Gero, 2011) and animation. The CSI tool was therefore selected over the standard NASA-TLX tool as the purpose of the CSI tool is more closely aligned with the objectives of this work package (evaluation of a creativity support tool).

The hypothesis concerning the CSI survey was that: 'The designers' rating of system usability will be better (higher CSI survey score) for the SAR system (TG) than for the 'standard' system (CG1) or AR system (CG2)'.

For the experiments with Stimulo, the same designer operated the system in the SAR condition and the AR condition. This designer was asked to complete the CSI survey twice - once for each condition. The aim had been for the designer to complete the CSI survey immediately after the session. However, due to time constraints on the day this was not possible. Instead, the designer completed the survey the day after. No survey was completed for the 'standard' condition, as no ICT tools were used, other than a few slides presented using a large computer monitor and PowerPoint software.

For the experiments with Artefice, a different designer operated the system in the SAR condition and the AR condition. Each designer completed the survey once.

3.2.3. Methodology for the follow-up survey

The follow-up survey featured four questions that were intended to capture qualitative feedback from the designers about the performance of the session. The four questions were:

- What were all the things that went well during the session?
- How did the [SAR/AR/other] tool you were using contribute to the positive aspects of the session you have described above?
- What were all the things that were challenging about the session?
- How did the [SAR/AR/other] tool you were using contribute to the challenging aspects of the session you have described above?

The survey was sent to the designers via email after the session.

3.3. GESTURE INTERACTIONS ANALYSIS IN CO-CREATIVE SESSIONS

This section presents the research method used to capture and analyse the gestural, artefact-centric interactions within our experiments. We analysed gestures in a *restricted sense*, which, as Visser (2010) explained: "... is about movement of hands and arms that are accompanied by speech - even if it's not always at the same time - and that are co-expressive with this one". The core of our analysis is to describe the activity *being performed* by actors using artefacts (tangible, digital, mixed, ephemeral) and especially to report what kind of gesture is performed by whom in the three experimental conditions (Standard condition CG1, AR condition CG2, SAR condition TG). The expected results are quantitative and intended to answer the following questions: Are End-Users effectively involved in these co-design sessions? Which artefacts are mainly used by End-Users and Designers during interactions they lead?

To this end, we first define the type of gestural data we can collect. We present the evolution of the categorisation of the artefacts, through the development of a coding book, which was the basis of our coding to proceed to the analysis of the artefact-centric interactions.

Secondly, we introduce our methodological approach that is based on a real-time coding tool that allow to identify quickly what kind of interaction is performed by the actors. It was developed in order to reduce the necessary time required to perform the data analysis (usually realized with a traditional post-session coding method). In the following sub-sections we present: the methodology used to make this 'on the fly' coding; the 'post-session coding' results; then a comparison between results from these two kinds of coding methods. Finally, we proceed with the verification of the reliability and the fidelity of the methodology used before to conclude that we can replace the post-session coding by this 'on the fly' method for the quantitative analysis.

3.3.1. Co-creative sessions gesture indicators

Based on a review of the scientific literature about interactions with artefacts during design sessions, we have been able to propose a classification of different kinds of gestures, which was presented in D4.1.

This gesture classification has been refined to facilitate the coding of gestures, especially regarding to ephemeral gestures - which do not refer to the artefact (no tangible, digital or mixed gesture) but which are movements with arms and hands used support the speech (communication gesture) or to simulate an object or an action of this object (virtual artefact). We also added a new categorisation - 'None' - to capture an interaction that involves an actor but no artefact or gesture.

The final gesture classification includes five categories:

- Tangible - gesture in reference to a mock-up, a prototype or any physical material which is at the heart of the collaboration interaction;
- Digital - gesture in reference to an electronic media usually displayed on any surface such as TV screen, laptop, tablet, smartphone or any ICT tool which is at the heart of the collaboration interaction;
- Mixed - gesture in reference to a physical prototype on which are projected digital media. It's especially the case of the SAR system where actors are likely to interact with the mixed artefact or with the interface of the software that allows modifications of the digital mock-up;
- Ephemeral – which includes two sub-categories: communication gestures and virtual artefacts:
    - A communication gesture is a movement in the air made by end-users or designers which is associated to the speech and aims at supporting it;
    - A virtual artefact is an imaginary object that is depicted or mimicked by a gesture in the air as described in D1.2;
- None - an interaction that involves an actor but no artefact or gesture.

The aim of this categorisation was to enable us to code all kinds of gestures that we can observe; and to minimise the subjective judgement required by coders to accurately and consistently categorise the gestures observed. This aim led us to develop a coding book, which lists all the aforementioned categories of artefact-centric interactions along with an associated set of coding rules. An example of the coding rules is: "*Rule 2: When several people speak at the same time, we choose to code the end-user rather than the designer. Concretely, the designer begins to talk: we code Designer. In the same time the end-user speaks (and the designer keeps talking), we code End-user. When the end-user stops talking and the designer speaks (he has never stopped), we code Designer again. We chose to give advantage to end-users."*
This final version of the coding book has been created based on tests that involved the collection of a large number of interactions that featured different actors and artefacts as well as different coding approaches (coding on-the-fly vs. post-session coding).

3.3.2. Methodology and validation of the method

The method has been validated in two steps completed during the ‘early trial sessions’ described in D4.1. First a Cohen’s Kappa analysis of inter-rate reliability has been performed among the coders on the on-the-fly and post-session coding to ensure the validity of the coding. Second a comparative analysis of the on-the-fly and post session results have been carried out based on a classical Student’s t-test. Finally, we show the time saving of or method.

On-the-fly coding

In D4.1 we presented a first version of the software developed by the consortium to assist coders during the co-creative design session. This tool, named “Observer”, allows the researcher to identify gestural artefact-centric interactions occurrences in real time using a specific procedure whilst they are observing a co-creative design session. This first version of real-time coding method required some improvements.

In order to reduce the cognitive load of the coders, we decided to involve two coders working in parallel, with the first coder focusing only on the actors involved in the interaction, whilst the second coder focused only on the types of artefacts involved in the interaction. Five artefacts categories are considered: Tangible, Digital, Mixed, Ephemeral and None (None meaning that the interaction is not supported by any artefact). An example of the two interfaces is shown below.

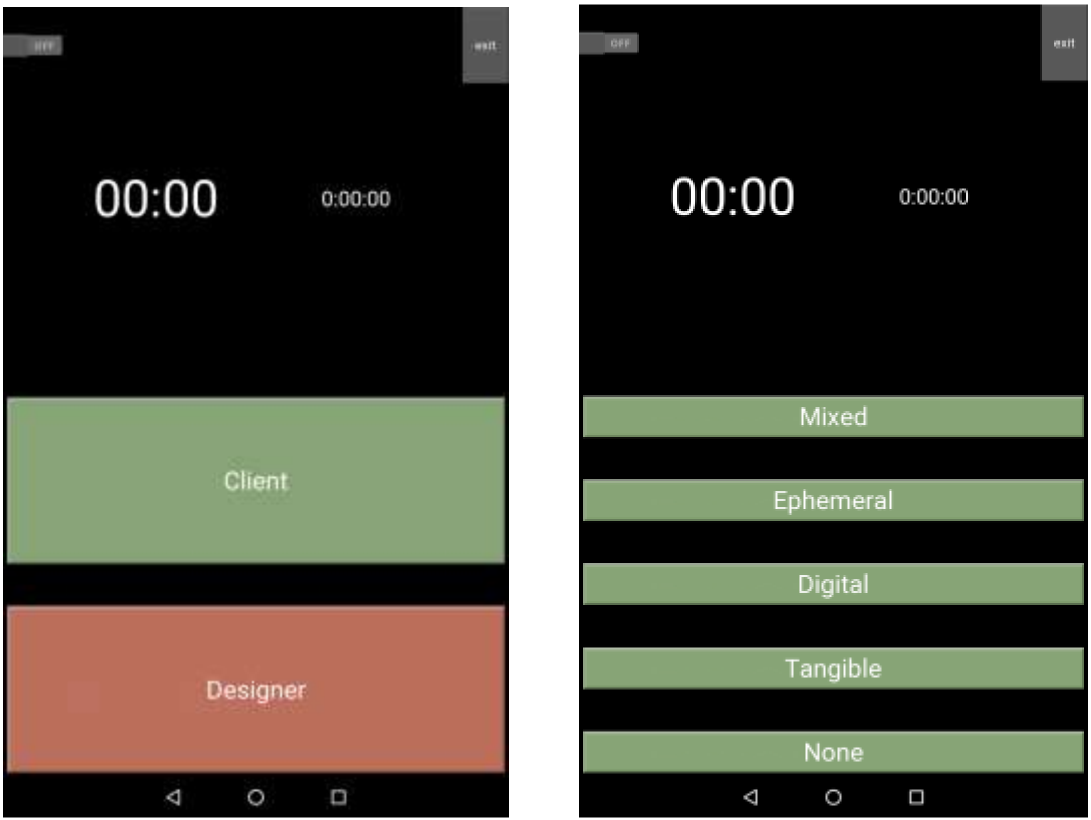


Figure 3: Observer interfaces used for coding the actors (left) and artefacts (right)

An automatic computation and a manual adjustment are necessary to obtain a usable set of data in the form of an Excel file that is used to make a comparison with the coding completed by the post-session coding method in order to validate our method.

Post session coding

To collect the quantitative data of gestures realized during the co-creative design sessions, we also proceeded to do a traditional post-session coding with two independent coders. The coders completed the coding of the entire Stimulo sessions whilst for the Artefice sessions we choose to randomly code two to three excerpts of each condition, which included 20-25 minutes of content in total.

Assessment of the robustness of the coding

Before the main WP4 experiments we carried out a usability test in order to verify that our coding on-the-fly was reliable. Specifically, the aim was to check that:

- it is possible (in a cognitive way) to code using this Observer tool (i.e. the coder is able to follow the stream of interactions and track down the events); and that,
- two encoders obtain comparable results (robustness) when coding the same session.

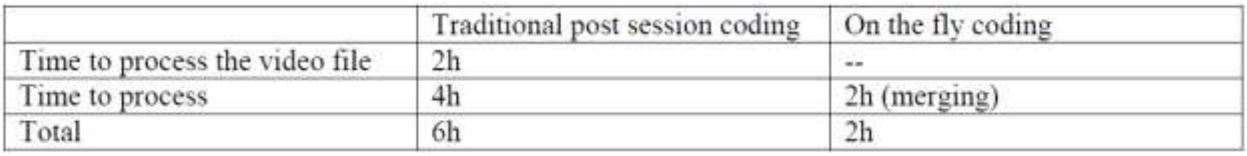
Table 5: Cohen's Kappa and percentage agreement of coding for each session analysed.

	Standard		SAR		AR	
ON THE FLY	Cohen's Kappa	% agreement	Cohen's Kappa	% agreement	Cohen's Kappa	% agreement
Actor	0,55	73	0,71	84	0,58	76
Artefact	0,62	74	0,59	71	0,30	49
POST SESSION	0,50	64	0,45	59	0,59	68

For this test, two researchers were invited to code, using the Observer tool, on different excerpts of previously captured sessions, which were available in video format. A first conclusion from this test was that Observer is easily usable for coding the actors. It is also usable for coding artefacts, although more effort is required because of the five types of artefact. A second conclusion was that two researchers coding three samples of different sessions lead to a fair convergence on the results obtained. We decided to compare these results using Cohen's Kappa index (see Table 5 above).

If we refer to the Cohen's Kappa scale (Landis and Koch, 1977), the results show substantial agreement (green), moderate agreement (yellow) and fair (red). We can see a reduced agreement level on the first excerpt of the AR session with the Artefact coding (0,30). We concluded that these results were satisfactory taking into account the radical improvement of processing efficiency and that the coding scheme and the Observer tool were sufficiently validated for use in the main experiments.

Table 6: Time taken to complete the coding of a 30 minutes section of session using the On-the-fly vs Traditional post-session coding approach.



	Traditional post session coding	On the fly coding
Time to process the video file	2h	--
Time to process	4h	2h (merging)
Total	6h	2h

While the On-the-fly method still requires some improvement, these results are provided in a much shorter time. The necessary time to do the coding of a 30 minutes episode of a session is approximately three time shorter for the On-the-fly approach compared to the traditional, post-session coding approach, including the post treatment of the on-the-fly data (Table 6).

3.4. SPOKEN INTERACTIONS ANALYSIS

As stated in Deliverable 4.1, part of the analysis of the experimental activities carried out in Tasks 4.3 and 4.4 focuses on spoken/verbal interactions in co-creative design sessions. This part of the analysis aims at extracting relevant information about the role technology plays in enabling or hindering creativity of co-designers when they gather to generate ideas.

The coding schemes to be used for such analysis consider three different facets that characterize the dialogues among different participants in the co-creative session. The first two characterize the content of the design solution in terms of, respectively, items and parameters (as shown in Table 7 and Table 8). These two coding schemes, on the one hand, provide elements of knowledge about the contents the SPARK platform should be capable of managing and processing. On the other hand, they also allow for the estimation of the degree of exploration of the design space. The third coding scheme focuses on the intentions of the co-designers participating in the creative session (Table 9). This helps when trying to distinguish between cognitive processes that mostly deal with idea generation from those that concern the evaluation of design proposals.

These coding schemes have been refined as detailed in Deliverable 4.1 (Sect 5.3.1) and they represent the latest update to the metrics initially developed for WP1.

Table 7: The refined coding scheme to map items in spoken interactions.

Text	What is expressed by words
Image	A computer generated picture (potentially vectorial)
Photograph	A photograph of a real object (non-vectorial image)
Logo	A graphic or textual representation of the brand identity
Icon	A graphic or textual representation related to marks of certification or similar
Background Motif	A texture or a set of elements characterizing the background of the design
System Parts	A reference to part of the entire system
Whole	A reference to the design proposal as a single entity

Table 8: The refined coding scheme to map parameters referring to items in spoken interactions.

Position	The parameter refers to the geographical location of the item on the design proposal or on one of its parts (i.e. [x,y] coordinates)
Orientation	The parameter refers to the degree of rotation of the item (i.e. θ coordinates)
Size	The parameter of the item describes refers to a change in its dimensions (without changing the aspect ratio)
Number	The parameter refers the quantity of items of the same kind to be added on or removed from the design proposal
Presence	The parameter refers to the introduction or the removal of an item within the design proposal
Colour	The parameter refers to the chromatic characteristics of an item
Reflectivity	The parameter refers to the capability of an item to show glossy or matt properties
Material	The parameter refers to the physical material that constitutes the item
Content	The parameter refers to the subject represented by the item
Shape	The parameter refers to the representational characteristics of an item (i.e. a change in the aspect ratio)
Sharpness	The parameter refers to the resolution and definition of the representation of the item

Table 9: The coding scheme to map the participants' intentions during the co-creative design session.

Analysis	A segment where the participant speaking is interpreting a design proposal, providing their judgement about it (positive or negative)
Synthesis	A segment where the participant speaking is proposing a solution (which can address the issues emerged -explicitly or not- during the analysis, opportunistic introduction of changes, or can be a complete new direction of development)
Choice	The selection between two or more alternatives to achieve a specific objective
Other	Other spoken interactions that do not pertain to one of the above codes for segments (e.g.: spoken interaction describing activities that occurred before the design session, instructions to deal with the ICT platform/tool interface…)

The coding schemes presented in Table 7, Table 8 and
Table 9 have been developed within Task 4.2, and they are an input for the activities of Task 4.3, 4.4 and 4.5 described in this deliverable. The planning of Tasks 4.3, 4.4 and 4.5 remains as was presented

in D4.1 and briefly reported in Section 3.1 of this document. The methodology for spoken interaction analysis is graphically displayed in Figure 4.

The spoken interactions occurred during the six co-creative design sessions have been video and audio recorded (individual microphones for speakers). This allows the team to carry out a typical design protocol analysis activity on the transcripts of the verbal utterances of co-designers. Design protocol analysis aims at extracting meaningful knowledge about thinking processes in design activities. A design session consists of a protocol of utterances and gestures (gaze…) which might include design moves (as actions/contents) that can be interpreted to unveil design thinking and creativity. From a viewpoint of data management and processing, design protocol analysis is about the transformation of qualitative data into an ordered dataset of nominal data (or differently, depending on the coding scheme) that allows for quantitative assessments. Design protocol analysis consists of several stages:

1. Recording the protocol;
2. Transcribing the protocol;
3. Segmenting the protocol;
4. Coding the protocol;
5. Analysis of the protocol.

These key activities of protocol analysis have been incorporated into the methodology presented in Figure 4. The recordings (first box of Figure 4) also provide an ex-post, privileged viewpoint on what happened during the co-creative sessions so that the data for the analysis can be reviewed (re-watched and re-listened) multiple times, ensuring an accurate transcription, segmentation and coding of the protocol. As the qualitative data available in recordings requires the coders to interpret its content, multiple interpretations by different coders are likely. Therefore, the reliability of the coding is critical for the meaningfulness of the results emerging out of the analysis.
Further details on the coding process and its validation are provided in Appendix II.

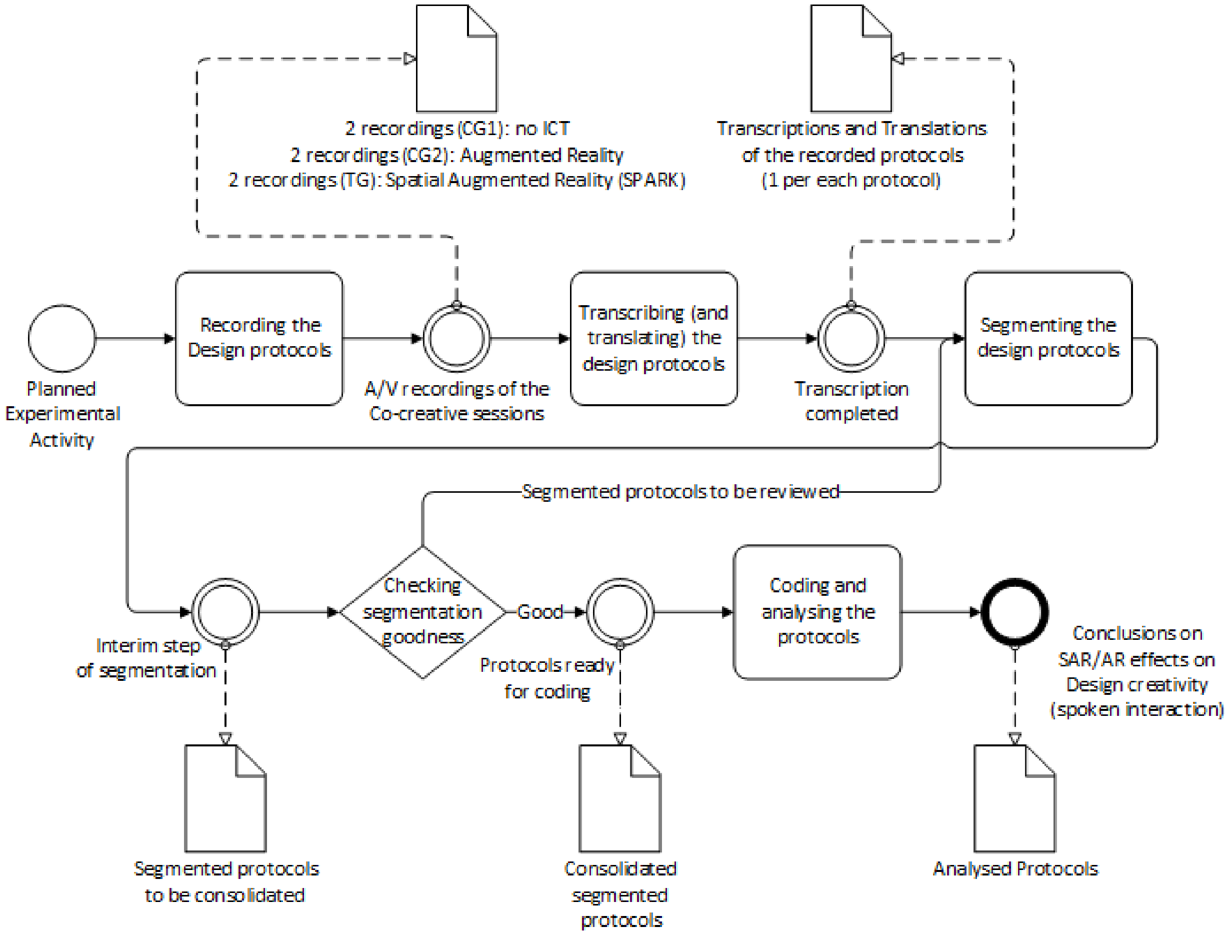

2 recordings (CG1): no ICT
2 recordings (CG2): Augmented Reality
2 recordings (TG): Spatial Augmented Reality (SPARK)
Transcriptions and Translations of the recorded protocols (1 per each protocol)
Planned Experimental Activity
Recording the Design protocols
A/V recordings of the Co-creative sessions
Transcribing (and translating) the design protocols
Transcription completed
Segmenting the design protocols
Segmented protocols to be reviewed
Interim step of segmentation
Checking segmentation goodness
Good
Protocols ready for coding
Coding and analysing the protocols
Conclusions on SAR/AR effects on Design creativity (spoken interaction)
Segmented protocols to be consolidated
Consolidated segmented protocols
Analysed Protocols

Figure 4: The methodology for spoken interaction analysis as activities and their output of data/information.

In order to explore the effects of the technologies (or their absence) during co-creative design sessions and carry out meaningful comparisons, the team working on the analysis of spoken interactions defined an initial set of assumptions based on a review of relevant literature.
The following bullets summarize the main assumptions behind the collaborative design session dynamics, without distinguishing between the conditions of being supported by ICT tools for design:

- A more fluent idea generation corresponds to a larger amount of design move/segments characterized by synthesis, regardless if they refer to a complete idea or part of it (Jones, 1984), (Torrance, 1972);
- A broader exploration of the design space corresponds to a larger set of ideas that are diverse from each other (Shah et al., 2003), (Boden, 2009);
- Within a more fluent production of ideas, the chances of finding an idea of good quality is higher. On the contrary, with low fluency, the chance of finding ideas of good quality is lower (Osborn, 1953). Independently from fluency, a good indication of quality of ideas depends on the matching between the problem space and the solution space as subsets of the design space (Dorst and Cross, 2001).
- The quality of interaction in a co-creative design session should be higher if all the participants actively participate in the session, by verbally expressing their thoughts, in the analysis of design concepts, in the synthesis of new ideas and changes to existing design, as well as in the choices between alternatives (elaborated upon the conclusions on knowledge externalization and combination as for Civi, 2000).

