Paper or Pixel? Comparing Paper- and Tool-Based Participatory Design Approaches

Matthias Heintz, Effie Law, Samaneh Soleimani

Paper or Pixel? Comparing Paper- and Tool-based Participatory Design Approaches

Matthias Heintz, Effie Lai-Chong Law, Samaneh Soleimani

University of Leicester,
University Road,
Leicester, LE1 7RH,
United Kingdom
mmh21@leicester.ac.uk, elaw@mcs.le.ac.uk, ss887@leicester.ac.uk

Abstract. Traditionally, in participatory design (PD) workshops, pens and paper are often used by participants to provide their design ideas. However, using a software tool to gather their feedback can have certain advantages. While some attempts to develop such tools have been undertaken, the basic question whether the tool-based approach is better or worse than its paper-based counterpart in terms of the quality of feedback gathered is rarely explored. We aim to address this research question by conducting three PD workshops with the paper-based and tool-based approach. In addition to the findings about the comparability of the two approaches, one of our main contributions to the future research on this question is the development of the coding scheme CAt+. It enables systematic comparisons of PD data collected with different methods and aims to support designers and developers to exploit PD results.

Keywords: Participatory design; Paper-based; Tool-based; Coding scheme

1 Introduction

Participatory design (PD) is a broad research area regarding the inclusion of prospective users in the design and development process for various physical as well as digital systems, products, and services ([1, 2, 3]). The goal is to gather their insights and input, especially design suggestions. There exists a proliferation of PD approaches, methods, techniques, and tools, for example: an expert designer co-creates with a single user to create a 3D mock-up from scratch using clays; a researcher elicits feedback on a simple 2D paper mock-up from a group of users using coloured pens [4]. In this paper, we present our research study on PD for "webapps", a collective term we use to refer to a variety of web-based applications, websites, and online portals.

With prevalent PD techniques (see [4] for an overview) feedback is typically elicited from participants in the form of verbal comments and sketches with the help of different materials and props such as storyboards, post-it notes, paper mock-ups, and acetates (see [5] for details). Using a software tool for PD instead, enables the user to interact with a prototype directly, creating a more realistic and engaging experience (e.g. [7, 8, 9, 10, 11]). This could result in more feedback of possibly better quality, as

adfa, p. 1, 2011.

compared to the paper-based approach. In addition, applying a software tool for data gathering can be advantageous for the data analysis, i.e. digitalisation of PD data at the time of capturing can enhance the effectiveness and efficiency of data processing (e.g. mitigating data loss; enabling software-supported data analysis).

However, there are also some constraints for using a software tool as compared to using pen-and-paper, e.g. computer access is needed; drawing on paper is more natural than on a screen [6].

When comparing the two approaches for PD, amount and quality of feedback is the most important aspect. Accordingly, we formulate two research questions (RQ):

RQ1: To what extent is the number of comments captured by tool-based PD activities on a specific webapp different from that by their paper-based counterparts?

RQ2: How are the comments captured by tool-based PD activities on a specific webapp qualitatively different from those by their paper-based counterparts?

The specific webapp for which our PD activities have been performed is a web-based portal under development in the European Go-Lab project. The goal of this project is to develop a portal [12] and online tools to facilitate the integration of online labs in science lessons. PD activities on the associated mock-ups have been performed with the target groups of the project, namely teachers and students from primary schools up to universities. According to the IBF Participatory Continuum Model [13] our PD activities lean towards the informant design [5]. When participants are used as informants for design decisions it is a normal PD practice to develop initial mock-ups and utilise them to gather user feedback and ideas. Specifically, myBalsamiq, the web-version of the prototyping software Balsamiq, which can be used to create interactive mock-ups from predefined or custom shapes (see section 3.1 for details), has been selected as the tool for our PD activities, because the mock-ups had been created with this tool. Therefore the results from the PD sessions would directly be visible and integrated in the design environment. Designers could not only instantly see it but also use it as a base or an input for the next re-design iteration. To further support the decision to use myBalsamiq for the PD activities and not only rely on the designers' choice of software tool, as it might be very good for design work but inappropriate for PD activities, we additionally performed a tool evaluation. From the results the use of myBalsamiq can also be inferred (see Related Work section for details). As it is not clear yet if the tool usage leads to results comparable to using the conventional way of gathering the feedback on paper, we aimed to answer the two research questions stated above to justify applying the tool over using paper in order to benefit from the described advantages. But it also goes the other way around: We are aware of some of myBalsamiq's shortcomings (e.g. the requirement for an Internet connection) and the question then becomes if paper is an appropriate way to be used as a back-up, e.g. in case of very limited Internet access.

To answer the two RQs, it is crucial to have a coding scheme for analysing and comparing PD user comments, thereby allowing systematic comparisons of the cross-media results (paper vs. digital). As such a scheme is lacking in existing literature, we have been motivated to develop one, which we named CAt+, to fill the gap. It comprises one Categorisation and three Attributes (Impact, Specificity, Uniqueness). By rating comments based on those attributes, it allows quantifying and assessing qualita-

tive characteristics of PD comments and thus enables a more thorough comparison (see section 5 for details). The coding scheme has been developed with generalizability in mind to enable other PD professionals to apply it for coding their participants' feedback. Although this generalizability of CAt+ is yet to be established, it can be considered as advancing a critical step for this specific area of PD data analysis. In this study we focus on the use of the coding scheme to compare data collected using the two different methods. In our future work, we will validate the impact of the scheme on enabling designers and developers to make sense of user feedback and implement changes.