The above assumptions are also formulated so that it is possible to eventually run meaningful comparisons with other SPARK-related metrics, both with reference to the final outcomes of the design process (Section 3.2, first three bullet points) and the interaction in co-creative sessions which occurs in gestural interactions (Section 3.3, last bullet point).

According to the above assumptions, the following metrics, based on the above coding scheme, appear to be reasonably suitable for the investigation of design creativity along the design process:

- Creativity in terms of interaction represents the degree of involvement of participants in co-design session. To be measured as the number of shifts in spoken interaction among the different participants. This is to be measured both in terms of dialogue shifts between participants (regardless of their role) and as shifts between classes of participants (Designers and Clients).
- Creativity in terms of fluency represents the quantity of new ideas generated during a co-creative session. To be measured counting the number of shifts between two subsequent design moves (source: Analysis or Choice, target: Synthesis). It is assumed that subsequent "Synthesis" segments refer to the same idea.
- Creativity in terms of divergent thinking can be considered as a preliminary measure of the exploration of the design space. In general, it can refer to the diversity of the explored solutions and, more in general, design topics. To be measured as the ratio of the protocol duration and the overall amount to items/parameters couples explored in that time. Quality of solution, in principle, can be measured as the matching between the explored problem and solution spaces, as the combination of items and parameters explored in design moves of, respectively, analysis and synthesis.
- Creativity in terms of convergent thinking, in turn, corresponds to the capability to make selections among various alternatives in design (whole design proposals or part of them). In this document it is proposed to measure convergent thinking by means of indexes which make use of the amount of design moves of choice.

These metrics will be used for the analysis of the coded protocols having a sufficiently high level of reliability (Fleiss' K>0,6 as for Landis & Koch, 1977). The results of the analysis of spoken interactions are in Section 0 of this deliverable.

3.5. DESIGN PROCESS EFFICIENCY METRICS

Whilst the majority of the data capture and analysis activities in WP4 aims at understanding the impacts and benefits of using the SPARK platform within single co-creative design sessions, another important aspect is to understand the cumulative impact of the SPARK platform over multiple co-creative design sessions on the overall efficiency of the design process. To assess this aspect, a set of design process efficiency metrics were defined in deliverable D4.1 and are summarised in Table 10.

Table 10: Summary of the design process efficiency metrics
(red colour refers to modifications applied after testing as detailed in Appendix III).

Metric title	Definition
Person-hours spent on project	All hours spent on project by design agency (including unbilled hours)
Lead time	Number of days between project start date and end of the Ideas Development phase
Total development cost	Direct costs incurred by design agency (Only up to End of layout - Ignoring post-production costs)
Cost of prototype production	Cost of preparing all design representations used in collaborative sessions or sent to end-user (materials and labour)
Re-work iterations	Total number of times that a design activity has to be completed again (as requested by management). Total number of all major design meetings within each design stage Total number of co-creative design sessions completed within each design stage (and no. of those of which are conducted with the SPARK platform) Number of versions released to the end-user, feedback received and acted upon

Within the timescale of WP4 the objective was to apply these metrics to a number of historical case studies provided by the industrial partners. These historical case studies were completed without the aid of the SPARK platform.
The main application of the design process efficiency metrics will be in WP5, when the industry partners will have the SPARK platform installed at their own premises and will therefore be able to use the SPARK platform for several sessions over the course of a real project.
Therefore, during WP5, new projects will be conducted by the industrial partners that will make use of the SPARK platform within several co-creative design sessions over the course of the project. The design process efficiency metrics will then be applied to these new projects to generate longitudinal case studies, which can then be compared with the benchmark data from the historical case studies. During WP5, the choice of projects for the new longitudinal case studies will be limited by: the range of projects that are being worked on by the industrial partners during the WP5 timescales; the suitability of the product for representation using SAR technology (e.g. not too big, not too small etc.); and the willingness of the end-user to participate in several co-creative design sessions using the SPARK platform. To maximise the probability that the projects selected for the new longitudinal case studies will be comparable in nature to the historical case studies collected, a range of representative projects from the industrial partners will be selected for analysis

Within WP4, the primary objectives for the application of the design process efficiency metrics were:

- To demonstrate that the design process efficiency metrics are practical to apply (the required data is available from the industry partners and can be gathered without excessive difficulty/effort);
- To apply the design process efficiency metrics to a variety of historical case studies (that had not used the SPARK platform) to provide some benchmark data for comparisons with the longitudinal case studies that will be completed during WP5.

The metrics were applied to three historical case studies at Artefice. The case studies were selected in order to be representative of typical projects completed by Artefice. Three projects were considered sufficient to provide a representative range for the type of project that would be completed with SPARK i.e. packaging for a new product where the brand strategy and brand identity have already been defined. All three projects involve products from the same end-user, who we will refer to as 'Food Inc.' to maintain the anonymity of the company. Food Inc. is a regular customer of Artefice and the company has expressed an interest in participating in the longitudinal case studies to be completed in WP5. If Food Inc. goes on to participate in the WP5 longitudinal studies, it will further enhance the comparability of the historical case studies (without SPARK) and the WP5 longitudinal case studies (with SPARK).

To collect the historical case study data a UBAH researcher visited Artefice's offices where they worked with Artefice staff to build a project timeline for each of the case studies. The timeline included details including:

- The scope of the project;
- The dates of the key project milestones (initial brief, stage-gate reviews, completion);
- The co-creative sessions completed;
- The number and type of design representations created to gather feedback from the end-user;
- The number of design iterations.

Additional data was collected concerning the person-hours spent on the project by Artefice staff, with breakdown by iteration and by type of staff (designer vs commercial manager) and other project costs. The combination of the project timeline and these supplementary financial data enabled the application of the design process efficiency metrics shown in Table 10. Examples of the project timelines collected are presented in Section 4.5.

No historical case studies were collected from Stimulo within the WP4 timeframe. The UBAH research team have presented the work completed with Artefice on the design process efficiency metrics to the Stimulo team and it was concluded that: there would be no problem in obtaining the data required to apply the metrics (first objective); and that it would be efficient to gather the historical case study data (second objective) as part of the activities to be completed at the Stimulo premises during WP5.

3.6. TECHNOLOGY FEEDBACK DISCUSSIONS

A lot of thought and care was taken during the planning of the experiments to ensure that the experimental situation for each of the sessions was as close as possible to a 'real life' co-creative session. The experiments were completed by professional designers working on real projects with appropriate end-users. This helped to ensure that the evaluation of the SPARK technology was realistic, but also offered the opportunity to gather very high-quality feedback on the technology from the participants. To make the most of this opportunity, technology feedback discussions were organised and held immediately after the sessions.

These discussions involved the designers that had participated in the session and the research team. There was no fixed structure to the sessions, but in general the designers provided a brief overview of how the session had gone from their perspective, then began to mention some of the specific technology-related aspects that had been successful or problematic. This then led to discussion between the designers and the technology development members of the research team as to how the technology could be enhanced to overcome the challenges encountered and build on the successful aspects of the technology.

Whilst these technology feedback discussions did not directly contribute to the objectives of this work package, they did provide excellent feedback on the technology and allowed the members of the research team involved in the technology development aspects to engage in very productive 'requirements discovery' discussions. The insights and requirements identified from these discussions are presented in Appendix III.

4. RESULTS AND DISCUSSION

4.1. OVERVIEW OF THE EXPERIMENTAL SESSIONS

The following provides a brief overview of each of the experimental sessions. To aim is to provide some context to the subsequent detailed analysis. The reports are based on the research teams' observations of the sessions as well as notes taken during informal feedback sessions held with the designers immediately after the sessions.

4.1.1. Summary of Stimulo sessions hosted at GINP

The SAR condition was conducted first. The session involved three end-users and was led by one designer (XM), whilst the second designer (JC) focused on operating the SAR user interface on a tablet PC. The participants were introduced to the designers and given a very brief overview of the SAR technology before the session began. An introduction to the session was provided by XM, which featured some background on the product, the objectives of the session, and the initial design proposals.

The end-users were interested in the technology but managed to stay focused on the task at hand and responded in a realistic and motivated manner to the questions and requests for ideas from the designer. JC was very quick in learning how to use the tablet PC user interface, but still required some help from the research team on a few occasions.

On three occasions during the session, the SAR user interface stopped responding, resulting in a delay of several minutes each time whilst the system was reset. Despite these technical difficulties, the team were able to generate several new design proposals that they were pleased with.
After the session was formally closed, the end-users were given the opportunity to ask further questions about the technology and play with it.

During the Technology Feedback Discussion, a number of observations were made by the designers and the research team. The designers noted that they were disappointed by the technical difficulties experienced during the session. They also noted that there were significant differences in the colours seen on the mixed prototype compared to the colours represented on the tablet PC user interface. The researchers observed that neither the designers nor the end-users had picked up or handled the mixed prototype to any significant extent. The designers explained that this was because they were worried that handling the prototype would cause further technology problems.

The AR session also involved three end-users and two designers and was conducted in a very similar manner. There were no significant technical problems in this session and the team were once again able to generate several new design proposals that they were pleased with. In the informal discussion with the designers after the session they noted that there were similar problems in relation to colour differences between the AR prototype and the colours shown on the tablet PC user interface. However, they did feel that the session had gone better because the end-users were able to move around more freely (without worrying about disrupting the projection) and the technology seemed more reliable. The researchers noted that, again, there was limited handling of the prototype during the session, although the designers commented that it becomes more difficult to handle the prototype whilst also holding on to the tablet to view the prototype. In the Technology Feedback Discussion, the designers discussed a number of new features that they felt would be useful for both the AR and SAR versions of the SPARK technology.

The session using standard types of design representation was conducted last and involved one designer and three end-users. The start of the session was similar to the previous two sessions. The designer used a combination of 3D printed components, Pantone colour swatches and a PowerPoint presentation on a large screen to present the pre-defined concepts and support the discussion. No significant challenges were encountered and the team was able to generate a number of new concepts.

Key points from Stimulo sessions:

- Enthusiastic participants who were interested in the technology;
- Technical difficulties caused significant disruption to the SAR session;
- Some significant differences identified between the colours displayed on the AR/SAR prototype and colours shown on the tablet PC user interface;
- Limited handling of the prototype during both the AR and SAR sessions;
- Good, detailed feedback from the designers concerning new technology features they require.

4.1.2. Summary of Artefice sessions hosted at POLIMI

For the Artefice sessions, each session involved a different pair of designers and different end-users, whilst the design task remained the same – as described in Section 2. The AR session was completed first, with input from two end-users. As with the Stimulo sessions, the designers were trained on how to use the SPARK tablet PC user interface before the session and the end-users were given a very brief introduction to the technology before the formal start of the session. The designers provided some background on the product and the objectives of the session at the start. Three pre-prepared concepts were introduced to the end-users and one was selected as the starting point for the design activity. After around 25 minutes, the team moved on to modifying the second pre-prepared concept, and this pattern was repeated for the third pre-prepared concept. By the end of the session the team had generated a variety of ideas that they were satisfied with.

Once again, a Technology Feedback Discussion was conducted involving the researchers and the designers. From this discussion, it was noted by the researchers that there had, again, been very little handling of the physical prototype. The designers noted the same problem with differences in colour rendering between the AR prototype and the representation in the tablet PC user interface.
The SAR session was completed next. For this session, extra time was spent before the formal start of the session to demonstrate to the end-users that the SAR prototype could be picked up and handled whilst maintaining the visualisation. This worked to some extent as at least one of the end-users (the one sitting closer to the mixed prototype) handled the prototype during the main session.

Early on in the session, one of the end-users expressed frustration that it was not possible to visualise changes to the geometry of the prototype, change the colour of graphical elements, nor incorporate graphical elements that had not been pre-loaded onto the Information System. This end-user suggested that these limitations to the scope of the session were a significant creative hindrance and limited the value of the technology. Actually, these limitations were already known by the SPARK development team and in fact the introduction of new assets is a feature available in the second release of the SPARK platform; however, they clearly affected the outcome of the test session.

Another problem encountered was that only the end-user sat closest to the prototype that could easily reach and handle the SAR prototype. This was mainly due to the set-up of the SAR system, as the 'projection volume' (the volume of space in which the projected content appears to be sharply focused) is relatively small and had been centred at one end of the rectangular table. This meant that the end-user furthest away could not easily reach the prototype, and even if they did pick it up and try to bring it closer to himself, the projected content would appear out of focus on the prototype once it was moved out of the projection volume. In the Technology Feedback Discussion with the designers after the SAR session, a detailed list of challenges they had encountered during the preparation of the session was discussed. The designers did also note that they saw significant potential for SAR technology in their work, as they could see how it could significantly reduce unproductive design iterations.

Finally, for the standard session the designers had prepared a variety of graphical elements in the form of stickers, which could be placed on the blank prototype to facilitate the discussion concerning the composition of the packaging and the positioning of the elements. They also worked directly on the blank prototype with coloured pens and pencils, to provide some background colour and to add text elements that had not been pre-prepared. There was good engagement from the end-users throughout the session and a number of new ideas were generated.

Key points from Artefice sessions:

- Some problems with the Information System connection resulted in incomplete session logs;
- A variety of limitations of the Information System identified by the designers and requests for new features generated;
- Some significant differences identified between the colours displayed on the AR/SAR prototype and colours shown on the tablet PC user interface;
- Some end-users frustrated by the limitations of the system;
- Designers can see potential of technology to significantly reduce unproductive design iterations;
- The physical arrangement of the participants around the table, and their position relative to the projection volume is an important consideration for SAR technology.

4.2. CO-CREATIVE SESSION PERFORMANCE METRICS

The following sub-sections provide details of the results of the application of the co-creative session performance metrics, the usability survey, and the follow-up survey sent to the designers.

4.2.1. Results of the co-creative session performance metrics

Hypothesis
Co-creative design sessions completed with SAR-based design representations (TG) will result in improved idea generation (quality, variety, quality and novelty), task progress and filtering effectiveness compared to similar sessions completed with standard design representations (CG1) or AR-based design representations (CG2)

Findings

- Stimulo SAR sessions resulted in improved idea generation (quantity, variety, quality and novelty) and task progress compared to sessions with standard design representations, but differences were not significant in some cases and were not better than AR sessions.
- Artefice SAR sessions resulted in improved idea generation in terms of quantity and quality of ideas compared to sessions with standard design representations, but was worse or the same in terms of variety and novelty of ideas, task progress and filtering effectiveness.

Conclusions
Overall, the SAR technology has shown good potential from the results of the co-creative performance metrics, but has not consistently outperformed the other conditions across all metrics and both companies.

Table 11 provides a summary of the session performance metrics results for all sessions. For every metric, a higher score indicates better performance.

Table 11: Summary of co-creative session performance metrics results.

	Stimulo			**Artefice**		
Metric title	**SAR**	**AR**	**Standard**	**SAR**	**AR**	**Standard**
Quantity of ideas	8	8	6	11	4	5
Variety of ideas	Original = 5 New = 1	Original = 1 New = 1	Original=4 New=1	Original = 2 New = 0	Original = 4 New = 0	Original = 5 New = 1
Quality of ideas	4	5	1	3	1	2
Novelty of ideas	=44/8= 5.5	=51/8= 6.4	=23/6= 3.83	=7/3= 2.3	=9/4= 2.3	=19/5= 3.8
Task Progress	1xHigh = 3 Total = 3	2xHigh = 6 1xMed = 2 Total = 8	1xMed = 2 Total = 2	2xHigh =6 Total=6	1xHigh =3 Total = 3	1xHigh = 3 1xMed = 2 1xLow = 1 Total = 6
Filtering Effectiveness	= 4/(8-1)= 0.57	=3/(8-5)= 1	=5/(6-1)= 1	=8/(11-1)= 0.8	=3/(4-2)= 1.5	=3/(5-3)= 1.5

For the Stimulo sessions, the SAR and AR condition performed best or joint best against the idea generation metrics (quantity, variety, quality, and novelty of ideas), task progress and filtering effectiveness metrics, with the AR condition generally offering the best performance. The novelty and quality of ideas stand out as the particular successes for the SAR and AR conditions over the standard condition at Stimulo. For the Artefice sessions, the SAR condition performed significantly better than the other two conditions in terms of the quantity of ideas metric. The SAR condition was also best or joint best on the quality of ideas metric and the task progress metric. However, the standard condition performed best or joint best in terms of the variety, novelty, task progress and filtering effectiveness metrics, but was not so good on the quantity and quality metrics. Overall, the SAR technology has shown good potential from the results of the co-creative performance metrics, but has not consistently outperformed the other conditions across all metrics and both companies.

4.2.2. Results of the usability assessment (CSI survey)

Hypothesis

The designers' rating of system usability will be better (higher CSI survey score) for the SAR system (TG) than for the 'standard' system (CG1) or AR system (CG2)

Findings

- Overall system usability was rated significantly better for the SAR system than the 'standard' system but was slightly worse than the AR system.
- Most important aspects of usability were 'Collaboration', 'Exploration' and 'Immersion'.
- SAR system scored poorly on 'Immersion' aspect

Conclusions

Designers were reasonably satisfied with the usability of the SAR system, but there is still room for improvement. Poor score for 'Immersion' aspect probably due to the technical difficulties encountered during the session.

Cherry & Latulipe (2014) state that the task, the tool, and the expertise/experience of the user are the three main variables that might affect score obtained when completing the CSI survey. Hence, it was important to only change one variable at a time (in this case, the tool). In the Artefice experiments, the designers were selected based on having similar level of general design experience and were all novice users of the AR/SAR user interface. In the Stimulo experiments, the same designer manipulated the user interface in both the AR and SAR activities, which enables a good comparison of those scores.

Table 12 presents the results of the CSI assessment. The average score is presented by condition and with the breakdown by CSI factor. The maximum score for each cell in the 'average score' columns is 10. The overall CSI score is provided in the final row and is the average from across the Stimulo and Artefice designers that participated in the same condition. A weighting factor is applied to scale the CSI score to a maximum of 100. Cherry and Latulipe (2014) suggest that scores can be interpreted similar to a typical educational grading scheme, where a score of 90 or above is an 'A' grade and implies that the tool supports the specific creativity task extremely well. Conversely, a score below 50 is an 'F' grade and suggests that the tool does not support creative work very well. On this basis, both the SAR and AR performed reasonably well (around a 'C' grade), whereas the conventional condition was considered poor ('F' grade).

When examining the results in more detail it is useful to consider the 'average count' column, which can be interpreted as a rating of the relative importance of each of the CSI factors. The maximum score is five. The 'collaboration' factor had the highest average count (4.1), followed by 'exploration' (3.7) and 'immersion' (3.0). The AR condition performs consistently well against these three most important factors. The SAR condition scored well on the ‘collaboration’ and ‘exploration’ factors but very poorly on the ‘immersion’ factor – the cause of this poor immersion score was probably the technical difficulties that interrupted the flow of the session, as discussed further in the following section.

Table 12: Results of the Creativity Support Index assessment.

	Average Score			Average count (all conditions)
Aspect	**SAR**	**AR**	**Standard**	
Collaboration	7.5	8.3	5.8	4.1
Exploration	7.0	7.3	4.0	3.7
Immersion	2.8	7.5	4.3	3.0
Expressiveness	4.5	5.0	4.5	2.0
Results worth the effort	7.8	6.3	5.3	1.3
Enjoyment	9.0	7.5	5.0	0.9
Overall CSI score	**67.2**	**74.7**	**49.0**	

4.2.3. Results of the follow-up survey

Findings

- Designers appreciated capability of the SAR technology to enhance collaboration with the end-users and the ability to make quick, real-time modifications on the prototype.
- SAR technology was useful for quickly eliminating poor ideas.
- Several technical challenges that require further attention, including poor reliability and differences in colour rendering between end-users view and designer’s view on tablet PC user interface.

Conclusions

Designers can see the potential value of SAR technology to support co-creative design sessions, but some significant technical challenges and limitations of the current prototype technology need to be addressed.

Table 13 provides a summary of the key points from the designer follow-up survey. The points were summarised by the researcher from the original responses provided by the designers.

Table 13: Summary of the follow-up survey responses.

Question	Condition		
	SAR	**AR**	**Standard**
What went well?	*Artefice:* Helps with detailed refinements. Helps to quickly rule out poor suggestions from end-users. *Stimulo:* Freedom to try many different ideas.	*Artefice:* Improved interaction with end-users. *Stimulo:* Good interaction/ communication with end-users.	*Artefice:* Good empathy with the end-users, who were willing and able to provide good input. *Stimulo:* End-users were positive and focused.
How did tool contribute to positive aspects?	*Artefice:* Quick, real-time modification of a tangible prototype facilitates co-creation. *Stimulo*: Able to generate and test some new ideas for colours and logo position.	*Artefice:* Real-time modification improved interaction with the end-users. *Stimulo:* Intuitive sharing of ideas between end-users and designer allowed quick iteration of concept	*Artefice:* Intuitive interaction method that enabled the end-users to participate in an uninhibited manner *Stimulo:* It was a basic way to support the engagement between the end-users and the designer.
What was challenging?	*Artefice:* Some end-users were frustrated by the limitations of the system. *Stimulo:* Technical problems had an impact. Limited interaction with the prototype. Major differences between the designer's view (tablet) and end-users' view (SAR).	*Artefice:* Various technical limitations and failures hindered and disrupted the session. *Stimulo:* More chaotic, less focused session with more random/trial and error - requires more pro-active facilitation.	*Artefice:* Limited range of elements and hand drawn elements limits the quality/realism/ fidelity of the final outcome. *Stimulo:* LED position options difficult to represent.
How did tool contribute to challenging aspects?	*Artefice:* Inability to add new assets during the session or modify standard elements limited the freedom of the session. SAR model lacks realism. *Stimulo:* 3D effect and on-the-fly changes within SAR is useful but need more simulation/user interaction features.	*Artefice:* Technical problems caused disruption. Limitations of the system discouraged the end-users. *Stimulo:* Tablet create a barrier to direct interaction with the real prototype.	*Artefice:* Limited range of pre-prepared elements meant that the final outcome did not entirely represent what was desired/discussed. *Stimulo:* Designer had to build a fourth concept (mixing elements of the three pre-prepared concepts).

The main perceived strengths of both the AR and SAR conditions was the enhanced collaboration with the end-users and the ability to make quick, real-time modifications to the prototype to try out new ideas. A notable example of this came in the SAR condition session with Artefice, in which the end-user made a proposal concerning the position of a logo. The designer was confident that this proposal would not enhance the design. In the follow-up survey she noted that with the SAR technology it was quick and simple to modify the prototype according to the proposal, show it to the end-user and get them to agree that it was not a good proposal before reverting to the original logo position. This type of idea elimination activity can be particularly helpful in co-creative design sessions in which the end-users are staff members from the manufacturer of the product, who might otherwise request further work to be completed on an idea before it is eventually rejected.

It terms of weaknesses, the lack of reliability of the AR and SAR systems was a problem for both companies, with designers commenting that the technical difficulties that occurred during the sessions disrupted the natural flow of the session, causing frustration. Some specific technical limitations were also identified. From Artefice's perspective, the SAR prototype lacked realism because it was too bright/reflective. For them, the AR condition was preferred as the quality and fidelity of the visualisation was considered to be better. From Stimulo's perspective, a key limitation was the significant differences in colour hue and shade that were observed when viewing the design representation on the designer's user interface compared to the view seen by the other participants when looking either at the SAR prototype or the AR prototype through the tablet. Given that selecting colours, materials and finishes was a key objective of the Stimulo sessions, this was a major problem.

4.3. GESTURE INTERACTIONS IN CO-CREATIVE SESSIONS

This section reports on the results from the gesture analysis and presents the discussion around the hypothesis tested. The quantitative analysis was based on the results obtained with the on-the-fly coding method, whilst the qualitative analysis was based on the post-session coding of the protocols.

4.3.1. Participation of end-users in the co-design sessions

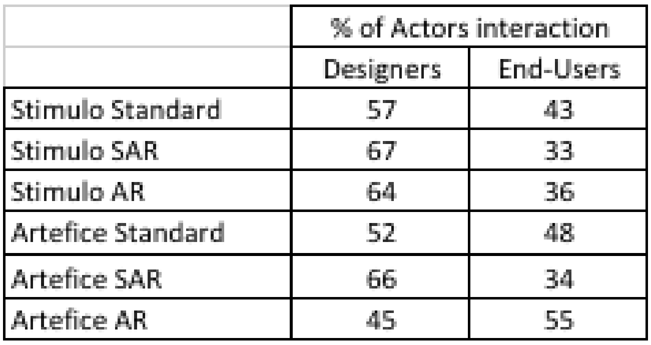
Hypothesis 1
End users' interactions increase from Standard to AR and SAR conditions.

Evidence
We computed the percentage of interactions in each condition for the two experiments.

Conclusions
The Spatial Augmented Reality condition does not favour the end-users' interactions in either the Stimulo or Artefice cases. However, we observe a comparable percentage of interactions which is encouraging for the SPARK project as it tends to demonstrate the suitability of the platform in a variety of co-creative design situations.

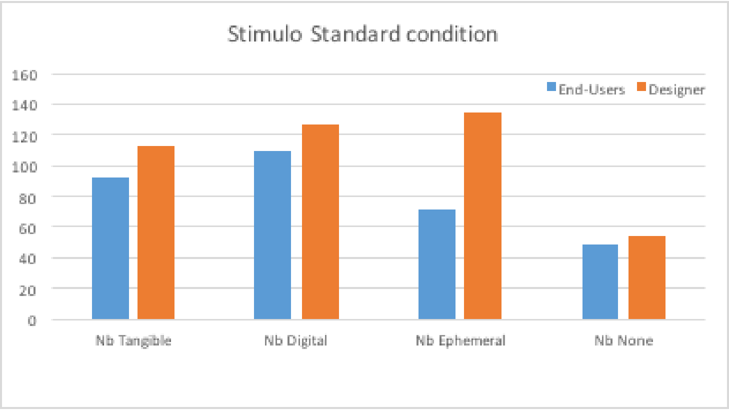
We compare in Table 14 the percentages of interactions initiated by Designers and End-users. We observe that designers perform a higher percentage of the interactions in most sessions. However, end-users initiate a significant proportion of interactions in all the sessions (between 33% and 48%). This tends to demonstrate that the sessions are real co-design sessions and that our experimental protocol is valid. Unfortunately, the Spatial Augmented Reality condition did not exhibit a higher percentage of interactions initiated by end-users. In fact, we observed a slightly lower proportion of end-users' interaction in the SAR condition, which is the reverse of the result expected. However, it should be kept in mind that we experienced technical problems for SAR condition, particularly in Grenoble. The results obtained are approximately the same in the Stimulo and Artefice sessions for the SAR and Standard conditions, which is not the case for the AR condition. The AR condition presents a significant discrepancy with regard to the other conditions and the Artefice session presents an opposite trend. Further experiments are necessary to investigate this point. The fact that we observe a comparable percentage of interaction with two very different sessions in terms of products and process is encouraging for the SPARK project as it tends to demonstrate the suitability of the platform in a variety of co-creative design situations. Despite some encouraging results, we cannot validate Hypothesis 1 with the available data.

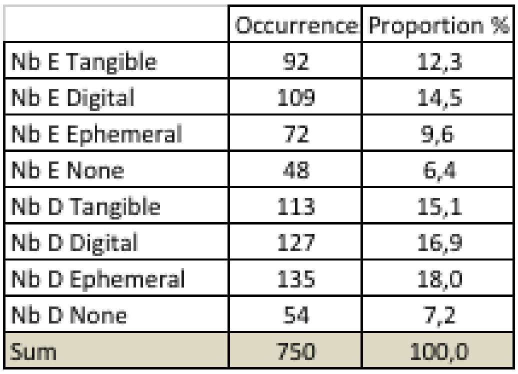
Table 14: Percentages of actors' interactions.



	% of Actors interaction	
	Designers	End-Users
Stimulo Standard	57	43
Stimulo SAR	67	33
Stimulo AR	64	36
Artefice Standard	52	48
Artefice SAR	66	34
Artefice AR	45	55

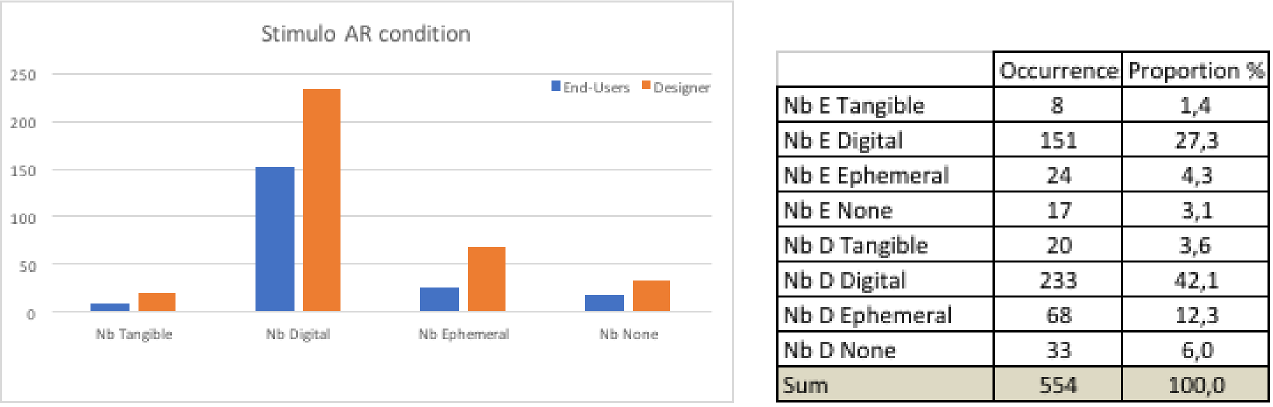
Regarding gestural artefact-centric interactions, we computed the proportions of each interaction type. The following results are based on the on-the-fly coding results. In order to facilitate the reading and the interpretation of the results, all the tables have been translated into graphs and presented together with the tables (Figures 5-10).





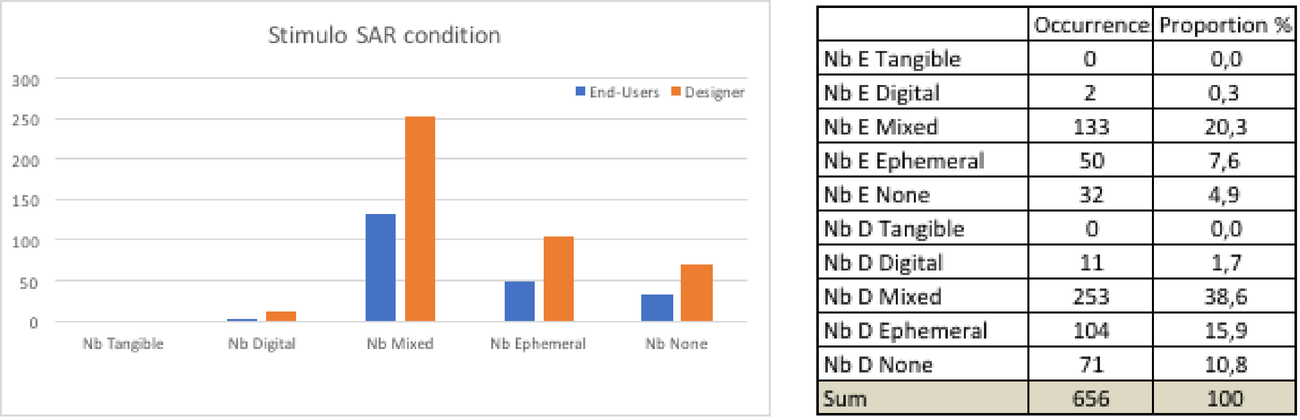
	Occurrence	Proportion %
Nb E Tangible	92	12,3
Nb E Digital	109	14,5
Nb E Ephemeral	72	9,6
Nb E None	48	6,4
Nb D Tangible	113	15,1
Nb D Digital	127	16,9
Nb D Ephemeral	135	18,0
Nb D None	54	7,2
Sum	750	100,0

Figure 5: Artefact-centric division for Stimulo standard condition.



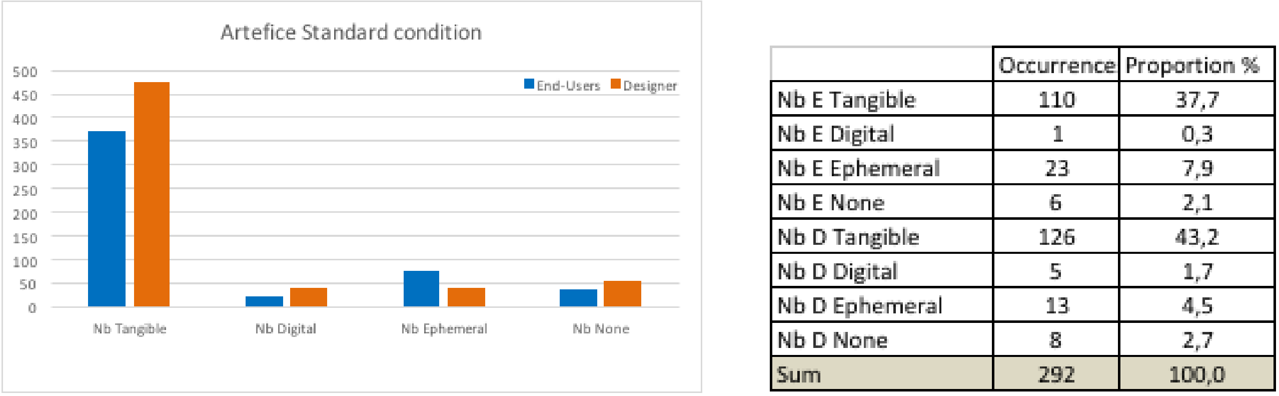
	Occurrence	Proportion %
Nb E Tangible	8	1,4
Nb E Digital	151	27,3
Nb E Ephemeral	24	4,3
Nb E None	17	3,1
Nb D Tangible	20	3,6
Nb D Digital	233	42,1
Nb D Ephemeral	68	12,3
Nb D None	33	6,0
Sum	554	100,0

Figure 6: Artefact-centric division for Stimulo AR condition.



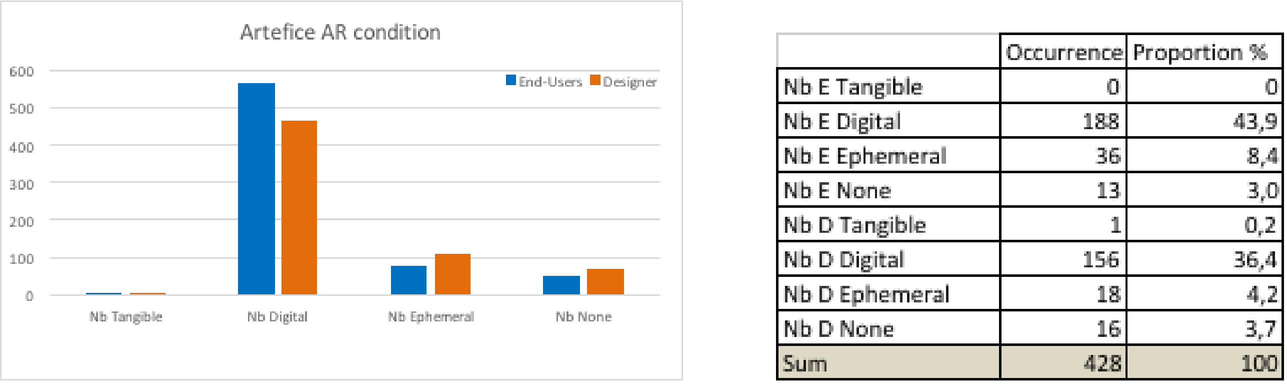
	Occurrence	Proportion %
Nb E Tangible	0	0,0
Nb E Digital	2	0,3
Nb E Mixed	133	20,3
Nb E Ephemeral	50	7,6
Nb E None	32	4,9
Nb D Tangible	0	0,0
Nb D Digital	11	1,7
Nb D Mixed	253	38,6
Nb D Ephemeral	104	15,9
Nb D None	71	10,8
Sum	656	100

Figure 7: Artefact-centric division for Stimulo SAR condition.



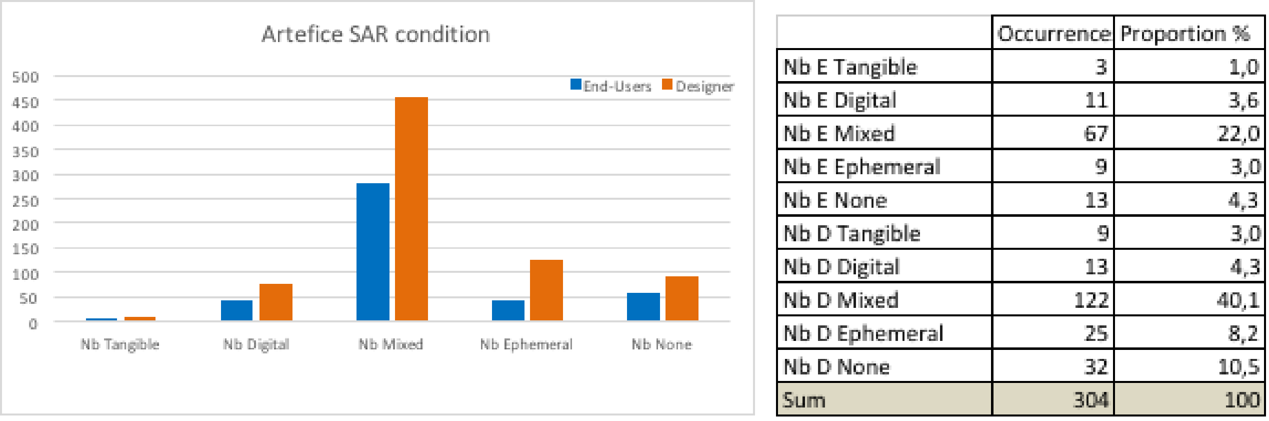
	Occurrence	Proportion %
Nb E Tangible	110	37,7
Nb E Digital	1	0,3
Nb E Ephemeral	23	7,9
Nb E None	6	2,1
Nb D Tangible	126	43,2
Nb D Digital	5	1,7
Nb D Ephemeral	13	4,5
Nb D None	8	2,7
Sum	292	100,0

Figure 8: Artefact-centric division for Artefice standard condition.



	Occurrence	Proportion %
Nb E Tangible	0	0
Nb E Digital	188	43,9
Nb E Ephemeral	36	8,4
Nb E None	13	3,0
Nb D Tangible	1	0,2
Nb D Digital	156	36,4
Nb D Ephemeral	18	4,2
Nb D None	16	3,7
Sum	428	100

Figure 9: Artefact-centric division for Artefice AR condition.



	Occurrence	Proportion %
Nb E Tangible	3	1,0
Nb E Digital	11	3,6
Nb E Mixed	67	22,0
Nb E Ephemeral	9	3,0
Nb E None	13	4,3
Nb D Tangible	9	3,0
Nb D Digital	13	4,3
Nb D Mixed	122	40,1
Nb D Ephemeral	25	8,2
Nb D None	32	10,5
Sum	304	100

Figure 10: Artefact-centric division for Artefice SAR condition.

In each of the six graphs we observe that the participants use the type of artefacts provided by experimental conditions (i.e. Digital for AR condition, Mixed for SAR condition and Tangible for Standard condition) except in

Figure 5 for the Stimulo Standard condition where Digital, Tangible and Ephemeral artefacts are used in the same proportion. This can be explained by the fact that within the Stimulo Standard condition a large monitor was used extensively to display design elements. As for hypothesis 1, this remark tends to reinforce the validity of our experimental setting by showing that the participants actually performed their task with the material they had and therefore what we measure is the impact of the technology and the physical setting. In the Spatial Augmented Reality condition, participants used mixed artefacts for between 20,3 and 40,1% of the interactions.

4.3.2. Evolution of virtual artefacts in co-design sessions

Hypothesis 2
The number of virtual artefacts referred to by participants will be lower in the SAR condition than in the AR and Standard conditions.

Evidence
Ephemeral gestures are considered as an indicator of the presence of virtual artefacts.

Conclusions
The profile of the graphs suggests that there is an effect of the technology on the co-creative design sessions.

Our protocol with on-the-fly-coding only allowed us to capture what we call ‘ephemeral artefacts’, which are types of gestures made by the participants during the experiments, specifically ‘virtual artefact’ gestures and ‘communication’ gestures. Virtual artefact gestures have been defined in D1.2 and can be summarised as a gesture simulating a function or the use of the product by the means of gestures in the air. Communication gestures are gestures that reinforce the speech and do not add more information. In our analysis, it was not possible to distinguish between the two types of gesture without referring to the analysis of the speech. However, in a first approximation we can consider that ephemeral gestures can be considered as a trend of virtual artefacts, even if we are unable to distinguish in which proposition they are present inside the ephemeral category. Therefore, if we look at the ephemeral gestures, we note that they are more represented in the Stimulo sessions than in the Artefice sessions (respectively 15,4 % against 7,4%). Further investigations regarding the structure and contents of the session may help to understand this difference. We may notice here an effect of the protocol itself.

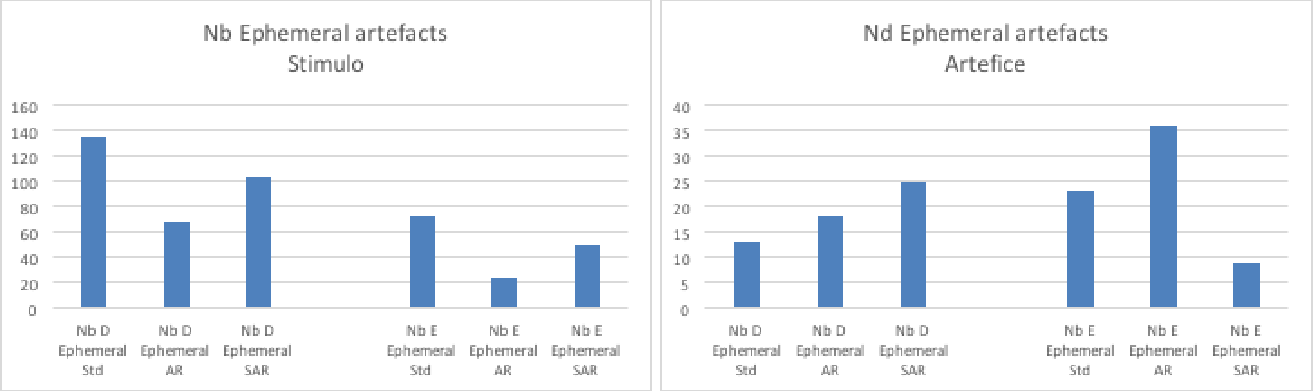


Figure 11: Occurrences of ephemeral gestures in the two conditions (D for Designer, E for End-users)

If we consider Figure 11, we can notice a trend of reduction of ephemeral gestures in SAR conditions, except for Artefice Designers, but without a clear explanation. Unfortunately, there is no clear evidence of the reduction of virtual artefacts in SAR sessions, as expected in our hypothesis. Besides, it is difficult to determine the cause of the variations. The technological factor cannot be considered as pre-eminent here, or at least, it is part of other factors such as the type of management of the sessions, the profile of the participants, etc. However, the profile of the graphs suggests that there is an effect of the technology. Except for Artefice SAR condition the trend is symmetric for the two experiments suggesting that there is the same effect of the condition on the designers and the end-users. However, without further research concerning the indicators allowing us to distinguish a virtual artefact from a gesture of communication we cannot conclude more on this point.