2 Related Work

Some studies (e.g., [14, 15]) compare paper- and tool-based approaches to collect user input empirically in contexts other than but somehow related to PD, including software inspection [16] and multimedia design [17]. These comparison studies focused mostly on quantitative results (e.g., the number of defects found during inspection; task completion time) and on subjective opinions of the participants about the use of a tool versus the use of paper for performing specific tasks. However, they hardly compared the quality of the results (except [17] where the richness of the user-generated screens was evaluated to some extent).

The drawbacks of exclusively quantitative (or qualitative) approaches are increasingly recognized in the field of HCI (e.g., [18]). For instance, the total number of user comments cannot tell whether the comments address the content or the user interface design and interaction concept of the system evaluated. One of the challenges of qualitative analysis is the identification or development of a viable coding scheme.

Various coding schemes have been developed for a variety of topics, ranging from user comments on machine learning [19], over student comments on the teaching performance of professors [20] to YouTube comments [21]. [22] applied PD in a school setting to plan and improve lessons together with students and developed a coding scheme to code the spoken comments made during discussions. As those schemes are tailored to specific topics and use cases, they are not general enough to be applied directly to coding comments from different domains, e.g. PD user comments on webapp design.

While paper-based approaches remain commonly used for PD of webapps, some research efforts have been undertaken to mimic properties of paper in software tools suitable for PD by supporting graphical as well as textual feedback. One of them is GABBEH [23], an electronic paper prototyping tool which enables users to comment on the current design. Another one is DisCo [24] an online tool supporting PD workshops where adults and children collaborate. However, GABBEH only works with the DENIM tool [25] and DisCo is not yet publicly available for use.

To the best of our knowledge, no study has been conducted to compare systematically a paper-based and tool-based approach to determine to what extent software tools can be used to support or even replace paper-based PD activities. This gap has motivated us to conduct such a study.

The first and foremost step of the planned study was to identify an appropriate paper-based approach and software tool to use. As our target group involves children (students) as well as adults (teachers), we decided to choose a paper-based approach appropriate for children, and the Layered Elaboration approach [26] was proved to serve this purpose (see section 3.1 for details). It was shown that this approach could somehow be evolved into an online tool (DisCo, see above). However, the results gathered with DisCo were not yet compared to those gathered with the paper-based Layered Elaboration approach.

Given the constraints of DisCo and GABBEH, we had to find an appropriate tool to compare with the Layered Elaboration approach. The use of myBalsamiq by the mock-up designers was a strong argument to also use it for PD. But for a proper comparison of paper- and tool-based method, a sound tool selection is necessary. To guide this, ten requirements (six user-based and four developer-based) were gathered from three main sources: a literature review (especially [2, 4, 5]), end-user survey results, and unstructured interviews with main developers of the portal. The requirements scoped our search that resulted in 21 tools of which only two can fulfil most of the ten requirements: *Appotate* and *MyBalsamiq*, with the latter having the higher maturity and usability (details are reported in [27]). Therefore myBalsamiq was chosen for comparing the effectiveness of the paper-based and tool-based PD approach in terms of the quality of feedback elicited.

3 Design of Empirical Study

3.1 PD Study with Interactive Mock-ups

With the Balsamiq software, three mock-ups were created that shared the same basic structure but differed in the complexity of the learning content to address students of different academic levels (Fig. 1).

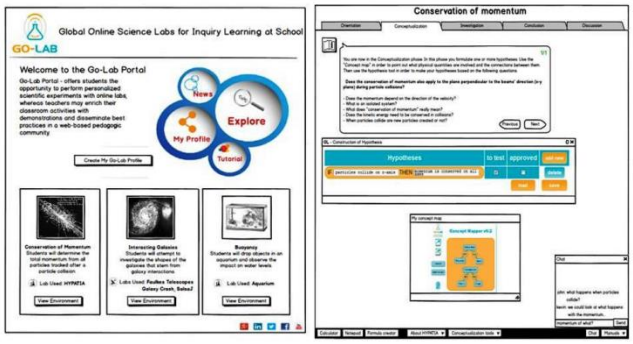

Fig. 1. Mock-ups of the portal: Homepage (left) and a webpage with learning content and tools provided by the teacher for students (right).

The PD work of the Go-Lab project has been implemented with two approaches: (1) tool-based with interactive mock-ups and (2) paper-based with printouts of those mock-ups. Details are described in the following.

Paper-based: Layered Elaboration is a more recent paper-based prototyping technique [26], which is simple to apply, and has the feature of keeping the initial prototype and comments from an iterative process intact. By overlaying different acetate sheets for the same mock-up printout, researchers can identify, for instance, which features have most frequently been commented on. The process starts by providing each individual or a small group of participants with a usage scenario, a set of ordered numbered printouts of the mock-up, a clipboard, and acetate sheets. Participants are asked to read through the scenario, put one acetate sheet on a printout, one after the other following the given order of printouts, and then provide feedback by annotating the acetate sheets with text and sketches, while working through the scenario on their own pace within a 45- or 60-minutes timeslot.