4.3.3. Level of interaction in the various experimental conditions

Hypothesis 3
There should be more interactions in the SAR condition than in the other conditions.

Evidence
We observed the number of interactions per minute in the different conditions.

Conclusions
We notice a relatively stable pace of interaction in all the sessions regardless the conditions. This tends to suggest that there is no effect of the conditions (including the technology) on the rhythm of the sessions.

If we look at the sum of the occurrences of each condition, there is no clear indication that SAR conditions results in a greater number of interactions. However, we notice that the average number of occurrences is higher in the Artefice company than in the Stimulo company. This is due to the duration of the sessions (40 min for Stimulo and 60 min for Artefice). Nevertheless, if we consider the pace of the interactions (Table 15) we observe that the number of interaction per minute is surprisingly stable with very different conditions.

Table 15: Number of occurrences per minutes.

Session	Number of occurrences per minutes
Stimulo Standard	18,8
Stimulo SAR	14,6
Stimulo AR	15,0
Artefice Standard	11,7
Artefice SAR	14,4
Artefice AR	19,0

Therefore, we cannot validate Hypothesis 3 with these results and we notice a relatively stable pace of interaction in all the sessions regardless the conditions. As a further study to perform, we can correlate this result with the quality and quantity of ideas generated during the sessions, so as to derive indications about the efficiency of the sessions.

Another very positive result is that we find a good participation of end-users in all the sessions with a complementarity between actors. The potential of Spatial Augmented Reality to reduce ephemeral gestures while conserving a good engagement in terms of interaction pace needs to be further checked. The dynamism allowed by instantaneous changes and versatility of the object may certainly be more emphasised in WP5 protocol. Experiments of WP5 should investigate this further.

4.4. SPOKEN INTERACTIONS

The results for the analysis of the spoken interactions have been generated, as stated in Section 3.4, by means of transcribed protocols that have then been coded by six coders working independently. As all of them provided their individual evaluation of the segmented protocols (six protocols, one for each of the recorded sessions), evaluating the degree of coherence among the answers of different coders allows the reliability of the analysis to be assessed. Table 16 shows the values of the Fleiss' Kappa statistics, which measures Inter-Rater Reliability (IRR) for the six considered protocols. Values above the threshold of 0,66 suggest that the coders have a substantial agreement among them (where a value of 0 or less is no agreement; 1 means perfect agreement).

Table 16: Fleiss' Kappa statistics for the measurement of the Inter-Rater Reliability.

	CONDITIONS	INTENTIONS	ITEMS	PARAMETERS
Artefice	AR	0,8626	0,9118	0,9179
	SAR	0,8766	0,8909	0,8922
	Standard	0,8881	0,8339	0,8351
Stimulo	AR	0,8989	0,9073	0,8924
	SAR	0,7463	0,9189	0,9192
	Standard	0,8475	0,8561	0,8309

The figures of Table 16 refer to the three coding schemes used for the analysis of the protocols: Intentions, Items and Parameters (as was described in Section 3.4). These figures show that, despite some differences among the coders' agreement, all the different evaluations provided converging and reliable results.

The next sub-sections detail the results as computed and interpreted consistently with the metrics and the hypotheses presented at the end of Section 3.4.

4.4.1. Degree of interaction within the design session – Involvement of the participants

Hypothesis I:
Creativity in terms of interactions represents the degree of involvement of participants in co-design session (this enables ideas to get cross-fertilized more frequently).

Purpose:
Verify if Spatial Augmented Reality improves communication among co-designers having different profiles (Designers and End-users).

Metrics:
Shifts between speakers and category of speakers during verbal interactions.

Evidence from experimental data analysis:

- The largest number of spoken interactions happened in co-creative sessions run with standard design representations.
- Shifts between categories of participants (Designers and End-users) appear to be less frequent with the use of Spatial Augmented Reality
- In general, technologies to support communication and shared design representation provide data against the initial hypothesis.

Implications from data:

- It should be checked if hidden/unconsidered factors might affect the dynamic of spoken interaction during co-creative design sessions (e.g. profile of participants, style of the co-creative design session leader…)
- It should be checked if the presence of a shared design representation that is made more easily available and accessible to all the participants reduces the need to exchange viewpoints, as it potentially enables an easier interpretation of the design proposal and facilitates a common understanding (which makes the shifts in spoken interaction less frequent, but more effective and efficient).

For what concerns the degree of interaction among the different participants, the measurement of their involvement was analysed through identifying the shifts between different speakers, as they have been recorded in the protocols. Table 17 and
Table 18 present the outcomes of this measurement. Both the tables contain the same data but are presented in different ways to highlight similarities and differences with reference to the nature of the design task (Packaging design – Artefice vs Product Design – Stimulo in Table 17) or the specific technology used to run the session, if any (Standard, AR and SAR conditions in
Table 18).

Table 17: Summary of the dialogue shifts as recorded from the spoken interactions of the six considered protocols – Data organized by domain of application: Packaging Design (Artefice) and Product Design (Stimulo).

	Artefice			**Stimulo**		
	Standard	**AR**	**SAR**	**Standard**	**AR**	**SAR**
Total lines	1875	1512	1216	430	410	603
Speaker shifts (Individual)	1378	1248	857	380	283	275
Speaker category shifts (Between Designers and End-users)	1067	706	530	282	168	223
Speaker Shifts /Total Lines	73,5%	82,5%	70,5%	88,4%	69,0%	45,6%
Speaker category shifts/Total Lines	56,9%	46,7%	43,6%	65,6%	41,0%	37,0%
Speaker category shifts/Speaker Shifts	77,4%	56,6%	61,8%	74,2%	59,4%	81,1%

Table 18: Summary of the dialogue shifts as recorded from the spoken interactions of the six considered protocols – Data organized by the technology used to support design (Standard conditions; AR, SAR).

	Standard		**Augmented Reality**		**Spatial Augmented Reality**	
	Artefice	**Stimulo**	**Artefice**	**Stimulo**	**Artefice**	**Stimulo**
Total lines	1875	430	1512	410	1216	603
Speaker shifts (Individual)	1378	380	1248	283	857	275
Speaker category shifts (Between Designers and End-users)	1067	282	706	168	530	223
Speaker Shifts /Total Lines	73,5%	88,4%	82,5%	69,0%	70,5%	45,6%
Speaker category shifts/Total Lines	56,9%	41,0%	46,7%	41,0%	43,6%	37,0%
Speaker category shifts/Speaker Shifts	77,4%	59,4%	56,6%	59,4%	61,8%	81,1%

Figure 12 shows the distribution of the measured values in terms of:

- Total lines - the overall number of considered segments;
- Speaker shifts (Individual) - the number of dialogue shifts between any two speakers;
- Speaker category shifts - the number of dialogue shifts between two speakers of different categories (Designer or End-user).

These data are organized according to the domain of application of the technology, therefore this figure aims at highlighting within domain differences according to the technology used within the session.

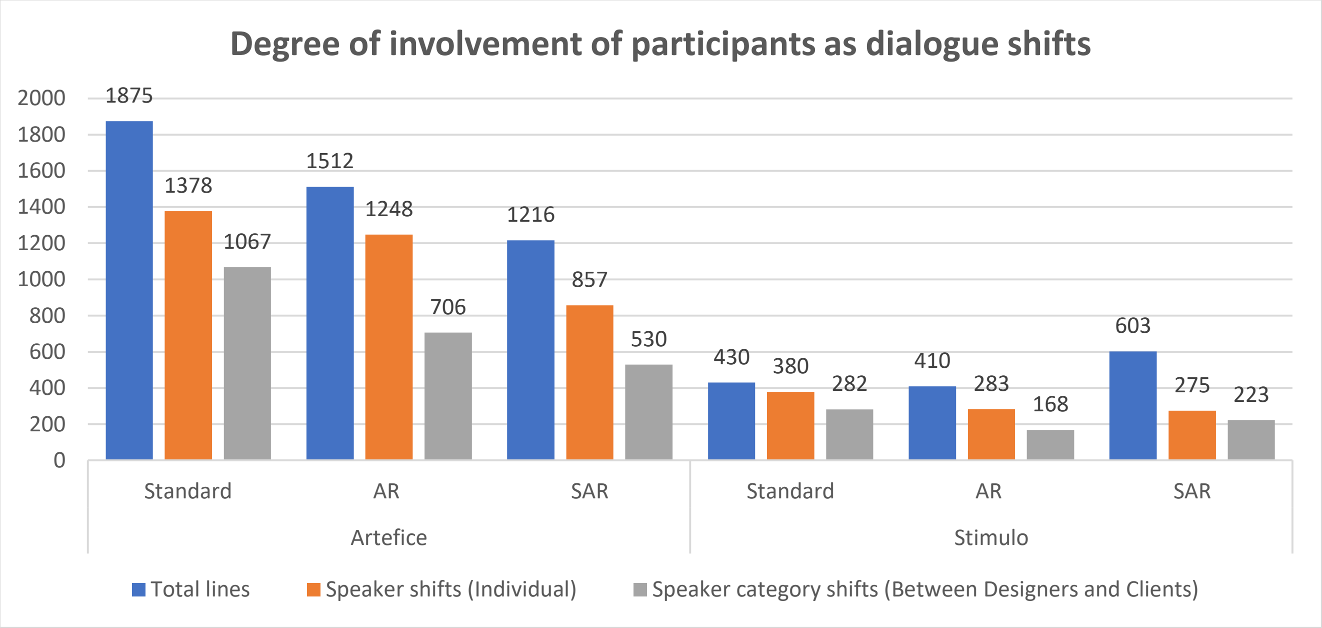


Figure 12: Observed data about the dialogue shifts in spoken interactions among the participants of creative sessions

These data show a large variability between the two design domains, which depends on, at least two factors. The two SPARK industry partners (Artefice and Stimulo) typically run creative design sessions that are different in terms of the design domain (packaging/product) and their duration (typical Artefice session duration is ~2 hours compared with ~1 hour for Stimulo). This is immediately visible by looking at the blue bars of Figure 12, which describe the length of the protocol in terms of the total number of lines it contains. Orange and Grey bars, respectively, refer to the shifts identified during the dialogue between any two individuals and between two different categories of individuals. The latter is a subset of the former.

This metrics aims at verifying the hypothesis that Spatial Augmented Reality is effective in enabling the communication among the different participants in the co-creative design session. This large variability among the sessions, as shown by observed values, does not allow to run meaningful comparisons.

The normalization of these data is reported in Figure 13 and Figure 14, which present the same data shown in the last three rows of, respectively, Table 17 and Table 18.

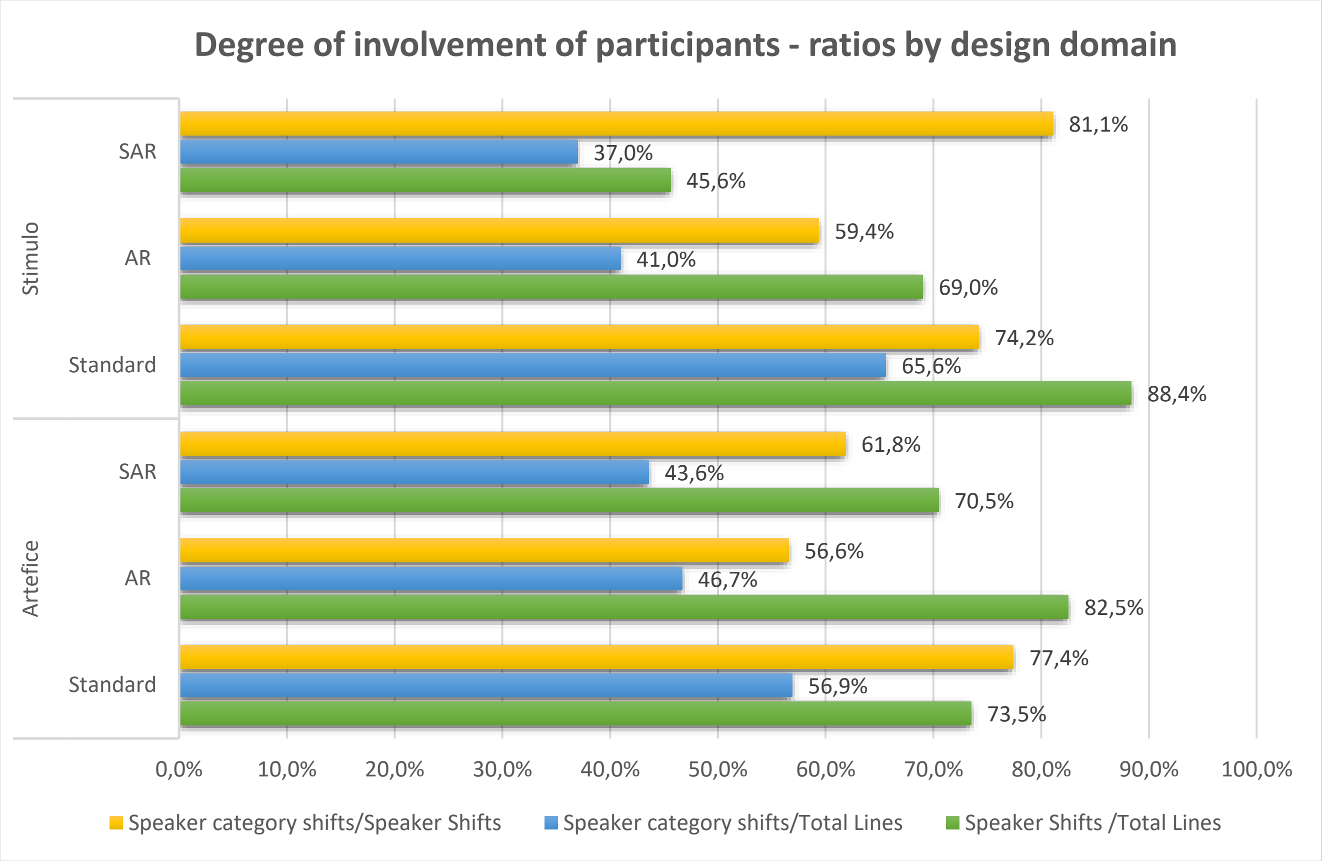


Figure 13: Ratios of dialogue shifts according to the design domain where they have been recorded

Figure 13 presents the normalized data organized according to the design domain in which the three different experimental conditions have been tested. This graph shows that there is no particular homogeneity among the different domain of applications, as each session, depending on the technology, provided different results from what was observed in sessions dealing with the same domain, but carried out with different means to support idea generation. For instance, the Standard condition provides different degrees of interactions when measured as speaker shifts in Packaging versus Product Design. A more regular behaviour can be observed by considering the blue lines of Figure 13, which represent the normalized dialogue shifts as they occur between different categories of participants (End-users and Designers). In both product and package design, the highest percentages of speaker category shifts were observed in the standard conditions (i.e. 65.6% and 56.9%), while the introduction of Spatial Augmented Reality appears to be a potentially inhibiting factor (i.e. 37.0% and 43.6%).

Such a counter intuitive outcome, at least with reference to the assumptions described in Section 3.4, suggests a reinterpretation of the observed data is necessary and also leads to some comments on the initially defined hypothesis. This result (blue lines) shows that the technology, whether AR or SAR, reduces the spoken interaction among participants. We suggest that this outcome could be a result of having a shared design representation that all the participants can visualize and comment on. This shared viewpoint on the design proposal, in fact, could be a good opportunity to reduce the verbal interactions. The presence of a shared design representation, that was clearly visible by all participants, allows them to skip a lot of the verbose verbal exchanges that are often required to create a common understanding of the design proposal within sessions that are not showing a shared design representation.

In order to highlight similarities and differences among the three conditions of the experiment, Figure 14 presents the same data of Figure 13, but organized according to the technology used in the session. This graph, unfortunately, shows that there are no clear differences/patterns when comparing the two co-creative sessions dealing with different design domains, as noted in the description of Figure 12.

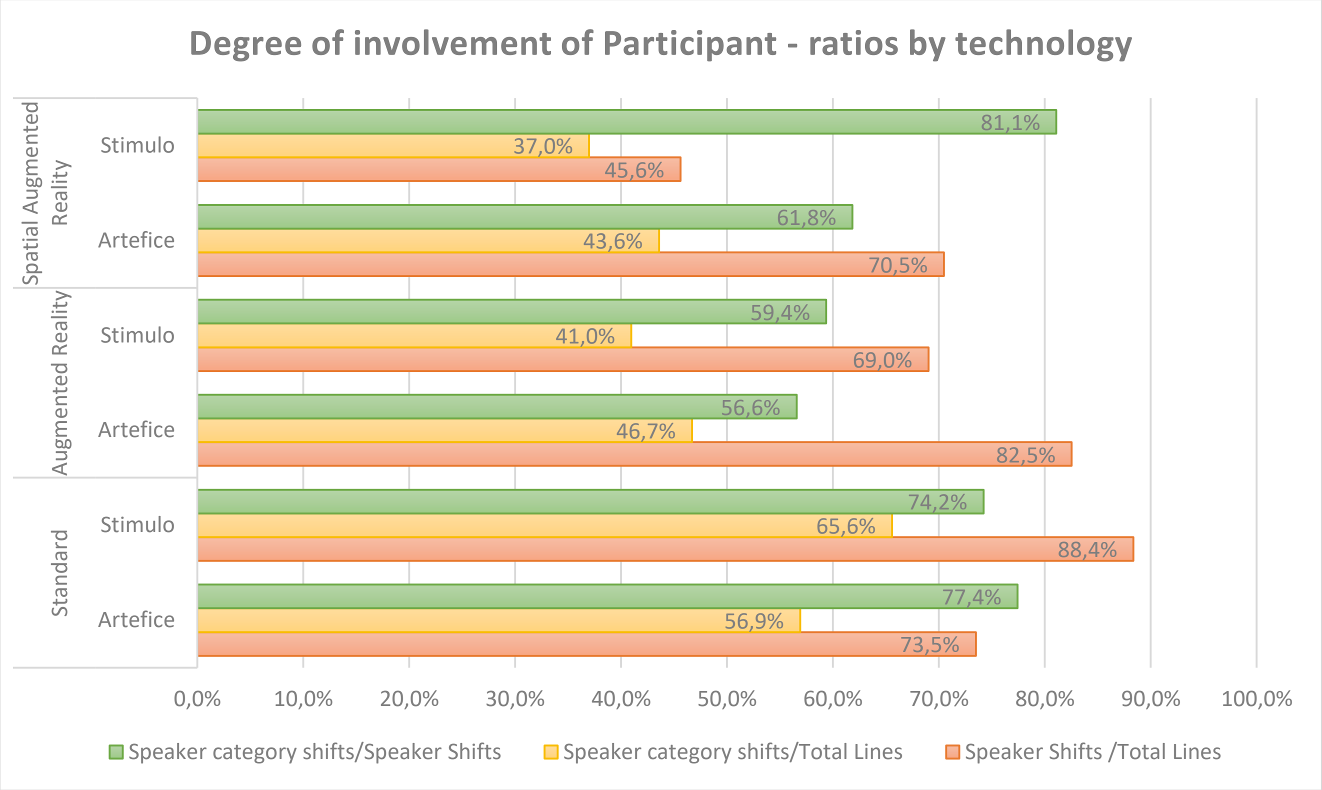


Figure 14: Dialogue shifts ratios organized by the technologies used in the experiments

Referring back to Table 17, it clearly appears that the largest amount of spoken interactions occurs in the standard sessions CG1. This also suggests that in a condition where more shared design representations (CG2 and TG) are made available to all the participants, the interaction is more efficient and reduces the need of exchanging various viewpoints (interpretations) on the design proposal to create a common understanding of it.

4.4.2. Fluency of idea generation

Hypothesis 2:
Creativity in terms of fluency represents the quantity of new ideas generated during a co-creative session.

Purpose:
Verify if using Spatial Augmented Reality (SAR) as the shared design representation enables a more fluent generation of ideas (Fluency increases).

Metrics:
shifts between intentions, from X (Analysis or Choice) to Synthesis

Evidence from experimental data analysis:

- Packaging design:
  - A: Overall amount of design moves: AR=0,8 x Standard; SAR=0,7 x AR
  - B: Overall amount of design moves of Synthesis: AR=0,9 x Std; SAR=0,75 x AR
  - Ratios (B/A) ➔ the three technologies perform similarly, SAR slightly better.
- Product Design:
  - A: Overall amount of design moves: AR= 0,9 x Std; SAR= 1,05 x Std
  - B: Overall amount of design moves of Synthesis: AR=1,3 x Std; SAR=0,9 x Std
  - Ratios (B/A) ➔ the three technologies perform differently: AR best, SAR worst.

Implications from data:

- Designers steering the co-creative design sessions, as well as participants, might strongly influence the results, as sessions having the same designers have less variability among technologies, overall. In such cases AR enables fluency more than SAR.
- For similar design tasks, Spatial Augmented Reality performs slightly better than its competitors (being it standard or simple Augmented Reality).

In a similar manner to the data presentation in Section 4.4.1, Table 19 and Table 20 present figures concerning the six recorded protocols. They provide data about the fluency in idea generation. Fluency has been here measured in terms of the design moves of Synthesis with reference to the parts of the design protocol that can be recognized as a design move (it means that the segments of the protocol that were coded as “Other” for the intentions are not considered within this analysis). Table 19, similarly to Table 17, presents the data organized with reference to the design domain of application (Packaging/Artefice; Product/Stimulo).

The condition to be verified with this metric is consistent with the overarching assumption stated at the end of Section 3.4: Spatial Augmented Reality should also allow for a more fluent idea generation process as it releases some cognitive load from working memory to process all the information dealing with the design solutions. Put another way, if I have an easy to understand design representation in front of me, I do not need to put a lot of cognitive effort into visualising or interpreting what the design solution looks like, so I can then dedicate more of my cognitive efforts to generating new ideas.

Table 20 presents the same data, but organized according to the technologies that characterize the three experimental conditions (Standard, AR and SAR) tested both on packaging and product design. This should help to highlight any possible similarity among the different technologies and then provide evidence for the identification of the impact on creativity of Spatial Augmented Reality.

Table 19: Fluency as design moves of Synthesis – Data organized according to the design domain of case studies

	Artefice			Stimulo		
	Standard	AR	SAR	Standard	AR	SAR
Total shifts (of intentions)	1013	803	571	205	184	213
Total shifts 2 Synthesis	175	155	115	35	46	32
% shifts 2 Synthesis on total shifts	17,3%	19,3%	20,1%	17,1%	25,0%	15,0%

Table 20: Fluency as design moves of Synthesis – Data organized according to the different experimental conditions (technology).

	Standard		Augmented Reality		Spatial Augmented Reality	
	Artefice	Stimulo	Artefice	Stimulo	Artefice	Stimulo
Total shifts (of intentions)	1013	205	803	184	571	213
Total shifts 2 Synthesis	175	35	155	46	115	32
% shifts 2 Synthesis on total shifts	17,3%	17,1%	19,3%	25,0%	20,1%	15,0%

The colour coding in Figure 15 is different to the previous figures. The blue bars in Figure 15 show the total number of shifts in the Intentions being discussed (without the segments that got coded as “Other” for Intentions). The orange bars show the shifts in Intentions from Analysis or Choice to Synthesis. In this way, the orange bars always represent a subset of the data represented by the blue bars. The point data (green diamonds) shows the number of shifts to Synthesis as a percentage of the total shifts in Intentions. This figure is subdivided into two halves. The one on the left organizes data consistently with Table 19 and the other one according to Table 20.

.

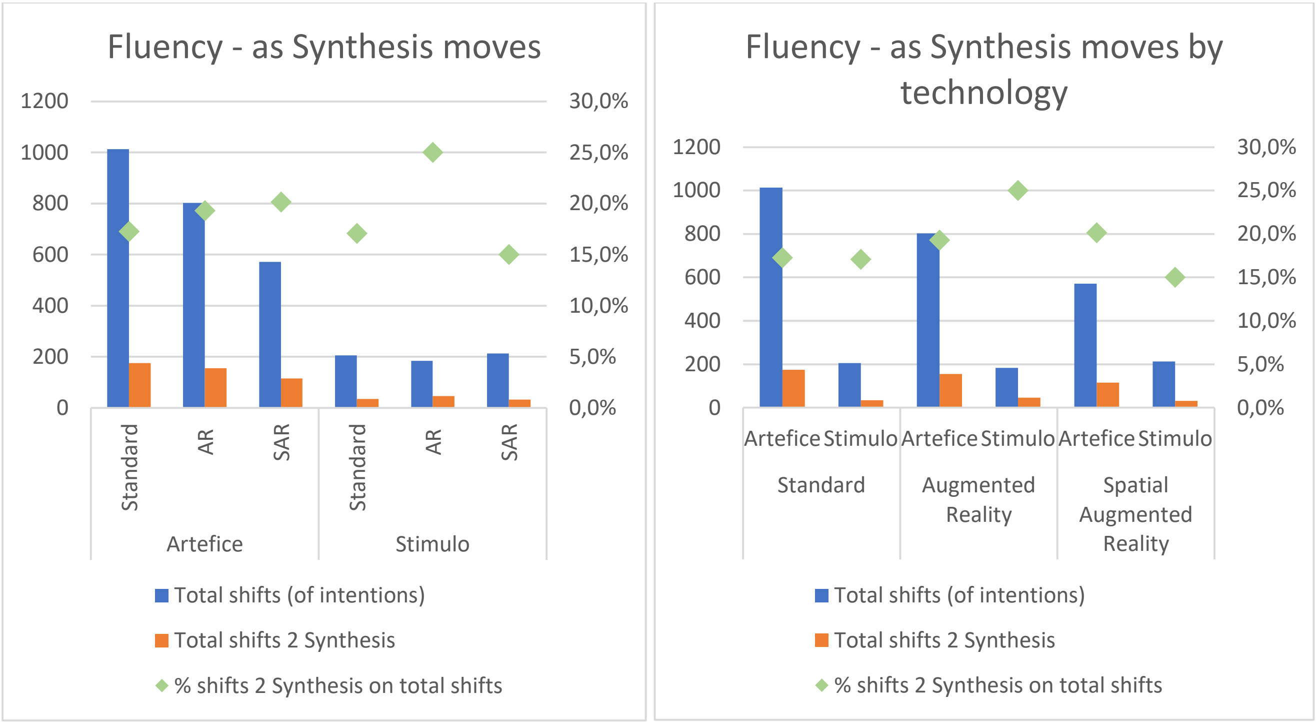


Figure 15: Fluency in idea generation as occurrences of design moves of Synthesis (Left – by design domain; Right – by technology)

Considering the blue and orange bars in both these figures, it clearly emerges that the two domains of application for the technology might significantly affect the results. In fact, the normalized amount of design moves of Synthesis (green diamonds) do not present significant differences in the three considered conditions when dealing with a packaging design task. In fact, despite these three sessions were characterized by a different amount of design moves (they differ with each other by 20/25%), the overall amount of design moves of Synthesis remains a consistent proportion, independent of the technology used in the session.

In contrast, we do see significant variations across the three conditions in the percentage of shifts to Synthesis in the product design domain, despite the fact that the overall amount of design moves is approximately the same in each condition. This is due to the different amount of design moves of Synthesis distributed among the three operating conditions (Standard, AR and SAR). Contrary to the initial expectations, the Spatial Augmented Reality performed a little worse. It is the best performing technology in packaging design (quite close to the performance of AR). However, for what concerns product design, AR overperforms all the alternatives in terms of fluency of ideas, while SAR is ranked last.

The different nature of these design task does not allow for a strong generalization of the results, which display variation with no specific regularity. The further demonstrations and showcase, together with interactive design sessions to be carried out with students and with professionals in WP5 should pay particular attention to the definition of further experimental testing activity where this counter intuitive finding can be checked.

Still with reference to the different operating conditions for the experiments described in this document, it is also worth mentioning that the sessions on product design involved the same two designers for all three conditions, but worked on different case study products. The consistency in the amount of design moves recorded in the product design session might be due to the fact that they were all completed by the same two designers. At the same time, however, it is also worth noticing that the co-creative design sessions for packaging design were held by three different pairs of designers, one per each co-creative session. The non-uniformity of results (differences of the order of 20-25%) for what concerns the overall amount of design moves in the packaging design sessions could be due to the different styles/working practices of the different pairs of designers. Some technical issues in the session on product design with SAR technologies (as the test was held with the first release of the platform, that has some SAR-IS communication bugs to be solved as they emerged just during the session) might also be one of the possible causes of this misalignment.

4.4.3. Exploration of the design space

Hypothesis 3:
Creativity in terms of quality of ideas depends on the matching between the exploration of the problem and the solution space (within the design space).

Purpose:
Verify if Spatial Augmented Reality (SAR) facilitates co-designers to efficiently explore design alternatives (both to identify requirements and propose design changes).

Metrics:
Correlations between couples of items and parameters in Analysis and Synthesis design moves.

Evidence from experimental data analysis:

- In packaging design, the correlation between analysis and Synthesis is stronger for Augmented Reality. SAR and Standard conditions show similar values of correlation.
- In product design all the conditions show very high values of correlation. SAR shows slightly lower correlation between design moves of Synthesis and analysis.

Implications from data:

- The match between problem and solution space is generally high → High quality ideas
- Differences between technologies are different considering Product and Packaging design. Augmented Reality typically performs better than its competitors.
- Further studies are needed to check if these differences are statistically significant and if the composition of the co-design team affects the results.

An efficient and effective exploration of the design space is a pre-condition for a creative design process that has a higher chance of providing ideas of higher potential (where ‘potential’ could be considered both in technological and innovation/adoption terms). The measurement of the exploration of the design space, in turn, should be considered with a comprehensive perspective. On the one hand, this mainly refers to the generation of diverse ideas (having a larger variety). On the other hand, however, it is possible to explore the design space by also considering the design moves that belong to the investigation of the proposed design solution, in order to identify issues to be solved and set new objectives. Put another way, the exploration of the design space might occur in case of design moves of Synthesis and Analysis. In the first case, it refers to the generation of ideas which are increasingly different from each other as the extent of the explored design space. In the second case, the investigation of the design space allows for the definition of new objectives or design targets that were not included within the design brief.

In order to measure the exploration of the design space during the co-creative design sessions, different indexes have been hypothesized to describe different nuances, as the exploration might entail different activities, as explained above.

The co-evolution of problems and solutions is a well-known concept in design creativity, as this kind of cognitive behaviour has been recorded and identified several times in design creativity and cognition literature (Dorst & Cross, 2001). Stemming from this concept, which focuses on the need to find a proper match between the problem and the solution space (within the design space), we expect that a careful analysis of the design proposal (in terms of items and parameters) should correspond to solutions that address the same items and components. This kind of recorded "matching" behaviour should correspond to co-creative design sessions where participants can properly set design targets and propose specific solutions for them.

Figure 16 shows several heat maps for the Artefice SAR session (TG). There is one heat map for each type of coded intentions (Synthesis, Analysis, Choice) plus one, at the end, that collects the whole set of intentions (sum of Synthesis, Analysis and Choice). Each of these heat maps aims at describing the extent of the exploration of the design space with reference to the amount of considered design moves dealing with a couple of items and parameters (segments coded as "Other" are not considered here, as they're also not to be coded for items and parameters). Numbers in the heat maps describe the amount of spoken interactions as design moves dealing with that couple of item and parameter. Dark coloured cells in the heat map mean that such dyad of item and parameter has been not considered during the co-creative design session. Colours from white to red progressively show spoken interactions that occur with higher frequency (white is low, red is high).

These heat maps, organized consistently with the Intentions of the co-designers, allow us to perform a preliminary check of coherence. As briefly mentioned above, this is based on the assumption that a careful analysis of the design proposal (which occurs in the problem space) allows for the generation (Synthesis) of more precise and targeted ideas (which occurs in the solution space). These ideas, in principle, and according to literature, should have higher chances of meeting the requirements defined during the stage of analysis of the solution. In other terms, the comparison between heat maps of Analysis and Synthesis allows the matching between the problem and the solution space (i.e. coloured cells in heat maps of Synthesis and Analysis should have similar nuances, to show that both the problem and the solution space received a similar exploration, assuming that this reflects in a better matching between the two).

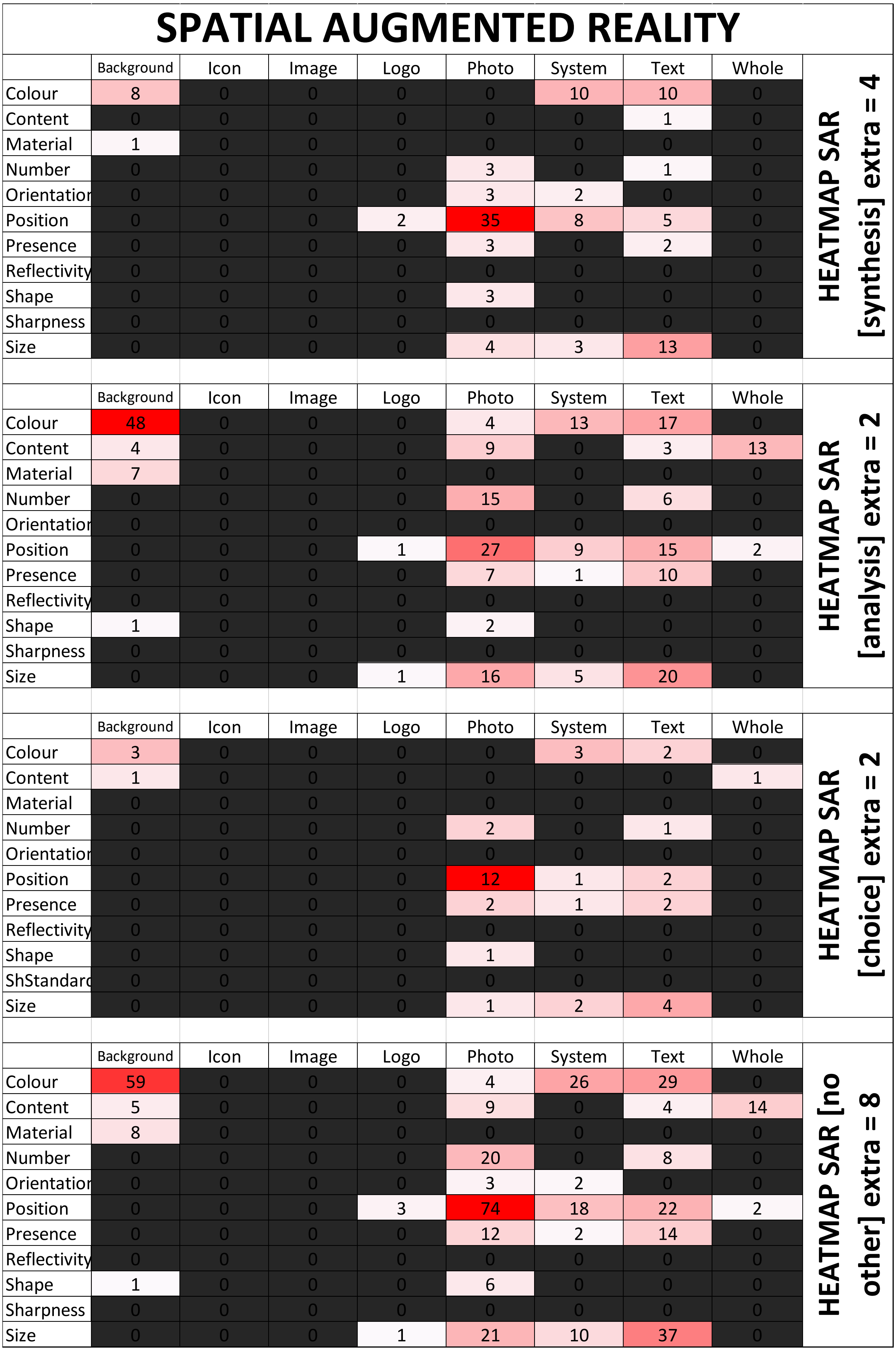
An overall perspective on the general degree of matching for what concerns the heat maps of Synthesis and Analysis is available in . The degree of matching has been computed, for reasons of convenience, by means of linear correlations between matrices of Analysis and Synthesis.

Table 21: Correlation values to compute the degree of agreement between the design moves of Synthesis and Analysis.

	ARTEFICE (Packaging design)			STIMULO (Product Design)		
	AR (CG2)	Standard (CG1)	SAR (TG)	AR (CG2)	Standard (CG1)	SAR (TG)
Correlation (Analysis vs Synthesis)	0,8173	0,6332	0,6843	0,9801	0,9746	0,9255

The above values show that there is a generally good agreement between the couples of items and parameters when they refer to design moves of Analysis and Synthesis. As the matching between the exploration of the problem and the solution spaces appear to be satisfactory, this result also suggests that the quality of solutions should be, in general, satisfactory. The degree of matching, however, does not represent an ex-post evaluation of real quality, as recognized by the designers during the ex-post evaluation of ideas done with the interviews documented in Section 4.2.1.

Comparing the technologies against each other, in each of the two design scenarios (packaging and product), the figures of Table 21 clearly show that product design allows for a higher matching, while the results for packaging design are less homogenous (AR significantly performs better than Standard and SAR conditions).



SPATIAL AUGMENTED REALITY

	Background	Icon	Image	Logo	Photo	System	Text	Whole	HEATMAP SAR [synthesis] extra = 4
Colour	8	0	0	0	0	10	10	0	
Content	0	0	0	0	0	0	1	0	
Material	1	0	0	0	0	0	0	0	
Number	0	0	0	0	3	0	1	0	
Orientatior	0	0	0	0	3	2	0	0	
Position	0	0	0	2	35	8	5	0	
Presence	0	0	0	0	3	0	2	0	
Reflectivity	0	0	0	0	0	0	0	0	
Shape	0	0	0	0	3	0	0	0	
Sharpness	0	0	0	0	0	0	0	0	
Size	0	0	0	0	4	3	13	0	

	Background	Icon	Image	Logo	Photo	System	Text	Whole	HEATMAP SAR [analysis] extra = 2
Colour	48	0	0	0	4	13	17	0	
Content	4	0	0	0	9	0	3	13	
Material	7	0	0	0	0	0	0	0	
Number	0	0	0	0	15	0	6	0	
Orientatior	0	0	0	0	0	0	0	0	
Position	0	0	0	1	27	9	15	2	
Presence	0	0	0	0	7	1	10	0	
Reflectivity	0	0	0	0	0	0	0	0	
Shape	1	0	0	0	2	0	0	0	
Sharpness	0	0	0	0	0	0	0	0	
Size	0	0	0	1	16	5	20	0	

	Background	Icon	Image	Logo	Photo	System	Text	Whole	HEATMAP SAR [choice] extra = 2
Colour	3	0	0	0	0	3	2	0	
Content	1	0	0	0	0	0	0	1	
Material	0	0	0	0	0	0	0	0	
Number	0	0	0	0	2	0	1	0	
Orientatior	0	0	0	0	0	0	0	0	
Position	0	0	0	0	12	1	2	0	
Presence	0	0	0	0	2	1	2	0	
Reflectivity	0	0	0	0	0	0	0	0	
Shape	0	0	0	0	1	0	0	0	
ShStandard	0	0	0	0	0	0	0	0	
Size	0	0	0	0	1	2	4	0	

	Background	Icon	Image	Logo	Photo	System	Text	Whole	HEATMAP SAR [no other] extra = 8
Colour	59	0	0	0	4	26	29	0	
Content	5	0	0	0	9	0	4	14	
Material	8	0	0	0	0	0	0	0	
Number	0	0	0	0	20	0	8	0	
Orientatior	0	0	0	0	3	2	0	0	
Position	0	0	0	3	74	18	22	2	
Presence	0	0	0	0	12	2	14	0	
Reflectivity	0	0	0	0	0	0	0	0	
Shape	1	0	0	0	6	0	0	0	
Sharpness	0	0	0	0	0	0	0	0	
Size	0	0	0	1	21	10	37	0	

Figure 16: An example of heat maps (SAR – Artefice) describing the amount of design modifications, requests of amendments or judgements in terms of couples of items and parameters. (From top to bottom: Synthesis, Analysis, Choice, Overall – Extra numbers refer to couples that have been not univocally described).

Hypothesis 4:
Creativity in terms of quality of exploration depends on the time spent refining solutions until the satisfactory configuration is found.

Purpose:
Verify if Spatial Augmented Reality (SAR) facilitates co-designers to efficiently explore design alternatives, as the capability to persist in the identification of good compositions or configurations.

Metrics:
Average time spent on a generic Item & Parameter (I&P) dyad considered during the whole protocol (Exploration index).

Evidence from experimental data analysis:

- Packaging design I&P combinations = 5 x Product design I&P combinations.
- Packaging design protocol duration = 2-2,5 x Product design protocol duration.
- The time spent for the exploration of I&P combinations in product design were not significantly different between technologies: AR = 1,15 x SAR; AR = 1,45 x Standard sessions.
- The time spent for the exploration of I&P combinations in packaging design is slightly longer with SAR than AR (10%) and SAR = 2 x Standard sessions.
- Along the co-creative design sessions for packaging design, the SAR setting shows a potential in allowing co-designers to spend longer exploring the rearrangement of solutions.

Implications from data:

- To verify if the measured persistence for the rearrangement of contents on top of the mixed prototype with the SAR setting produces the same results for packaging design.

In order to further measure the extent of the exploration, the co-creative protocols have been also analysed in terms of the average time spent on the rearrangements of items on the prototype (being it mixed as for SAR and AR), completely virtual (CG1 – Stimulo) or completely real (CG1-Artefice). These data are presented in Table 22, Table 23 and Table 24, where they are collected by design domain and technology.

Table 22: Unique combinations of items/parameters as coded, for the six protocols. Data organized by design domain (Artefice – packaging design. Stimulo – product design).

	Artefice			Stimulo		
	Standard	**AR**	**SAR**	**Standard**	**AR**	**SAR**
Time duration of the whole protocol (in seconds) [A]	5115	3554	4913	2108	1820	3178
Number of item&parameter combinations (per item in Synthesis design moves) [B]	24	27	21	5	3	6
Exploration index [A/B]	213	132	234	422	607	530

Table 23: Unique combinations of items/parameters as coded, for the six protocols. Data organized by technology (Standard, AR, SAR).

	Standard		**Augmented Reality**		**Spatial Augmented Reality**	
	Artefice	**Stimulo**	**Artefice**	**Stimulo**	**Artefice**	**Stimulo**
Time duration of the whole protocol (in seconds) [A]	5115	2108	3554	1820	4913	3178
Number of item&parameter combinations (per item in Synthesis design moves) [B]	24	5	27	3	21	6
Exploration index [A/B]	213	422	132	607	234	530

Table 24: Unique combinations of items/parameters as coded, for the three protocols on packaging design. Each of these protocols has been further split into three shorter episodes, consistently with the design proposal the participants started to propose new solutions from.

	Artefice								
	Std 1	**Std 2**	**Std 3**	**AR1**	**AR2**	**AR3**	**SAR1**	**SAR2**	**SAR3**
Time duration of the episode (in seconds) [A]	1967	1501	1647	747	1676	1131	1340	1974	1599
Number of item&parameter combinations (per item in Synthesis design moves) [B]	6	9	9	8	10	9	7	9	5
Exploration index [A/B]	328	167	183	93	168	126	191	219	320

Table 22 shows that, in general, the interaction with item & parameter couples is shorter in packaging design than in product design. Table 23, in turn, shows that the technology helps the co-designers to persist in searching for an arrangement of contents that is satisfactory (except for Artefice/AR). To this purpose, Table 24 aims at providing a more fine-grained viewpoint on what happened in the packaging design sessions. Indeed, those protocols have been macro-segmented into three episodes each. Each of these episodes corresponds to the exploration of a specific family of concepts, that stems from the different initial proposals (respectively numbered 1, 2 and 3).

Figure 17 shows the contents of Table 24 as histograms and a green dot, whose values on the y-axis (right side) measures the seconds during which the co-designers persisted rearranging the same couple of items and parameters.