Tool-based: Participants are provided with computer access and work individually or in a small team of two or three people. They are introduced to the mock-up and then shown how to modify the mock-ups using different options provided in the myBalsamiq editor (Fig. 2). As with the paper-based approach participants are given a usage scenario specific to the mock-up to follow and are asked to give feedback while following the scenario. As the tool adds the user feedback elements in a transparent layer on top of the prototype, leaving the original prototype intact, it is comparable to the layered elaboration approach. Among the feedback options there are yellow "virtual sticky notes", which are added to the prototype by dragging them from the menu, to give textual feedback. Such notes are also commonly used in paper-based PD methods [28]. As with the paper-based method the participants progress through the scenario on their own pace within a 45- or 60-minute timeslot.

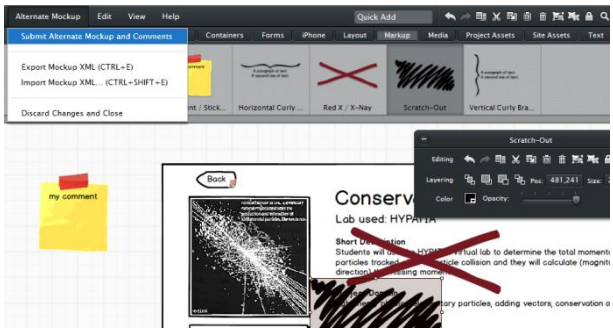

Fig. 2. The main feedback functions of myBalsamiq (screenshot of myBalsamiq taken and included with the permission of Balsamiq Studios, LLC)

3.2 Participants and Procedure

The mock-ups were used in PD workshops in two schools in a European country (School 1 and School 2) and in an international teachers programme in Europe (Teachers Programme) to gather ideas for the improvement of the prototypes and data for the comparison of the paper-based and tool-based approach. The workshops were started with two presentations (i.e. one briefly on the project and one on the PD process), followed by about a one-hour session of hands-on activities with the mock-ups.

School 1: The first PD workshop took place in a high school. Thirteen science students (mean age: 17 years old) were randomly divided into two groups with six using the Layered Elaboration approach (paper-based) and seven using myBalsamiq (tool-based). Two researchers were present to provide support, and each observed one of the two student groups.

The Layered Elaboration technique is typically applied on a group basis and with young children, albeit being applicable to other age groups as well. Because of the low number of participants, they did not work in groups at this event but gave feedback individually, to maximize the number of datasets, with no intention to use it as an intervention variable to compare the results from the two settings. Although the Layered Elaboration approach was altered slightly in this case, this did not affect the comparability of the data collected with the two PD approaches, as the participants using the tool to give feedback also did not work in groups at this event.

School 2: The second PD workshop took place in an elementary (or primary) school. The procedure was the same as in School 1, except having the students working in groups instead of individually, thus following the Layered Elaboration approach as initially described. 28 students (mean age: 10 years old) were randomly assigned to one of the approaches; 13 (in three groups of three and two groups of two) used the paper-based approach and 15 (in five groups of three) used the tool-based approach to give feedback. Because of the larger number of participants four researchers (two per group) were present this time, observing and providing support.

Teachers Programme: The third PD workshop was organized as one of the activities in a programme for international high school teachers. It took place in a research institute in Europe and involved 51 science teachers from 29 countries worldwide. The participants were split in groups of three. Feedback data from 8 groups working with the paper-based approach and 6 groups using the myBalsamiq tool were collected.

4 Data Analysis

All data were digitalised for further analysis. For the paper-based data, a set of the mock-up printouts and all the annotated acetates were scanned. With the use of the Gimp software the acetate part was removed from the scanned images, making the area where there was no drawing transparent again. Then all feedback from a single session was digitally layered onto the scanned mock-ups (Fig. 3).

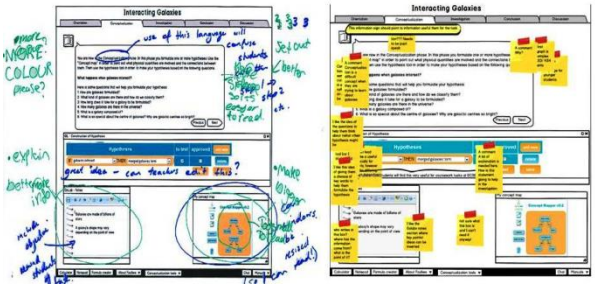

Fig. 3. Superimposed feedback gathered using the paper- (left) and tool-based (right) approach.

For further data analysis, all comments, including textual and graphical, were recorded in Microsoft Excel sheets with two columns. The first one labelled "source" contains "participant and screen ID" allowing the retrieval of the original feedback from the digital files. The second column records the textual user comment enhanced with researcher-generated details to make it easier to understand (e.g. description of the position or target) or a description of the drawing. Comments covering several ideas were split into individual rows during this step to prepare for further analysis.

5 Origin and Description of Coding Scheme CAt+

Each user comment was coded by two HCI researchers (i.e. fully crossed design [29]) with about two and six years of experience in usability research. Content analysis was applied to generate categories while coding [30]. Both researchers coded the comments in chunks of about twenty, introducing new category identifiers and definitions where necessary. To make sure that all comments were coded appropriately, earlier comments were revisited whenever a new category was introduced. The results were then compared and in case of discrepancy the researchers discussed till a consensus was reached.