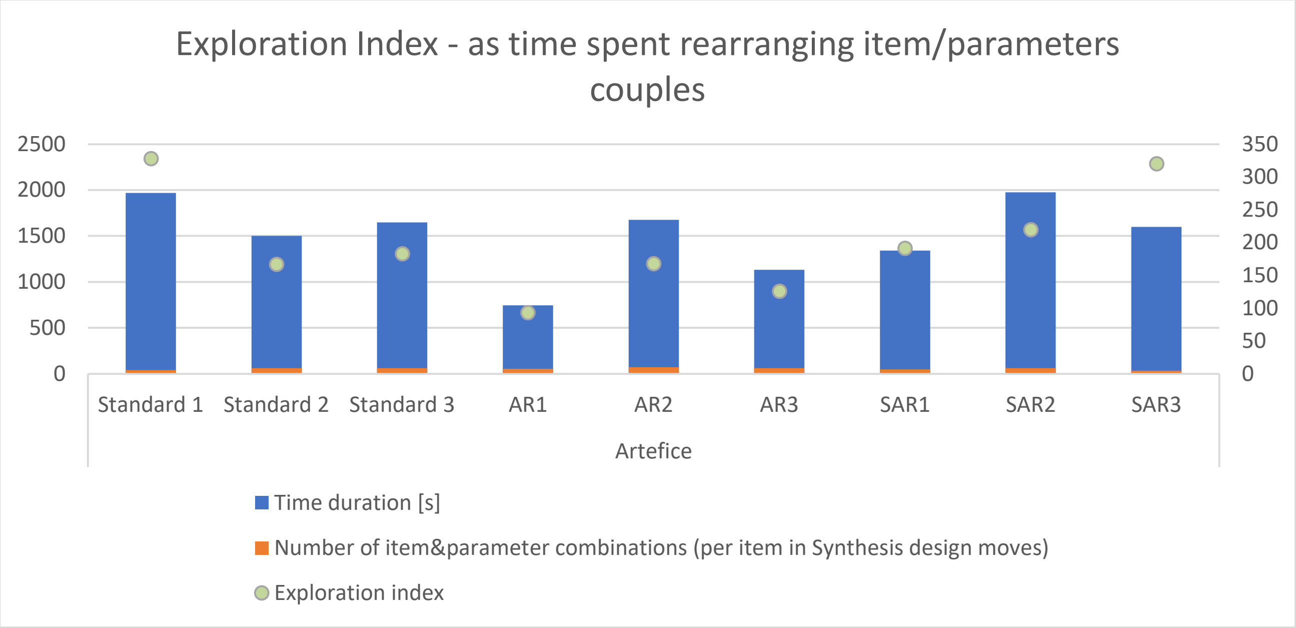


Figure 17: Time spend rearranging I&P couples and Exploration Index for the Artefice sessions, with breakdown by the three initial design proposals.

From Figure 17 it should be noted that, except for the first episode of the session carried out with standard means (Standard 1), participants in the SAR condition generally worked for a longer time with each dyad of item and parameter (green dots). This is finding is consistent with the hypothesis proposed. The value measured during the episode named Standard I represent an exception. This can be explained by the large amount of design moves of Analysis and the few design moves of Synthesis that characterized that episode, this makes the behaviour measured in episode Standard 1 not comparable with that of the SAR conditions (while episodes Standard 2 and 3 show a more uniform behaviour in terms of alternation of Analysis and Synthesis moves).

4.4.4. Convergent thinking in design

Hypothesis 5:
Creativity in terms of convergent thinking corresponds to the capability to make selections among various alternatives in design (whole design proposals or part of them).

Purpose:
Verify if Spatial Augmented Reality (SAR) facilitates co-designers to select contents to be used for the configuration of product interface and the composition of the packaging.

Metrics:
Ratio between the design moves of Choice and the overall amount of design moves.

Evidence from experimental data analysis:

- Differences between technologies are less marked for packaging design and more significant for product design.
- Almost 10% of the design session of Packaging design is focused on the choices of options with SAR and Standard settings (SAR slightly better than Standard Sessions). With the AR setting the ratio drops to 7,7%.
- In packaging design sessions, SAR boosts the selection of alternatives four times more than the other conditions.

Implications from data:

- SAR, with its shared design representation as mixed prototype, supports convergent thinking better than the other conditions.

The selection of concepts (or part of them, as in this specific case) is a typical creative activity as it involves the understanding of the overall objectives of the design (or part of it) and the specific solutions that have been developed to address those objectives. An efficient creative process allows the co-creative session participants to make the selection of the most promising concepts in a quick and effective way. The main expectations for this hypothesis concerns the improved capabilities of a SAR platform to facilitate the selection of concepts, hence convergent thinking. The shared design representation, which is visible by all the participants should be capable of removing interpretation ambiguities through a process of knowledge externalization which becomes concrete (real-like) on top of the design object's surface (that in the SPARK setting is to be kept blank with very small markers for its tracking).

Table 25 presents the amount of design moves of Choice as a percentage of the total number of lines captured in each protocol and the total number of design moves of all types.

Table 26, in turn, presents the same data rearranged so that the comparison between technologies used for the experimental campaign becomes evident.

Table 25: Choice design moves used as a metrics to describe convergent thinking – arranged by design domain.

	Artefice			Stimulo		
	Standard	AR	SAR	Standard	AR	SAR
Total lines [Tot]	1875	1512	1216	430	410	603
Design moves [Tot-Other]	1013	803	571	205	184	213
Total Choice	93	62	56	8	8	34
% of Choice on total lines	5,0%	4,1%	4,6%	1,9%	2,0%	5,6%
% of Choice on design moves	9,2%	7,7%	9,8%	3,9%	4,3%	16,0%

Table 26: Design moves of 'Choice' used as a metrics to describe convergent thinking – arranged by technology support.

	Standard		Augmented Reality		Spatial Augmented Reality	
	Artefice	Stimulo	Artefice	Stimulo	Artefice	Stimulo
Total lines [Tot]	1875	430	1512	410	1216	603
Design moves [Tot – other]	1013	205	803	184	571	213
Total Choice	93	8	62	2	56	34
% of Choice on total lines	5,0%	1,9%	4,1%	0,5%	4,6%	5,6%
% of Choice on design moves	9,2%	3,9%	7,7%	1,1%	9,8%	16,0%

Figure 18, in turn, graphically presents the data of the above two tables (Left – Table 25, Right – Table 26). The orange bars on the left side of Figure 18 show that, for packaging design, the number of design moves of choice is not so significantly different with reference to the overall number of segments in the protocols. On the other hand, for product design, the number of design moves of Choice was negligible in the Standard and AR condition, but non-negligible for the SAR condition.

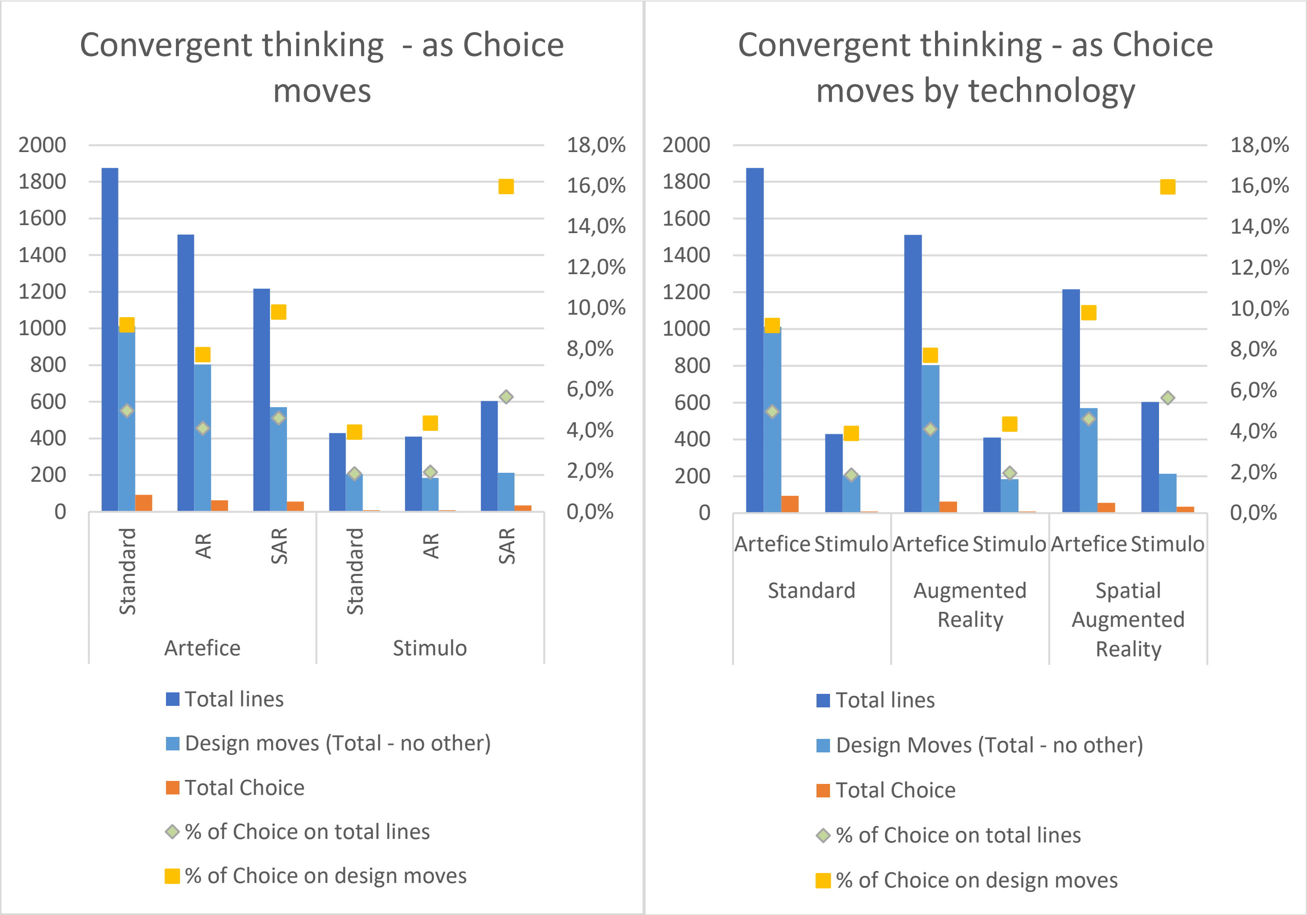


Figure 18: Observed data about the dialogue shifts in spoken interactions among the participants of creative sessions

The benefits of Spatial Augmented Reality in supporting convergent thinking can be better appreciated by looking at the yellow dots on the right graph of Figure 18, which computes the ratio of design moves of Choice with reference to the overall number of design moves (i.e. the calculation of this ratio takes into account the overall number of segments coded as Synthesis, Analysis and Choice. It intentionally overlooks segments coded "other", as they do not refer to the design proposal). In both the Artefice and Stimulo sessions, the SAR condition resulted in the highest percentage of design moves of Choice – 9.8% and 16.0% respectively.

4.5. DESIGN PROCESS EFFICIENCY METRICS

Findings:

- Design process efficiency metrics successfully applied to three historic case studies at Artefice.
- Up to 21 design iterations within the Ideas Production and Ideas Development phases.
- Several minor modifications made to the definitions of the metrics.

Conclusions

The relevance and practicality of the design process efficiency metrics has been validated. There seems to be significant scope to reduce the number of iterations in projects, without compromising on the quality of the final design output. The historical case studies will provide benchmark data for the longitudinal case studies to be completed during WP5

The design process efficiency metrics were successfully applied to three historical case studies completed by Artefice with their end-user 'Food Inc.' (not the real name). The brand strategy and visual identity for Food Inc. had already been defined through a previous project and so each of the projects concerned the application of this visual identity to new products.

The first project was the design of packaging for two flavours of a new pizza product. The project timeline is shown in Figure 19.

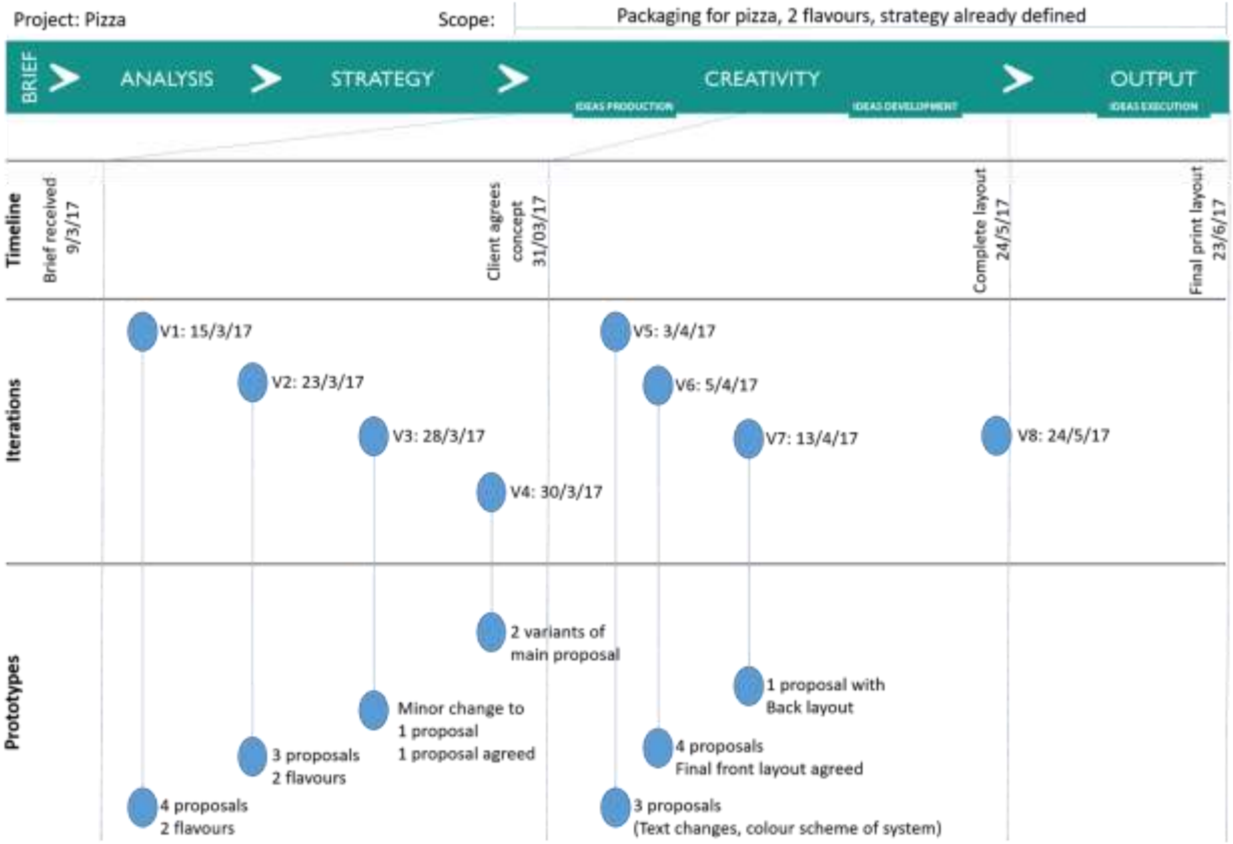


Figure 19: Project timeline for pizza project.

The second project involved the design of packaging for five flavours and two pack sizes of yoghurt. The project timeline is shown in Figure 20.

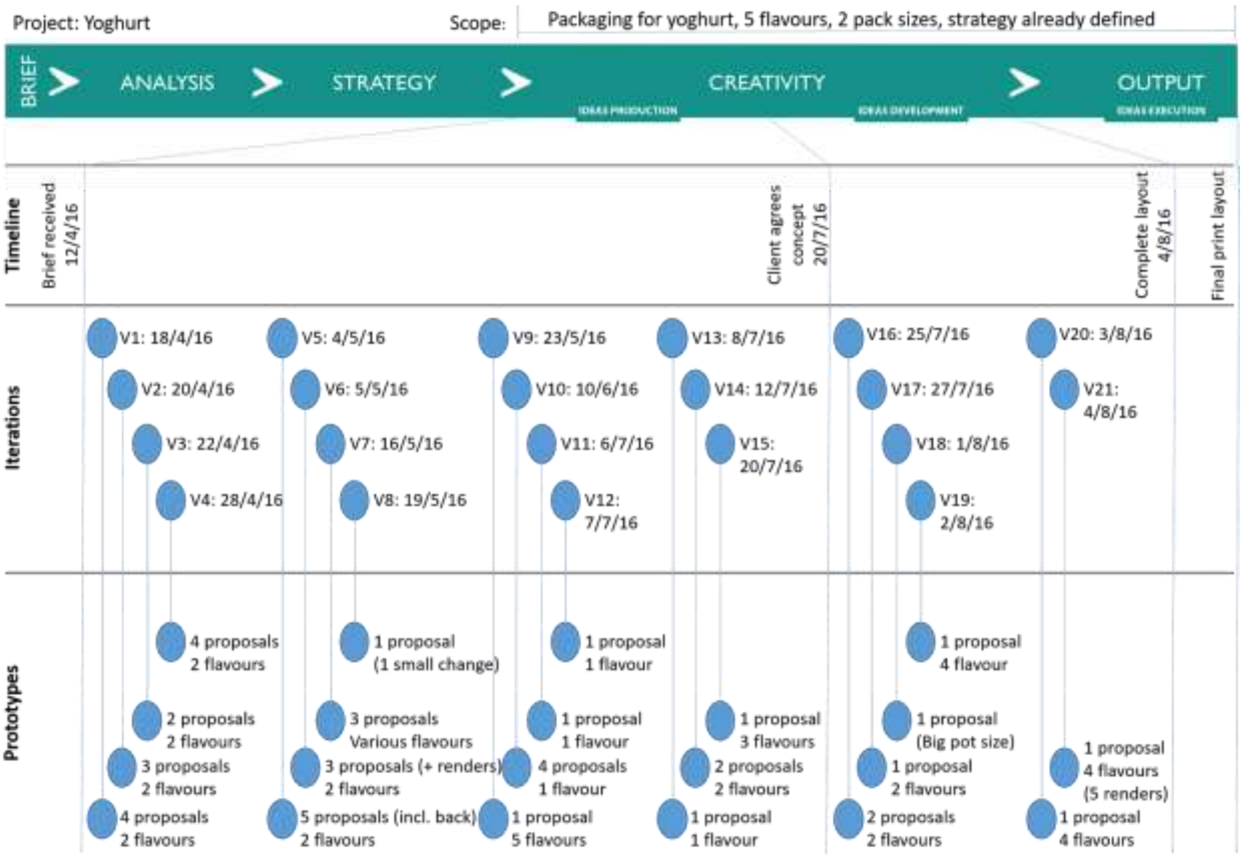


Figure 20: Project timeline for yoghurt project.

The third project involved the design of packaging for four soup flavours. The project timeline is shown in Figure 21.

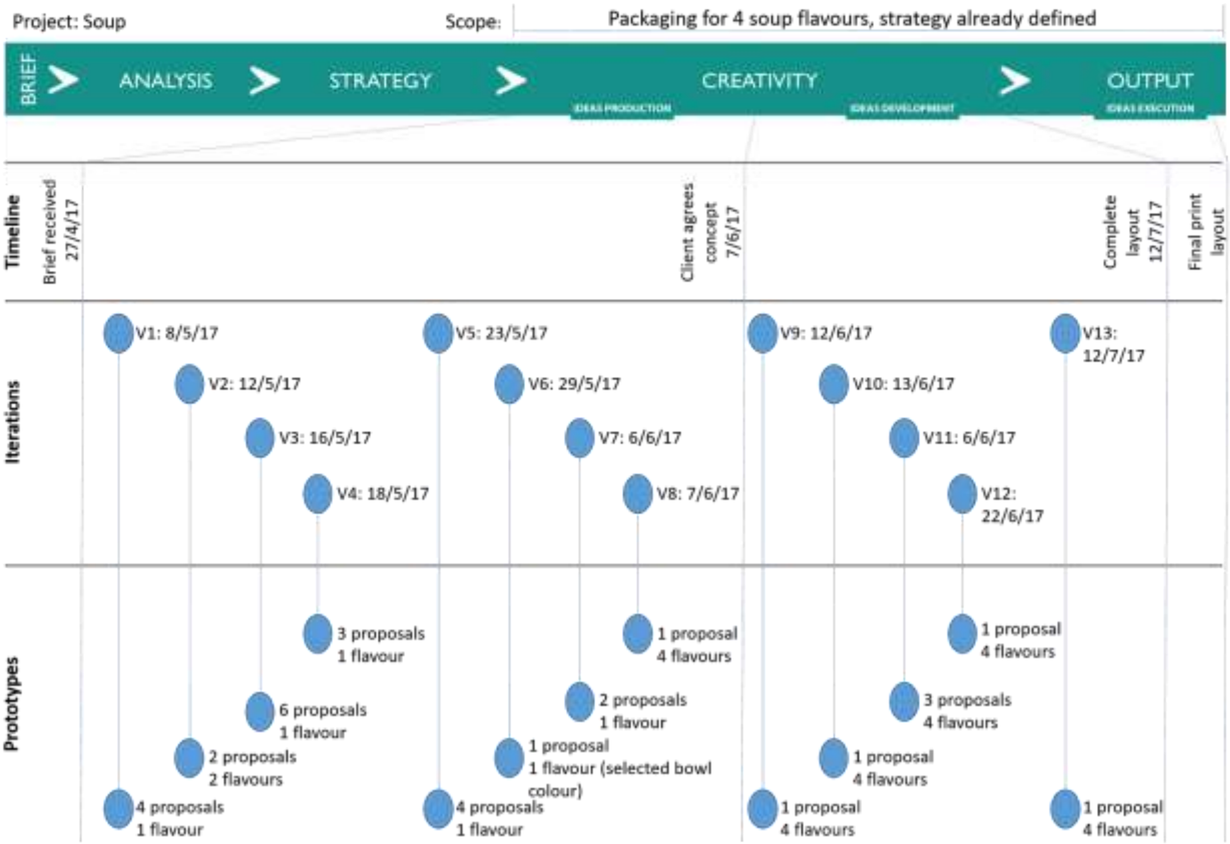


Figure 21: Project timeline for soup project.