Classifying comments based on their content can help to get an idea about the information contained, but not necessarily enables the comparison of the two approaches used to create the comments. For instance, comments on design are not necessarily "better" than comments on functionality; on this basis no conclusion can be drawn which approach is "better". Hence, a broader set of codes with meaningful ratings on the measurable quality of feedback such as specificity (the more specific, the better) was needed. Based on the literature on user defect classification systems (e.g. [31]) and downstream utility (supporting developers in addressing user feedback, e.g. [32]), we identified the following three major attributes: Impact – the extent to which the mock-up will be changed by addressing the idea expressed in the comment; Specificity (regarding target, reasoning, and solution) – the detailedness and thoroughness of the comment in terms of explicitly stating the target, reasoning, and possible solution to make an improvement; Uniqueness – the distinctiveness of the idea expressed in the comment. The initial coding scheme containing categorisation was thus completed by including those Attributes with values and definitions.

Accordingly, we name the coding scheme CAt+: **C**ategories plus **At**tributes, with the **plus** of Attributes compared to other coding schemes that only apply Category (or content) based coding.

5.1 Categories

Table 1 shows the categories of CAt+. Each rating is composed of a sub-category together with the main category, e.g. "Content-Add" (i.e. no feedback is just rated as "Content").

Table 1. Name, description, and example for each category of the rating scheme.

Category	Description	Example
Content	Comments on the learning material.	
Add	Request for more	"Put some text on this page."
Amount of text	Comment on the number of words used	"Shorten the text."
Change	Request for alteration (i.e. what is written), including typos	"However the questions don't seem to link in with the overall subject of the page."
Language	Comment on the wording (i.e. how it is written)	"Use more child friendly language."
Missing description	Request for explanation	"Unclear what to do with these questions."
Positive statement	Supportive comment	"I like the idea of having a video"; "Helpful instructions"
Remove	Request for deletion	"You don't need it to say in three minutes"
Terminology	Comment on only single words and their definition	"What does Buoyancy mean?"
Design	Comments on the visual appeal of the mock-up.	
Add	Request for new graphical elements or sound	"add crashing sounds"
Colourful	Comment on the colour or aesthetics	"The tabs could be more colourful."
Friendly layout	Comment on the suitability for children	"Kid friendly layout"
Negative statement	Criticism	"I do not like how the lines meet, it looks messy."
Not specific	General comment without detailed information	"Nothing to grab my attention."
Positive statement	Supportive comment	"Good use of colour to engage the students."
Screen layout	Comment on the positioning, order and size of elements on the page	"Make this bigger to fill the page?"
Terminology	Comment on only single words	e.g. button labels: "Complicated word ["Conceptualization" tab]"
Text layout	Comment on the format (size, style, colour, etc.) of the writing	"Better font, bigger font."
Remove	Request for deletion	"Don't need this [Page 2/2]."

Visual	Comment on the form/shape/sharpness of elements other than text and images	"… a different symbol could be used."
Functionality	Comments on interactivity of the mock-up.	
Add	Request for more things to do (e.g. buttons or apps)	"Maybe include zoom in and out buttons […]."
Missing description	Request for explanation	"What is this for? [Resize element on video]"
Positive statement	Supportive comment	"Good system dragging and getting the answer."
Remove	Request for deletion	"I don't think you will need the calculator."
Picture	Comments on the pictures in the mock-up.	
Layout	Comment on the positioning, order and size of pictures on the page	"… have bigger pictures so you can see better"
Missing description	Request for explanation (including all "picture unclear" comments)	"Try to describe the photos more so we know what they are."
Unknown	Comments of which the coders could not make sense or it was not clear how to address this comment (e.g. could be redesign or adding content or functionality).	
Not understandable	Comments of which the coder could not make sense	"Isn't the video"
Unreadable	Comments which or important parts of which could not be deciphered.	
Irrelevant	Comments not related to the mock-up itself	"less scribleing (sic!) [feedback to scratched out feedback from another participant]"

5.2 Attribute: Impact

The impact rating specifies how much of the user interface would change if this comment is addressed. Its possible values from 0 to 4 are defined as follows:

- No changes suggested (code: '0'). There is nothing which could have an impact on the mock-up if implemented (e.g. "good idea").
- Change affecting one element (code: '1'). As implementing the suggested changes would only influence a small part or a single element, the impact of this comment on the whole prototype is small (e.g. "the next [button] needs to be in a different colour to make it clearer.").
- Change affecting several elements (code: '2'). As multiple parts of a mock-up page would change if this feedback is addressed, the impact of this comment is medium (e.g. "Do titles for pictures and stuff").
- Change affecting the page on a level larger than element (code: '3'), e.g. by adding/removing an element to/from the page, which would cause a change of the layout of the other elements as well. As significant parts or even the whole mock-up page would change if this suggestion is implemented, the comment is rated as having a high impact (e.g. "Add some thing (sic!) here [white space on the right]."). If the comment does not specify an element, it is assumed that the whole webpage is the target and thus affected (e.g. "More colour").

- Change affecting several pages (code: ‘4’). As implementing the changes suggested in this comment would change various parts of the whole prototype, its impact is rated as very high (e.g. “log out option”).