Comparing the three project timelines, it can be seen that there is a significant number of iterations in each of the projects (at least eight per project), but that the yoghurt project in particular had a very high number of iterations - 21 in total from the start of the 'Ideas Production' phase to the completion of the 'Ideas Development' phase. Whilst some iteration in the design process is considered beneficial as it helps to ensure the final design is mature and well-refined, Artefice believes that there is significant scope to reduce the number of iterations in projects such as these, whilst still achieving a very high quality final design. The WP5 longitudinal studies will establish if SPARK can help to achieve this goal.

During the application of the metrics it was necessary to introduce a number of small refinements to the definitions of the metrics. First, the 'project lead time' metric was changed so that the end point considered was the end of the 'Ideas Development' phase instead of the launch of the product. This change was made because sometimes there can be long delays in the 'Ideas Execution' phase that are beyond the control of Artefice and are not related to design issues. For example, in one project there was a long delay as the end-user waited for the nutritional analysis of the product to be completed (which is required for the nutritional label on the packaging).

Secondly, concerning the 'cost of prototype production' metric, it was found that within this project the normal way of working between Artefice and Food Inc. did not involve any physical co-creative meetings. Instead, Artefice would send a variety of digital design representations (2D images, PowerPoint presentations, 3D renders etc.) to the end-user. The Food Inc team would then review the proposals within an internal meeting and before providing feedback to the commercial manager at Artefice by email or through a telephone discussion. The definition of the 'cost of prototype production' metric therefore had to be adapted to allow for this scenario.

Thirdly, the 're-work iterations' was modified to include a sub-metric on the number of versions of the design proposals sent to the end-user, feedback received and acted upon. For Artefice, this replaces the sub-metric on the number of co-creative sessions completed as no co-creative sessions were completed.

The complete, revised design process efficiency metrics were presented in Table 10, with changes highlighted in red.

In summary, the design process efficiency metrics have been validated through application to three historical case studies at Artefice. The results suggest that there is potential for design process improvement, particularly concerning the number of design iterations. The historical case studies will provide benchmark data for the longitudinal case studies to be completed during WP5, when a complete analysis of design process efficiency with and without the use of the SPARK platform will be conducted.

5. CONCLUSION

One of the primary objectives of the experimental activities reported in this deliverable was to contribute to SPARK Objective 3 - "Study and analyse how, and to what extent, the SAR technology can stimulate and enhance design creativity through a comparison against a pre-defined metrics in real operational environments." Whilst there is further testing to be completed within WP5 before firm conclusions can be reached, the analysis presented in the preceding sections demonstrates that we now have a much better understanding of 'how' SAR technology supports creativity and have been able to apply our pre-defined metrics to tentatively measure the 'extent' of that support.

The co-creative session performance metrics identified that, for the Stimulo sessions, the SAR and AR conditions performed best or joint best against the idea generation, task progress and filtering effectiveness metrics, with particular improvements in terms of the novelty and quality of ideas. This was despite the technical problems encountered during the SAR session, which were explicitly mentioned by designers as having caused disruption to the session. Within the Artefice sessions, the SAR condition performed significantly better than the other two conditions in terms of the quantity of ideas metric and was also best or joint best on the quality of ideas metric and the task progress metric. However, the standard condition performed best or joint best in terms of the variety, novelty, task progress and filtering effectiveness metrics.

The results of usability assessment and the follow-up survey indicate that that designers perceive the SPARK SAR technology to be more effective for co-creative design sessions than the use of standard design representations. In particular, designers seem to appreciate the freedom to try out many different ideas, and quickly filter out poor ideas. However, there are clearly areas for improvement, such as the sense of immersion in the tool when using the SAR technology. Overall, based on the results of the co-creative performance metrics, the SAR technology has shown good potential for supporting enhanced creativity within co-creative design sessions but has not consistently outperformed the other conditions across all metrics and both companies.

The gesture analysis made by the GINP team, found that the Spatial Augmented Reality condition encourage the end-users to interact in 33% of cases. It is less than in the Standard and the Augmented Reality conditions but, as it is said previously, the quantity and the quality of ideas metrics are better in the SAR conditions. So, if the percentage of end-users' interactions is lower in the SAR conditions, this does not prevent the emergence of ideas. Furthermore, the average number of occurrences of interactions in the SAR conditions is constant between Stimulo and Artefice with 14 to 15 occurrences per minutes. We observe a pretty stable pace of interactions in all the sessions, with a small drop in the number of occurrences in the SAR condition, which does not negatively impact the quality of interactions.

Concerning the use of the tools, we observed that whatever the type of artefact provided (except for the standard Stimulo session), the actors widely used the tool corresponding to the condition. This kind of observation allows us to conclude that, even if the SPARK platform is a new co-creative design tool, the actors used the technology in similar way to more traditional tools. The actors were able to adapt and translate knowledge of co-creative design collaboration to the SAR, and our results show that the quality of sessions are equivalent or better, even though it is a new task.

The analysis of spoken interactions, in turn, also provide preliminary evidence that SAR might play a role in supporting co-creative design. In contrast to our initial hypothesis, AR and SAR technologies, provides counterintuitive effects on communication. The consortium expected that SAR would foster spoken interaction as a source to cross fertilize creativity among participants. However, the evidence shows that spoken interactions and shifts among categories of participants (End-users and Designers) occur less frequently. However, we suggest that a shared design representation that is freely available to all the participants with a good viewpoint on it, reduces the need of sharing thoughts in order to align the viewpoints among co-creative sessions' participants.

With reference to the fluency of idea generation, indeed, SAR proved to be effective at least in those cases where the comparison is homogeneous among the same design tasks (run with different technological settings), as for packaging design in the experiments described in the present deliverable. This preliminary conclusion partially supports the previous finding about a less intense but equally or more effective communication among participants (who can focus on making modifications to the design proposal, rather than spending additional time developing a shared understanding the current state of the design proposal).

For what concerns the quality of the proposed solutions, there was no evidence to suggest that any of the technologies outperformed the others. In general, Augmented Reality (through more traditional visualization means, not projections) allows for a more precise match between the exploration of the problem space (to identify new requirements for the solution under investigation) and the solution space (as the amount of changes introduced in the design proposal).

Still with reference to divergent thinking (specifically, the exploration of the design space), it is worth noting that Spatial Augmented Reality enables the participants of a co-creative design session to spend longer in rearranging contents until the optimal configuration is found. For those sessions, which allowed a sub-decomposition of protocols, SAR also showed that it prevents any fatiguing effect in design, as during the co-creative session people tend to spend more and more time rearranging contents (expressed as couples of items and parameters).

With reference to convergent thinking, the SAR setting provided good evidence supporting the hypothesis that this kind of technology allows for a more effective and efficiency selection of contents for the solution, so that co-creative designers can quickly select what to reuse and rearrange and what to discard from the initial set of solutions.

The present report accounts for the results of the experiments carried out with the first release of the platform. While being functional, the prototype SAR technology still suffers from technical limitations in terms of reliability and functionality. It is therefore extremely difficult to consider the present results as representative of a typical SAR co-design situation. Nevertheless, we have been able to successfully carry out this first round of experiment and usability tests and this has provided important feed-back and input for the rest of the project.

6. REFERENCES

Boden, M. A. (2009). Computer models of creativity. AI Magazine, 30(3), 23.

Bruner, J.S. (1999). Culture, Self, and Other. *Sémiotique des cultures et sciences cognitives*. Colloque inaugural de l'Institut Ferdinand Saussure, Genève, 20-23 Juin 1999.

Cherry, E. and Latulipe, C. (2014) 'Quantifying the Creativity Support of Digital Tools through the Creativity Support Index', *ACM Transactions on Computer-Human Interaction*, 21(4), pp. 1–25. doi: 10.1145/2617588.

Civi, E. (2000). Knowledge management as a competitive asset: a review. Marketing Intelligence & Planning, 18(4), 166-174.

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. Design studies, 22(5), 425-437.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52(C), pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. Biometrics 33:159 {74}.

Miller, G., A. (1956). The magical number Seven: Plus or minus two: Some Limits on Our Capacity for Processing Information. *Psychological Review,* 63 (2):81-97 (1956).

Osborn, A. F. (1953). Applied imagination. Oxford, England: Scribner's.

Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. Design studies, 24(2), 111-134.

Tang, H. H., Lee, Y. Y., & Gero, J. S. (2011). Comparing collaborative co-located and distributed design processes in digital and traditional sketching environments: A protocol study using the function–behaviour–structure coding scheme. *Design Studies*, 32(1), 1-29.

Torrance, E. (1972). Predictive validity of the Torrance tests of creative thinking. The Journal of Creative Behavior, 6(4), 236-262.

Visser, W. (2010). L’utilisation du geste dans des réunions de conception architecturale. Conférence “Design & Complexity. DRS 2010”. Montréal (Canada).

Zaman, L., Stuerzlinger, W., Neugebauer, C., Woodbury, R., Elkhaldi, M., Shireen, N., & Terry, M. (2015). Gem-ni: A system for creating and managing alternatives in generative design. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1201-1210). ACM.

APPENDIX I: METHODOLOGY AND ASSESSMENT OF THE ON-THE-FLY METHOD

We present here an extension of Section 3.3.2 for the reader who wants more details on the validation of the method.

On-the-fly capture

In D4.1 we presented a first version of the software developed by the consortium to assist coders during the co-creative design session. This tool, named "Observer", allows, with a direct observation of the collaborative activity, to identify gestural artefact-centric interactions occurrences using a specific procedure. This first version of real-time coding method required some improvements.
In order to reduce the cognitive load of the coders, we decided to involve two of them in parallel, each coding a specific aspect (actors and artefacts). Five artefacts categories are considered: Tangible, Digital, Mixed, Ephemeral and None (None meaning that the interaction is not supported by any artefact).

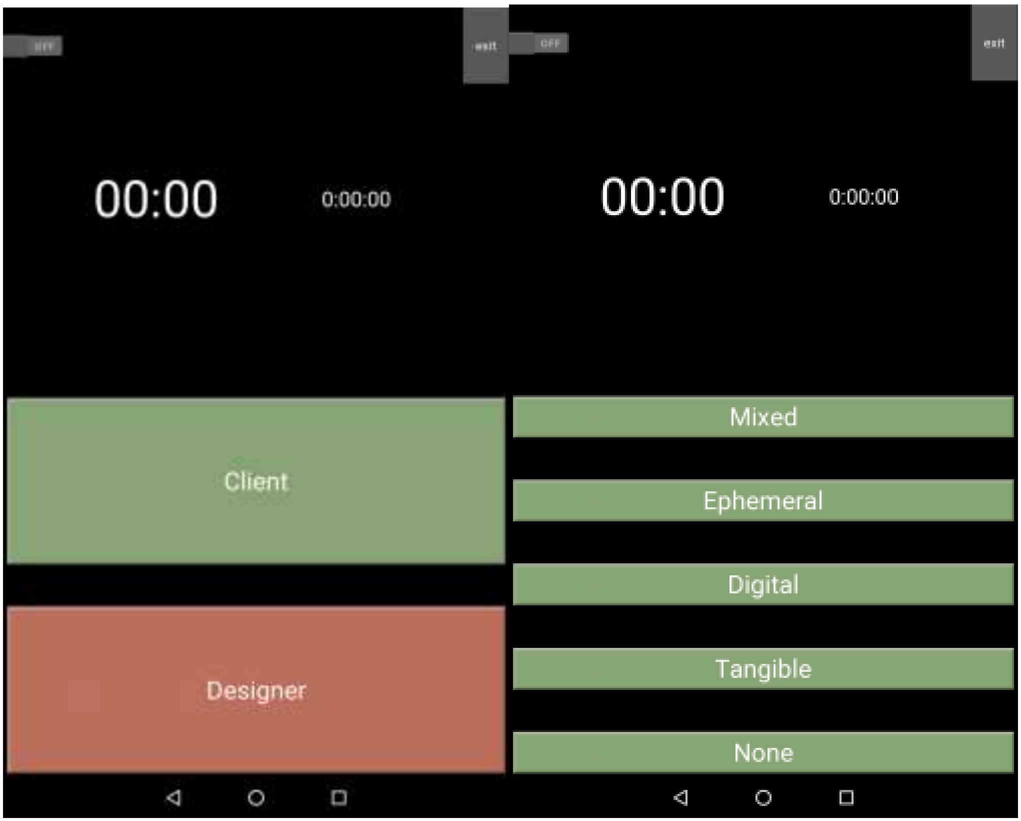
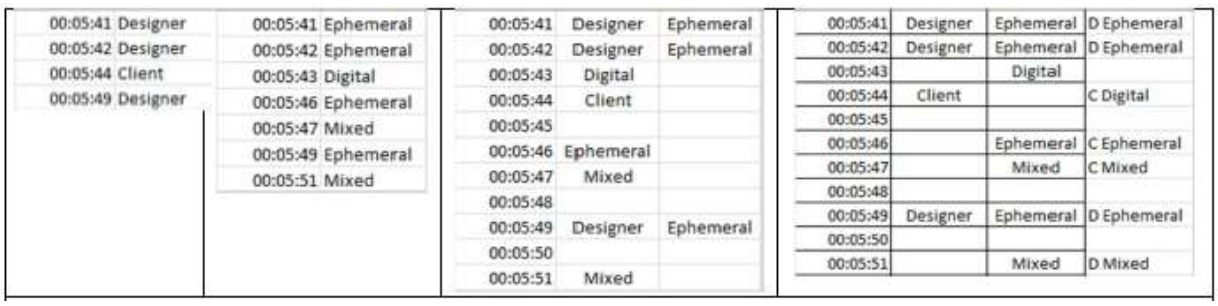


Figure 1.1: Observer interfaces according to actors and artefacts appearances

An automatic treatment and a manual adjustment are necessary to obtain a usable set of data in the form of an excel file that we can compute a comparative analysis with post-session coding in order to validate our method.

The first coding step completed, we process to the analysis of the data thanks to a software program we designed to merge the two separated coding results into a single set of data. This program, basically, analyses the content of each file in order to define the latest timestamp. Then it builds a data structure with a second-based timeline. Events of each data file are integrated in this new structure in separated columns. Finally, the data are saved in a csv file which can be imported in a wide range of corpus analysis software. We can then get a quantitative description of the interaction occurrences of the concerned session.

Figure 1.2 shows all steps of the analysis to get a final merging of the results. It shows excerpts of the coder's tables including timestamps of recorded events associated to the nature of the event (actor's intervention and artefact-centric interaction); a first merged table incrementing every seconds the occurrence of actors on the one hand and on the over hand, the occurrence of artefacts supporting the interaction if there is one at this time.



00:05:41	Designer
00:05:42	Designer
00:05:44	Client
00:05:49	Designer

00:05:41	Ephemeral
00:05:42	Ephemeral
00:05:43	Digital
00:05:46	Ephemeral
00:05:47	Mixed
00:05:49	Ephemeral
00:05:51	Mixed

00:05:41	Designer	Ephemeral
00:05:42	Designer	Ephemeral
00:05:43	Digital	
00:05:44	Client	
00:05:45		
00:05:46	Ephemeral	
00:05:47	Mixed	
00:05:48		
00:05:49	Designer	Ephemeral
00:05:50		
00:05:51	Mixed	

00:05:41	Designer	Ephemeral	D Ephemeral
00:05:42	Designer	Ephemeral	D Ephemeral
00:05:43		Digital	
00:05:44	Client		C Digital
00:05:45			
00:05:46		Ephemeral	C Ephemeral
00:05:47		Mixed	C Mixed
00:05:48			
00:05:49	Designer	Ephemeral	D Ephemeral
00:05:50			
00:05:51		Mixed	D Mixed

Input 1	Input 2	Automatic output of merging program	Manual building of combined events (last column)

Figure 1.2: From left to right: Actors table, Artefacts table, Automatized merged table, Final results.

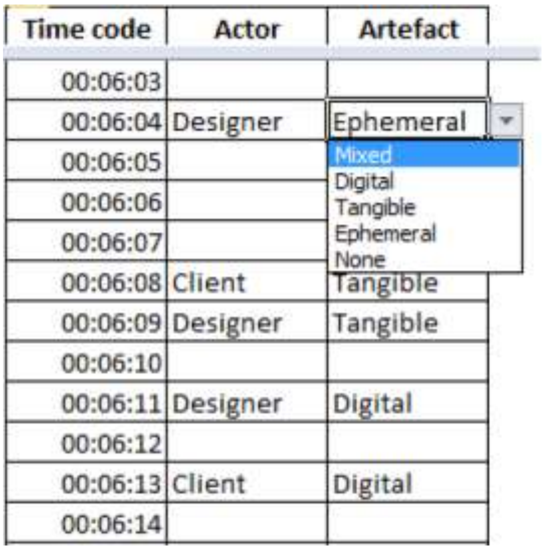
The last step is done manually. It consists in combining the actors and artefacts items according to a number of predefined rules. These rules have been defined in parallel of the splitting process elaboration. Actually, observers are coding separately the sessions. In order to ease the coding process and lighten the cognitive load, some rules were defined such as "not coding consecutives interactions of same nature". Consequently, some events might be marked by one observer (change of actor for instance) and not by the other observer. It leads to a half-empty line in the merged file that need to be reconstructed. This process, thanks to the automatic merging of the data set, is considerably fasten (see later the process performance). This last step (event combination) could be automatized, but we considered so far that a human eye is worthy in this activity in (rare) case of conflicts between data.

As we mentioned previously, we dedicated part of WP4 resources to the development of a coding On-the-fly method which would be comparable to a post-session coding to collect the quantitative data and would significantly speed up the analysis process. We will now validate coding approach by performing a comparison between the On-the-fly approach and a classical post-session coding through the analysis of several segments selected for that purpose. This is the aim of the next section of this Appendix.

Post session analysis

To collect the quantitative data of gestures realized during collaboration design sessions, we also proceeded to a traditional post-session coding. Two aims were associated to this post-session coding. The first one is to compare post-session coding results with the On-the-fly coding results to conclude to a potential equivalence of data collected thanks to these two methods. The second one is to make a qualitative analysis of data to answer the main research question which is "Does the SPARK platform (SAR condition) improve the performance of the co-creative design process compared to the typical design situations using other technologies?"

This post sessions coding is based on video and audio captures of the design scene. Cameras recorded four points of view to catch as much information as possible. Participants were equipped with a lapel microphone to record clearly their speech. Audio and video information combined, the researchers could use these files to proceed to the post-session coding. This coding is based on watching the scene and reporting on an Excel file the same information as the one gathered by Observer. Note that in contrary to the On-the-fly coding, each occurrence of an actor is immediately accompanied by an artefact which is the support of his intervention Figure 1.3 gives you a preview of the Excel file obtained for each session.



Time code	Actor	Artefact
00:06:03		
00:06:04	Designer	Ephemeral
00:06:05		
00:06:06		
00:06:07		
00:06:08	Client	Tangible
00:06:09	Designer	Tangible
00:06:10		
00:06:11	Designer	Digital
00:06:12		
00:06:13	Client	Digital
00:06:14		

Figure1.3: Post-session coding results.

The coder realized the completeness of the coding Post Session for the Stimulo sessions. On the other hand, all the Artefice sessions are very long and should require many hours of coding. That's why we choose to code randomly two to three excerpts of each condition, the totality of which is around 20-25 minutes. Later, to make the comparison between the coding On-the-fly and the coding Post session, we will compare all the conditions between them for Stimulo but for Artefice, we will compare only the parts which were coded in the Post session with the corresponding parts of the coding On-the-fly.

Robustness of the coding

A usability test was conducted, in order to verify that our coding on-the-fly is reliable:

- it is possible (in a cognitive way) to code using this Observer tool (i.e. the coder is able to follow the stream of interactions and track down the events); and that,
- two encoders obtain comparable results (robustness) when coding the same session.

To run this test, two researchers were invited to code, using the Observer tool, on different excerpts of previously captured sessions which were available in video format. Researchers were then standing in front of a computer, with their tablet, and had to code the session displayed on the screen, without any control on the stream (real time play). For each sample, each one of the coders coded first the actors who intervene (using interface Figure 1.1 left) and, in a second time, coded the artefacts that support the interaction (using interface Figure 1.1, right).
A first result from this test was that Observer is usable easily for coding the actors. It is also usable for coding artefact, although more effort is required because of the five types of artefact. A second conclusion was that two researchers coding three samples of different sessions lead to a good convergence on the results obtained. We decided to compare these results using Cohen's Kappa index.

This allowed us to measure inter-coder agreement, to identify and discuss the disagreements in the coding, and to compile examples of coding difficulties and final decisions that was taken on how to code these examples. To improve the coding reliability a coding book was then built, based on the enhancements made to the initial coding scheme following the reviews and discussions. It includes definitions and coding rules that help coders to converge to a consistent coding. These rules aimed at facilitating the coding, while being in line with the main analysis objectives (co-creativity assessment). Here are some examples of such rules:

- Prioritize end-users' interactions when designer and end-user talk at the same time,
- Even short interventions, just made of interjection such as 'OK' or 'Yeah', have to be coded,
- Prioritize ephemeral artefact comparatively to others types of artefacts, meaning that we code Ephemeral when an actor gestures around a Tangible, Digital or Mixed artefact.

Table 1.1: Cohen's Kappa for each session analysis.

	Standard		SAR		AR	
ON THE FLY	Cohen's Kappa	% agreement	Cohen's Kappa	% agreement	Cohen's Kappa	% agreement
Actor	0,55	73	0,71	84	0,58	76
Artefact	0,62	74	0,59	71	0,30	49
POST SESSION	0,50	64	0,45	59	0,59	68

If we refer to the Cohen's Kappa scale (Landis and Koch, 1977), the results show substantial agreement (green), moderate agreement (yellow) and fair (red). We can see a reduced agreement level on the first excerpt of the AR session with the Artefact coding (0,30). Artefact real time coding is the most difficult to grasp and requires substantial training. However, we considered the level of convergence sufficient as the two other samples show moderate to substantial agreement (SAR and Standard sessions) level.

To further improve inter-coder agreement, some rules have been defined to prevent the overload of the coder with repetitive events. For example, we decided not to code the reiteration of the same kind of interaction made after each other by the same actor. This might apply when a designer is interacting, stops speaking, and then re-initiates an interaction - the actors' coder will not code twice the designer. However, if two different designers carry on over each other, this will lead to several occurrences of the same actor type.