5.3 Attribute: Specificity

The specificity of a comment indicates how detailed the feedback is. This influences how easily and fast the designer or developer can assess and address the feedback. If the target (e.g. an interface object) is specified, the developer is able to identify which part of the mock-up should be changed. If the reasoning for a comment is given, the developer may find a solution, even if none has been specified by the participant. If a solution is specified by the participant the developer can decide to implement it or take it as further guidance in finding a feasible solution. If such information is not specified, it might still be possible and reasonable for the designers or developers to make an educated guess. But if the informativeness of a comment is too low, even guessing might not be possible.

Accordingly, a Specificity rating consists of six sub-ratings based on the three aspects of the comment discussed: Target, Reasoning, and Solution, and if they are stated or guessable. The aspects “Target/Reasoning/Solution stated” can have the value 0, 0.5 or 1. If the respective information is given explicitly in the feedback, the rating is 1. If it is somewhat clear what the participant means, the rating is 0.5. If no information is given, the rating is 0. The aspects “Guessability of Target/Reasoning/Solution” have been introduced to rate if this information can be guessed (1) or not (0). If it is not necessary to guess this information (most of the time because it has been clearly stated), this aspect is rated as 1 (i.e. “it is guessable”).

5.4 Attribute: Uniqueness

Each idea or issue is either coded as 1 if it has not been mentioned before or as 0 if it is a duplicate. By adding up the coding over all comments, the number of distinct comments can be known.

6 Results and Discussion

6.1 RQ1: the Number of Comments

Through the PD activities, 701 valid comments (720 in total of which 19 were not related to the mock-up, e.g., a participant wrote a message to the researchers on the acetate “Sorry about my messy handwriting!”) were given by participants using the paper-based approach. 190 valid comments (191 in total of which one was unrelated to the mock-up) were gathered using the tool-based approach.

When comparing the number of valid comments per individual or per group, the paper-based approach resulted in more than twice (School 2: 15.2 compared to 7.0) or even three times (School 1: 51.0 compared to 15.1; Teachers Programme: 27.0 com-

pared to 8.3) as many comments per individual/group as did the tool-based approach. Due to the limited space, Table 2 only shows the results grouped by the main categories. Table 3 shows the percentage (to account for and offset the vast difference in number of results gathered with paper and tool) distribution of comments to categories for the three PD events, for the categories with a difference in the percentage values larger than 5% between the paper-based and tool-based approach in at least one event.

Table 2. Number of comments per participant (School 1) or group of participants (School 2, Teachers Programme) for each of the main categories and in total.

Main Category	**Comments per participant/group**					
	School 1		**School 2**		**Teachers Programme**	
	Paper	**Tool**	**Paper**	**Tool**	**Paper**	**Tool**
Content	8.5	3.0	3.7	3.6	13.1	5.8
Design	14.5	8.0	6.3	0.8	4.4	1.0
Functionality	21.0	2.7	2.3	1.4	5.9	1.2
Picture	0	0.3	0.5	0.4	1.9	0
Unknown	6.7	1.1	1.8	0.8	0.8	0.2
Irrelevant	0.3	0	0.6	0.0	1.1	0.2
Total	**51.0**	**15.1**	**15.2**	**7.0**	**27.0**	**8.3**

Table 3. Distribution of comments to categories (with a difference of more than 5% between paper and tool for at least one of the three result sets) in both Schools and the Teachers Programme (in %).

Category	**Percentage of comments**					
	School 1		**School 2**		**Teachers Programme**	
	Paper	**Tool**	**Paper**	**Tool**	**Paper**	**Tool**
Content	16.8	19.8	25.3	51.4	50.7	71.4
Add	5.3	11.3	6.8	2.9	27.5	26.5
Language	1.3	0	5.8	14.3	1.9	4.1
Remove	0.3	0	1.6	0	1.0	6.1
Terminology	3.0	0	4.2	22.9	2.4	8.2
Design	28.6	52.8	43.2	11.4	16.4	12.2
Colourful	5.9	9.4	8.4	2.9	1.4	0
Text layout	1.3	11.3	12.1	0	4.3	2.0
Functionality	41.4	17.9	15.8	20.0	22.7	14.3
Add	23.7	9.4	3.2	17.1	7.7	14.3
Missing description	6.9	3.8	11.1	2.9	7.7	0
Picture	0	1.9	3.7	5.7	7.2	0
Missing description	0	0.9	3.7	5.7	6.3	0
Unknown	13.2	7.5	12.1	11.4	2.9	2.0

Summing up, the empirical data of all the three PD workshops with students and teachers indicate that the paper-based approach was much more effective in terms of eliciting comments.

When looking at the percentage results per category presented in Table 3 most of the differences between paper and tool in one event are contradicted by the results of another event. Thus no advantage for paper or tool can be identified for most of the categories. Two exceptions are "Functionality – Missing description" and Unknown, where on a percentage basis (far) more comments have been created when the paper-based method was used to gather the feedback as when the tool-based method was used. For the missing description of functionality this might be explained by the paper being non-interactive and presenting the mock-up out of context (e.g. not on a computer screen; not in a browser). Thus users might have a harder time to identify the functionality of screen elements and therefore give the feedback that a description would be needed. The differences in "Unknown" can be partly explained by unreadable comments but more often by comments on the paper, where the target was unclear and therefore the problem could not be understood. We assume this happened less with the tool, because to give textual feedback there, the participants had to put a yellow sticky note, which they mostly put onto the screen element causing the issue, thus at least giving a hint regarding the target.

6.2 RQ2: the Quality of Comments

All comments categorized either as 'irrelevant' or 'unknown' (either not understandable or unreadable) were removed from further analysis, as they do not contain useful information for the designers and developers. Therefore 809 comments (632 paper- and 177 tool-based) were further analysed with regard to the three Attributes of CAt+.

Impact. As presented in Table 4, most of the changes proposed by the participants have an Impact of either 1 or 3 – affecting a single element or the whole page. This might imply that the participants tended to perceive the mock-up from a holistic perspective, although they zoomed in to explore specific elements (e.g. the next button) in detail. Regarding the comparison of paper and tool it can be seen that paper elicited (slightly) more feedback coded as 1 as compared to the tool results with this impact coding. Participants being more willing to give feedback on small details with paper, might imply that giving feedback with the tool needs slightly more effort, which was more likely spent on ideas with bigger impact.

Table 4. Distribution of Impact rating (in %).

Impact	Percentage of further analysed comments					
	School 1		School 2		Teachers Programme	
	Paper	Tool	Paper	Tool	Paper	Tool
0	15.9	10.2	9.6	0.0	7.5	6.3
1	28.4	22.5	40.7	32.3	32.3	27.1
2	1.5	1.0	7.2	25.8	9.5	14.6
3	35.2	58.2	31.1	32.3	32.8	33.3
4	18.9	8.2	11.4	9.7	17.9	18.8

Specificity. Table 5 presents the percentage of results were the corresponding sub-rating was clearly stated. When looking at the Target it can be seen that this is more specific in paper-based comments than tool-based. This may be explained by the fact that paper allows for a variety of ways to highlight a target (e.g. by drawing a circle around or an arrow pointing towards something, or by underlining text), where the tool used in the evaluation was restricted to a predefined set of feedback elements. Users mainly attached virtual sticky notes, which might explain the lower precision and could be approached by enhancing the functionality of the tool.

On the other hand the use of sticky notes might explain the higher specificity of the Reasoning found for the tool-based comments, as the text field on the note might have invited the participant to further elaborate.

Table 5. Comments coded as very specific for each sub-rating (in %).

	Percentage of further analysed comments					
	School 1		School 2		Teachers Programme	
	Paper	Tool	Paper	Tool	Paper	Tool
Target	86.1	65.3	78.0	61.3	83.6	68.8
Reasoning	23.7	41.8	26.8	32.3	32.3	45.8
Solution	26.8	20.4	6.6	22.6	40.8	41.7

Uniqueness. To determine how many duplicated ideas have been generated with the paper- and tool-based approach, the comments have been rated based on their uniqueness. The results are shown in Table 6.

When comparing the percentage of duplicates in the paper- and tool-based results, it can be noticed, that it is mostly higher for paper. If one assumes that there is a limited pool of possible ideas the participants can come up with, the higher total number of comments for paper also explains the higher percentage of duplicates as it becomes less likely to come up with a unique idea with an increasing number of comments. The exception of this for School 1 might partly be explained by having the highest number of comments for the tool-based method throughout the three events, but is still unexpected.

Table 6. Results of the Uniqueness rating.

	School 1		School 2		Teachers Programme	
	Paper	**Tool**	**Paper**	**Tool**	**Paper**	**Tool**
No. of comments	264	98	167	31	201	48
No. of **unique** comments	225	80	137	30	178	47
% of duplicates	14.8	18.4	17.9	3.3	11.4	2.1

Inter-rater reliability. As two researchers were involved in coding the participants' comments, weighted Cohen's kappa [33] was calculated to determine the inter-rater reliability for the different coding criteria. For the Categorization the weight was determined based on the agreement about the main- and sub-category. If only the sub-category differed, a weight of 1 was used (as there was at least agreement about the main category of the comment), if the main category differed, a weight of 2 was applied. For Impact and Specificity, the weight has been determined by the difference between higher and lower value. For Uniqueness, the standard weight was used. For all ratings the value of Weighted Cohen's kappa was above 0.7, ranging from 0.72 (for Uniqueness) to 0.88 (for Specificity – Solution). Although the kappa rating magnitude guidelines in the literature are inconsistent (e.g. [34, 35]), with all values being above 0.7 we are still confident that our results are reasonable or even good.

Pearson's $\chi 2$ analysis of category rating. Some inferential statistics on the results were performed to check whether the observed differences are significant. As we have categorical data, we used Chi square tests [36] to verify the null hypothesis (H_0):

H_0: the number of comments in each of the coding dimensions (categories, impact, specificity, uniqueness) is independent of the method used to elicit and capture them (paper or tool).

Table 7 shows the results. To get expected values larger than 5 (as required by the Chi square test, e.g. [36]) it was necessary to combine some of the results. For categories Picture, Unknown, and Irrelevant were combined to "Other". For Impact 0, 1, and 2 were combined to "less than page level" and 3 and 4 to "page level and above". For Specificity the results were combined into three groups, very specific (two or more sub-ratings that are very specific), specific (one sub-rating that is very specific), and unspecific (no sub-rating that is very specific). For Uniqueness no combination was possible, therefore the result for School 2 is included in Table 7 although the requirement for Chi square was not met, as the expected value for "not unique" in "tool" was less than 5 (~4.85). A Fisher's Exact test has therefore been performed for the latter, confirming the rejection of H_0 in this case.