Comparison and assessment of the On-the-fly method

After realizing the coding of sessions using the two alternative methods, we compared these results. For this comparison, we focused on two kinds of information:

- percentage of actors' interventions obtained by each method,
- percentage of interactions quoted for each category of the scheme (artefacts).

Results of this comparison are shown in Table 1.2 and 1.3

Regarding to the Stimulo session we note that the percentage of actor's interventions are very close for the AR and the SAR condition but differ significantly in the Standard condition (Table 1.3). On the other hand, in experimentations led with Artefice, in the all three conditions, the percentage of interaction of actors with the On-the-fly coding is similar to the percentage of interaction of actors with the post-session coding (Table 1.2).

Table1.2: Percentage of actors' interactions in the Stimulo session.

	% of Designers interaction		**% of End-users interaction**	
	On the fly	Post Processing	On the fly	Post Processing
Standard condition	57,2	49,3	42,8	50,7
SAR condition	66,9	67,2	33,1	32,8
AR condition	63,9	63,6	36,1	36,4

Table 1.3: Percentage of actors' interactions in the Artefice session.

	% of Designers interaction		**% of End-users interaction**	
	On the fly	Post Processing	On the fly	Post Processing
Standard condition	52,1	54,2	47,9	45,8
SAR condition	66,1	63,2	33,9	36,8
AR condition	44,6	47,3	55,4	52,7

According to these two tables, we can say that the results from the coding On-the-fly are quite the same than the results from a traditional post-session coding. Except the standard condition for the Stimulo session, the difference in the other session is less than 2%, which a very good result in terms of our qualitative intended purpose.

If we pay attention to the percentage of gesture artefact-centric occurrences obtained from the 'on the fly' coding comparatively to the post-session coding, following results seem also to be very close and give us positive argument to consider that an On-the-fly coding approach is relevant and accurate enough.

Table 1.4: Artefacts occurrences during the Stimulo Standard session.

	% of Designers interaction		% of End-users interaction	
	On the fly	Post Processing	On the fly	Post Processing
Digital Artefact-centric	17	14	15	18
Tangible Artefact-centric	15	13	12	16
Mixed Artefact-centric	0	0	0	0
Ephemeral gesture	18	15	10	7
None	7	8	6	9

Table 1.5: Artefacts occurrences during the Stimulo SAR session.

	% of Designers interaction		% of End-users interaction	
	On the fly	Post Processing	On the fly	Post Processing
Digital Artefact-centric	2	3	0	0
Tangible Artefact-centric	0	0	0	0
Mixed Artefact-centric	38	37	20	19
Ephemeral gesture	16	15	8	5
None	11	13	5	8

Table 1.6: Artefacts occurrences during the Stimulo AR session.

	% of Designers interaction		% of End-users interaction	
	On the fly	Post Processing	On the fly	Post Processing
Digital Artefact-centric	42	42	27	26
Tangible Artefact-centric	4	2	2	2
Mixed Artefact-centric	0	0	0	0
Ephemeral gesture	12	14	4	5
None	6	6	3	3

Table 1.7: Artefacts occurrences during the Artefice Standard session.

	% of Designers interaction		% of End-users interaction	
	On the fly	Post Processing	On the fly	Post Processing
Digital Artefact-centric	2	1	0	0
Tangible Artefact-centric	43	42	38	29
Mixed Artefact-centric	0	0	0	0
Ephemeral gesture	4	6	8	11
None	3	5	2	6

Table 1.8: Artefacts occurrences during the Artefice SAR session.

	% of Designers interaction		% of End-users interaction	
	On the fly	Post Processing	On the fly	Post Processing
Digital Artefact-centric	4	2	4	5
Tangible Artefact-centric	3	1	1	1
Mixed Artefact-centric	40	43	22	22
Ephemeral gesture	8	9	3	3
None	11	8	4	6

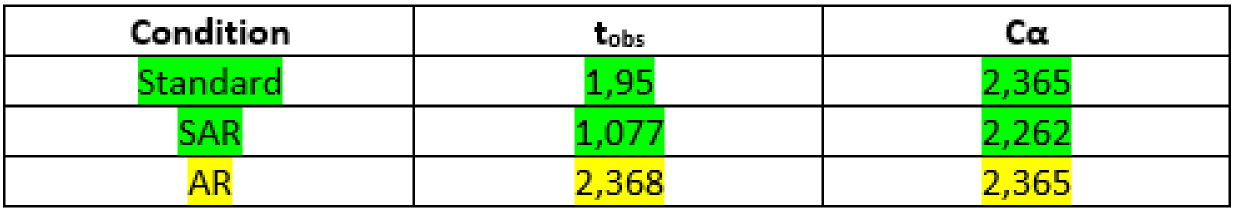
Table 1.9: Artefacts occurrences during the Artefice AR session.

	% of Designers interaction		% of End-users interaction	
	On the fly	Post Processing	On the fly	Post Processing
Digital Artefact-centric	36	31	44	36
Tangible Artefact-centric	0	0	0	0
Mixed Artefact-centric	0	0	0	0
Ephemeral gesture	4	11	8	12
None	4	5	3	5

To go further in the analysis of the similitude between these two coding methods, we proceed to a statistical test of averages comparison (Student's t-test). Through the statistical test we wanted to observe if there is a difference between the number of occurrences in the On-the-Fly coding compared to the Post-session coding. In other words, the aim of this statistical test is to argue against a similarity of the results obtained with these two kinds of coding methods. One factor is compared:

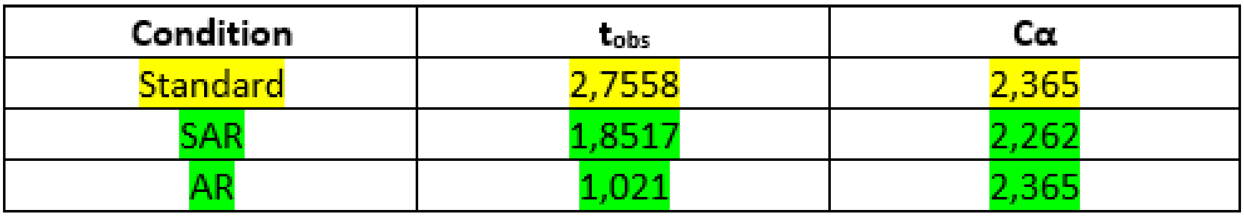
the type of session in two different modalities (On-the-Fly; Post Session). The samples of these two variables are matched and there is a bilateral test. With a p.value < 0,05, we obtained the following results:

Table 1.10: Results of the Student's t-test for the Stimulo sessions (validation if t_{obs}<Cα).



Condition	t_{obs}	Cα
Standard	1,95	2,365
SAR	1,077	2,262
AR	2,368	2,365

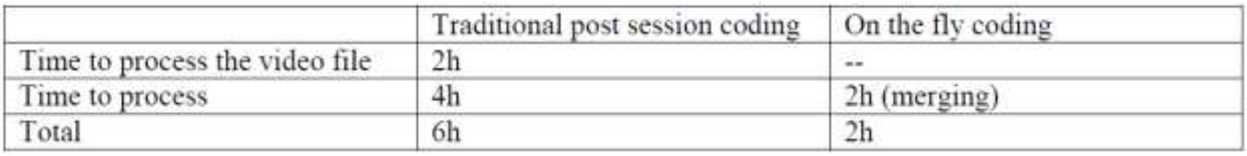
Table1.11: Results of the Student's t-test for the Artefice sessions (validation if t_{obs}<Cα)



Condition	t_{obs}	Cα
Standard	2,7558	2,365
SAR	1,8517	2,262
AR	1,021	2,365

We can conclude to similar results between the On-the-fly coding method and the Post session method when t_{obs}<Cα. Two conditions among six don't show that the coding On-the-fly is the same coding than using a traditional post-session coding. But, some results seem to be very close to be validated (yellow lines table 1.11 and1.12) and we can explain this difference. In particularly during the standard session of Artefice the length of design sessions (1:30) that was the last of a three session day, a potential lack of attention in the live coders due to fatigue, may have impaired the quality of the on-the-fly coding. This is certainly a factor we have to take into account in the future.
While the method On-the-fly still needs improvement, this statistical test is a supplementary argument to claim that quantitative data are very close between these two methods.

Table 1.12: Time taken to complete the coding of a 30 minutes section of session using the On-the-fly vs Traditional post-session coding approach.



	Traditional post session coding	On the fly coding
Time to process the video file	2h	--
Time to process	4h	2h (merging)
Total	6h	2h

Moreover, these results are provided in a much shorter time. If we do not take into account, on one hand the time of the session itself, and on the other hand the necessary time for quantitative analyses made from the encodings in Excel files that are common to both methods, the necessary time to do the coding of a 30 minutes episode of a session is approximately three time shorter for the On-the-fly approach compared to the traditional, post-session coding approach (Table 1.12).

APPENDIX II: DETAILS ON CODING SPOKEN INTERACTIONS

The six A/V recordings described in section 3.4 have been distributed to six Masters (MS)-level students in Mechanical Engineering that transcribed and, for the sessions conducted in Italian language, translated them into English (second box of Figure 4). They were instructed to transcribe all the audible utterances, even when the mixed audio track included overlapping voices. In these instances, they were allowed to re-listen to the individual audio track, as recorded by the lapel microphone. The transcribers were also instructed to capture utterances in lines, with a new line for each new speaker and with a length that is consistent with what might be presented on television subtitles.

In previous tasks, the team working on spoken interaction analysis developed tools to support non-experts to carry out the coding process (see Deliverable 4.1 - coding tables for items and parameters). These tables have also been applied during the post-processing of the experimental data from Task 4.3 and 4.4 by the same set of six students that transcribed and translated the protocols. In fact, they were also enrolled as coders of the six protocols (one protocol for each of the co-creative design sessions presented in Section 4.1).
The involvement of students (who we can consider as 'non-expert coders'), however, requires some adjustments in the standard pipeline of design protocol analysis. Non-expert coders, in fact, received pre-segmented protocols to code (third box of Figure 4). The reliability of coding, then, depends on the accuracy and the reliability of the segmentation process. The third box of Figure 4 and the diamond (4th step) clarify that this process has been carried out iteratively in order to have sufficient agreement among expert analysts about the appropriateness of the segmentation. Two expert analysts segmented the transcriptions of the six sessions according to the following rules:

- Every time a co-designer begins to speak and says something, this corresponds to a new utterance;
- Every utterance is to be segmented so that every segment should not be coded with more than one code for intentions (while multiple codes for items and parameters are allowed).

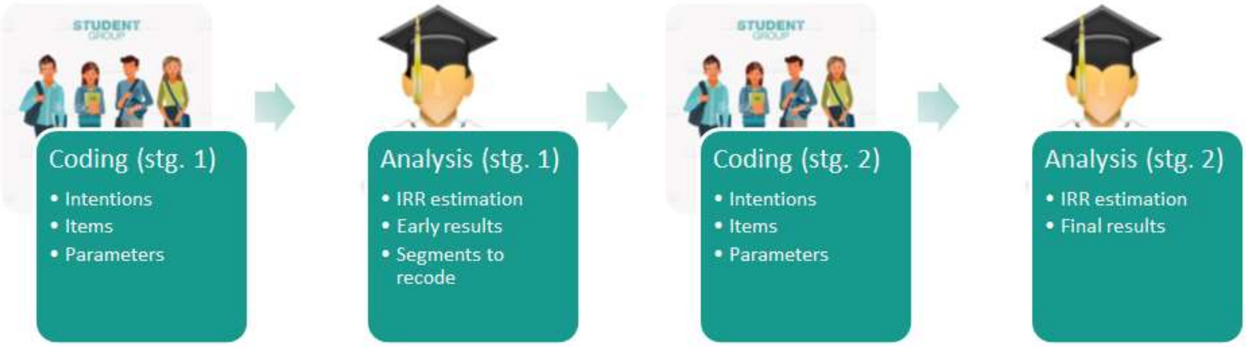
The first expert analyst segmented the protocols. The second one double checked the proposed segmentation and highlighted potential misalignments, which have been then solved by arbitration in order to produce, in total, six segmented protocols with a full agreement among expert analysts.

The six segmented protocols have been distributed to the six MS students for a first stage of coding so that each student had to code four protocols in terms of intentions, items and parameters. The coders were also exposed to a brief introduction to the codes of the mentioned coding schemes and the overall rationale of the analysis and of the project as a whole. The coders were told that each segment should normally receive not more than one code for intentions, while multiple codes are allowed for items and parameters. In order to check also the appropriateness of the segmentation process, the coders were also told not to force content that could be coded to multiple codes for intentions into a unique choice. Indeed, they had to highlight potentially ambiguous segments that had to be more appropriately fragmented. Just two coders highlighted potential needs of splitting the provided segments into more fine grained one. This happened for one segment each; actually, the same segment in the same protocol. This segment was split and reprocessed by the two coders who did not consider the original segment as two different ones.

The outcomes of this first stage of coding allowed the expert analysts to have a preliminary estimation of the results and, more importantly, of the reliability of coding through the assessment of the Inter-Rater Reliability (IRR). The results from the first stage of coding showed a substantial robustness of the segmentation process, as very few or no occurrences of ambiguous segments appeared.

The six coders processed the parts of the protocols that had resulted in lower degrees of agreement from the first stage of coding, in order to consolidate the coding for those 'challenging segments'. The second round of coding required them to interpret and define intentions, items and parameters for the challenging segments. Two students were involved in this stage for each protocol, so as to reduce the bias from previously considered protocols.

Fig. II.1 shows the detailed overall sequence of activities about the coding and analysis stage of Figure 4. Table II.1 displays the distribution of students with reference to the stage of activity and according to the protocols they analysed in each of them.



Coding (stg. 1)
- Intentions
- Items
- Parameters

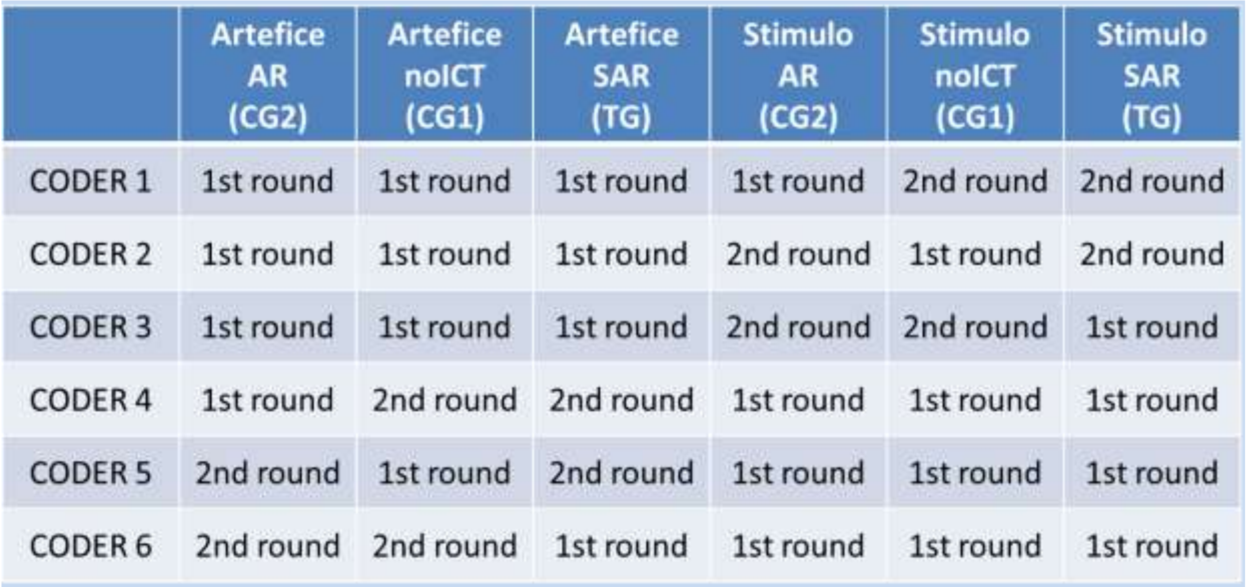
Analysis (stg. 1)
- IRR estimation
- Early results
- Segments to recode

Coding (stg. 2)
- Intentions
- Items
- Parameters

Analysis (stg. 2)
- IRR estimation
- Final results

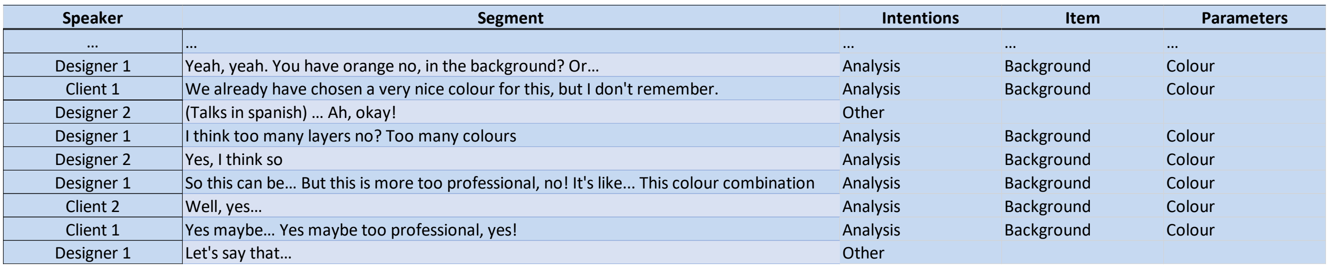
Figure II.1: The different stages and activities of the coding process and the related analysis for its results.

Table II.1: Distribution of the six non-expert coders with reference to the stage of coding and the design session/protocol.



	Arteface AR (CG2)	Arteface noICT (CG1)	Arteface SAR (TG)	Stimulo AR (CG2)	Stimulo noICT (CG1)	Stimulo SAR (TG)
CODER 1	1st round	1st round	1st round	1st round	2nd round	2nd round
CODER 2	1st round	1st round	1st round	2nd round	1st round	2nd round
CODER 3	1st round	1st round	1st round	2nd round	2nd round	1st round
CODER 4	1st round	2nd round	2nd round	1st round	1st round	1st round
CODER 5	2nd round	1st round	2nd round	1st round	1st round	1st round
CODER 6	2nd round	2nd round	1st round	1st round	1st round	1st round

Having checked the Inter-Rater Reliability (IRR) of the coding of the protocols and demonstrated that it was sufficiently reliable, the next stage of the analysis involved trying to identify links between the structured nominal and time related data obtained from the protocols and metrics of design thinking and creativity.



Speaker	Segment	Intentions	Item	Parameters
…	…	…	…	…
Designer 1	Yeah, yeah. You have orange no, in the background? Or…	Analysis	Background	Colour
Client 1	We already have chosen a very nice colour for this, but I don't remember.	Analysis	Background	Colour
Designer 2	(Talks in spanish) … Ah, okay!	Other		
Designer 1	I think too many layers no? Too many colours	Analysis	Background	Colour
Designer 2	Yes, I think so	Analysis	Background	Colour
Designer 1	So this can be… But this is more too professional, no! It's like… This colour combination	Analysis	Background	Colour
Client 2	Well, yes…	Analysis	Background	Colour
Client 1	Yes maybe… Yes maybe too professional, yes!	Analysis	Background	Colour
Designer 1	Let's say that…	Other		

Figure II.2: An excerpt of coded protocol (10 segments coded for intentions, items, parameters with the speaker – first column – responsible for the utterance).

APPENDIX III: INSIGHTS FOR CONSIDERATION WITHIN THE TECHNOLOGY DEVELOPMENT ACTIVITIES

The key pieces of feedback to emerge from the technology feedback discussions with designers after the experimental sessions are described below, organised by system module.

Information System

- Need a simple '**undo button'** to undo any operation made in the IS 3D editor interface. The current method is too time consuming.
- Uploading **multiple assets** to the IS assets library is fast and convenient, they can be easily dragged and dropped on the upload area.
- However, the overall IS interface (when you are in the 3D prototype editor) works at **low speed** after using it for a while (10-20 minutes).
- The organization of the **sessions and prototypes** is confusing. Inside one session you can prepare multiple prototypes but you can only use one during the live session. There is no way to select another prototype inside a session when it is in progress.
- It should be possible to **add new elements (2D assets)** when the session is in progress mode - for instance, go to Shutterstock.com, download an image, then use it straight away in the SAR model.
- The assets **tag function** is perceived as useful when running the design session but during the last WP4 tests it has not been used because it was not worth spending the time to tag the assets when preparing the session.
- The possibility to **group different elements (2D assets)** that are already placed on the 3D model and to apply a modification to all elements within that group would save time.
- The **selection of an element (2D asset)** on the 3D model is not very 'precise'. For instance, sometimes a texture is in the top layer but when the user clicks on it, another texture that is placed in a lower layer is selected.
- The function **move to layer** involves too many steps and could be simplified.
- The Information System **UI is not very clear**. Designers suggested to base the design of UI icons on those found in commonly used design software packages.
- Integration with **professional tools,** such as Photoshop and Illustrator, to prepare and modify assets already added to the library and to the prototype to avoid exporting and uploading again multiple versions of an asset.

SAR module

- Want to be able to work with **multiple prototypes** so that you could compare different concepts side by side.
- Need more **manual control of colour brightness and saturation**.
- Would like to be able to **represent user interface elements**, such as blinking LED lights. This would enable the system to give a better representation of the user experience as well as the basic appearance of the product.
- Would also be useful to be able to **play sounds**, again to help represent the full user experience.
- The **quality of projection** needs to be good, but does not need to be perfect as the system will generally be used for idea exploration activities, not final approval of layout.
- The tablet PC user interface worked well generally, but **precise positioning and orientation of elements** was tricky.

- **Colour 'eyedropper' tool** would be useful to copy a colour from one element to another.
- Would like more control of the **scaling of textures.**

This feedback has been presented to the technology development team and has been used to inform the development priorities for the version 2 and version 3 releases of the SPARK platform. The CSI survey will be applied again during the testing of the next iteration of the SPARK platform during WP5 and the results compared with those obtained in WP4.