As can be seen in Table 7 the results are only affirmative for Categories and Specificity ratings. For Categories, H_0 has to be rejected, meaning that the different meth-

ods influence the number of comments in different categories. For Specificity, H_0 is not rejected, meaning that the specificity of a comment is independent from the method used for feedback gathering. For Impact and Uniqueness the results of the Chi square tests are ambivalent.

Table 7. $\chi 2$ values for independence of number of comments per rating dimension of CAt+ on method (for all three events). Shading of cells indicates, where H_0 was rejected. *Contains one expected value less than 5!

	School 1	**School 2**	**Teachers Programme**
Categories	$\chi 2$ (3, n=412)=27.19, p<.001	$\chi 2$ (3, n=233)=15.51, p<.05	$\chi 2$ (3, n=266)=8.41, p<.05
Impact	$\chi 2$ (1, n=362)=4.32, p<.05	$\chi 2$ (1, n=198)=0.004, p>.05	$\chi 2$ (1, n=249)=0.03, p>.05
Specificity	$\chi 2$ (2, n=362)=3.52, p>.05 (ns)	$\chi 2$ (2, n=198)=1.96, p>.05 (ns)	$\chi 2$ (2, n=249)=4.44, p>.05 (ns)
Uniqueness	$\chi 2$ (1, n=362)=220.24, p>.05 (ns)	$\chi 2$ (1, n=198)=4.30, p<.05 *	$\chi 2$ (1, n=248)=20.57, p<.001

7 Conclusion and Future Work

When comparing the number of comments given, one can easily recognize that the paper-based approach created much more results than did the tool-based approach, even after normalizing it based on the number of participants or groups using the respective approach. The answer to RQ1 is then clear. One reason for this difference lies in the tool chosen for the evaluation. During the PD sessions with the students and teachers, it showed that some usability problems of myBalsamiq had undermined the efficiency of giving comments.

The results of the descriptive statistics suggest that no consistent trends or meaningful patterns in terms of the qualitative differences of the comments (i.e., Categorization and three Attributes) can be observed. The answer to RQ2 remains ambivalent.

Nonetheless, the number of students and teachers involved in the current study was relatively small. We are aware of this limitation. More studies, involving more participants would be needed to further substantiate the results presented in this paper.

Given the vast difference in number of comments gathered and the dependence of feedback gathered on the method used (at least for categories), we have to conclude that myBalsamiq cannot be used to replace the paper-based method. This interesting finding and the advantages we see in using a tool, have motivated us to develop a more usable PD online tool that can enhance the value of Participatory Design in general and be a valid substitute for the currently conventional and common paper-based method, to be confirmed by further evaluations.

Furthermore, our coding scheme CAt+ for PD data analysis can be useful for other researchers in this field and future comparison studies. Furthermore, we will focus not only on using the coded results to compare paper- and tool-based performance but also demonstrating how the coding can be applied by developers and designers to

make better sense of user feedback and address it more effectively. For instance, sorting the list of comments in descending order of Impact or Specificity rating could support them in dealing with the most important feedback first. This approach can also be automated as a kind of tool-supported content analysis in the future.

In summary, with the groundwork built as reported in this paper, our future work will comprise three main strands: (i) To further validate and substantiate our coding scheme CAt+ in different settings (i.e., co-located, distributed, individual, group-based) involving different stakeholders (teachers, students, researchers, designers, developers); (ii) To evaluate the downstream utility of coded PD results by examining whether and how such results support a development team in their redesign work; (iii) To develop a usable PD online tool enabling participants to give comments with ease and even fun.

Acknowledgements. This work was partially funded by the European Union in the context of the Go-Lab project (Grant Agreement no. 317601) under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D (FP7). This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

References

1. Greenbaum, J. and Kyng, M. Design at work: Cooperative design of computer systems. Hillsdale NJ: Erlbaum, 1991.
2. Muller, M. J. Participatory design: The third space in HCI (revised). In J. Jacko and A. Sears (eds.), Handbook of HCI (2nd Ed.). Erlbaum, 2007.
3. Schuler, D. and Namioka, A. Participatory design: principles and practices. 1993.
4. Sanders, E. B. N., Brandt, E., and Binder, T. A framework for organizing the tools and techniques of participatory design. Proc. of the 11th Biennial Participatory Design Conference (2010), 195–198.
5. Walsh, G., Foss, E., Yip, J., and Druin, A. FACIT PD: a framework for analysis and creation of intergenerational techniques for participatory design. In Proc. of CHI'13. ACM (2013), 2893-2902.
6. Weibel, N., Signer, B., Norrie, M. C., Hofstetter, H., Jetter, H. C., and Reiterer, H. PaperSketch: a paper-digital collaborative remote sketching tool. Proc. of the 16th international conference on Intelligent user interfaces. ACM (2011), 155-164.
7. Lin, J., Newman, M. W., Hong, J. I., and Landay, J. A. DENIM: finding a tighter fit between tools and practice for Web site design. In Proc. of the SIGCHI conference on Human factors in computing systems. ACM (2000), 510-517.
8. Rogers, Y., Sharp, H., and Preece, J. Interaction design: beyond human-computer interaction. John Wiley & Sons, 2011.
9. Sundar, S. S., Oh, J., Bellur, S., Jia, H., and Kim, H. S. Interactivity as self-expression: a field experiment with customization and blogging. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2012), 395-404.

10. Teo, H. H., Oh, L. B., Liu, C., and Wei, K. K. An empirical study of the effects of interactivity on web user attitude. International Journal of Human-Computer Studies, 58(3) (2003), 281-305.
11. Zhao, L. and Lu, Y. Enhancing perceived interactivity through network externalities: An empirical study on micro-blogging service satisfaction and continuance intention. Decision Support Systems, 53(4) (2012), 825-834.
12. Govaerts, S., Cao, Y., Vozniuk, A., Holzer, A. C., Garbi Zutin, D., San Cristóbal Ruiz, E., Bollen, L., Manske, S., Faltin, N. and Salzmann, C. Towards an Online Lab Portal for Inquiry-based STEM Learning at School. Proc. of ICWL 2013, 2013.
13. Read, J. C., Gregory, P., MacFarlane, S. J., McManus, B., Gray, P., and Patel, R. An Investigation of Participatory Design with Children – Informant, Balanced and Facilitated Design, Interaction Design and Children (2002), 53-64.
14. Hundhausen, C., Trent, S., Balkar, A. and Nuur, M. The design and experimental evaluation of a tool to support the construction and wizard-of-oz testing of low fidelity prototypes IEEE Symposium on Visual Languages and Human-Centric Computing, 2008. VL/HCC 2008. 2008, 86-90.
15. Segura, V. C. V. B., Barbosa, S. D. J. and Simões, F. P. UISKEI: A Sketch-based Prototyping Tool for Defining and Evaluating User Interface Behavior Proc. of the International Working Conference on Advanced Visual Interfaces, ACM (2012), 18-25.
16. MacDonald, F. and Miller, J. A Comparison of Tool-Based and Paper-Based Software Inspection. Empirical Software Engineering, Kluwer Academic Publishers, 3 (1998), 233-253.
17. Bailey, B. P. and Konstan, J. A. Are Informal Tools Better?: Comparing DEMAIS, Pencil and Paper, and Authorware for Early Multimedia Design. Proc. of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2003, 313-320.
18. Law, E. L.-C., van Schaik, P., and Roto, V. Attitudes towards user experience (UX) measurement. International Journal of Human-Computer Studies, 72(6) (2014), 526-541.
19. Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., and Herlocker, J. Toward harnessing user feedback for machine learning. In Proc. of the 12th international conference on Intelligent user interfaces (IUI '07). ACM, New York, NY, USA, 2007, 82-91.
20. Kindred, J. and Mohammed, S. N. "He Will Crush You Like an Academic Ninja!": Exploring Teacher Ratings on Ratemyprofessors.com. Journal of Computer-Mediated Communication, 10: 00 (2005).
21. Madden, A., Ruthven, I., and McMenemy, D. A classification scheme for content analyses of YouTube video comments Journal of Documentation, 69 (2013), 693-714.
22. Könings, K. D., Brand-Gruwel, S., and van Merriënboer, J. J. An approach to participatory instructional design in secondary education: an exploratory study. Educational Research, 52 (2010), 45-59.
23. Naghsh, A. M. and Andy, D. GABBEH: A tool to support collaboration in electronic paper prototyping. CSCW'04 the ACM Conference on Computer Supported Cooperative Work, Chicago, USA, 2004.
24. Walsh, G., Druin, A., Guha, M. L., Bonsignore, E., Foss, E., Yip, J. C., ... and Brown, R. DisCo: a co-design online tool for asynchronous distributed child and adult design partners. In Proc. of International Conference on Interaction Design and Children (IDC'11). ACM (2012), 11-19.
25. Newman, M. W., Lin, J., Hong, J. I., and Landay, J. A. DENIM: An informal web site design tool inspired by observations of practice. Human-Computer Interaction, 18(3) (2003), 259-324.

26. Walsh, G., Druin, A., Guha, M. L., Foss, E., Golub, E., Hatley, L., ... and Franckel, S. Layered elaboration: a new technique for co-design with children. In Proc. CHI'10. ACM (2010), 1237-1240.
27. Heintz, M., Law, E. L.-C., Govaerts, S., Holzer, A. and Gillet, D. Pdot: Participatory Design Online Tool, CHI '14 Extended Abstracts on Human Factors in Computing Systems, ACM (2014), 2581-2586 .
28. Druin, A. Cooperative inquiry: Developing new technologies for children with children. In Proc. of CHI'99. ACM Press, 1999, 592-599.
29. Hallgren, K. A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology, 2012, 23-34.
30. Krippendorff, K. Content analysis: An introduction to its methodology (2nd ed.). London: SAGE, 2004.
31. Vilbergsdottir, S. G., Hvannberg, E. T., and Law, E. L.-C. Assessing the reliability, validity and acceptance of a classification scheme of usability problems (CUP). Journal of Systems and Software, 87 (2014), 18-37.
32. Hornbaek, K. and Stage, J. The Interplay Between Usability Evaluation and User Interaction Design International Journal of Human-Computer Interaction, 21 (2006), 117-123.
33. Fleiss, J. L. and Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33 (1973), 613–619.
34. Altman, D. G. Practical statistics for medical research. London: Chapman and Hall, 1991.
35. Fleiss, J. L., Levin B., and Paik, M. C. Statistical methods for rates and proportions, 3rd ed. Hoboken: John Wiley & Sons, 2003.
36. Maltby, J. and Liza D. Early success in statistics. Pearson Education, 2002.
37. McDonald, J. H. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland, 2014.