Comparison of Paper- and Tool-based Participatory Design Approaches: A Case Study with PDotCapturer

Matthias Heintz
University of Leicester
University Road, Leicester
LE1 7RH, United Kingdom
mmh21@leicester.ac.uk

Effie Law
University of Leicester
University Road, Leicester
LE1 7RH, United Kingdom
lcl9@leicester.ac.uk

Pamela Andrade
University of Leicester
University Road, Leicester
LE1 7RH, United Kingdom
pyas2@leicester.ac.uk

In the current Participatory Design (PD) practice, paper-based approaches are commonly applied. However, the use of software tools can potentially enhance the PD process and outcome, given their flexibility for prototype presentation, data storage, retrieval and analysis. Driven by this assumption, we developed the specific-purpose PD online tool PDotCapturer. To evaluate the effectiveness of PDotCapturer, as compared with a paper-based approach, we conducted a within-subjects study where 33 participants were asked to provide feedback on e-learning prototypes, using either paper or PDotCapturer in one session and vice-versa in another session. The evaluation criteria were quantity and quality of the participants' comments. For qualitative data analysis, we applied the refined coding scheme CAt+. Results show that using PDotCapturer partly led to feedback of higher quality, but the number of comments was higher for the paper-based approach. Our work contributes to the HCI community by augmenting the repertoire of the PD toolbox.

Participatory Design, paper, tool, comparison, CAt+, PDotCapturer.

1. INTRODUCTION

Participatory Design (PD) aims to actively include end-users in the design and development process to ensure that the resulting product is not only tailored to their needs but also shaped by their input. This can, for example, lead to a stronger feeling of ownership and higher acceptance of a product by its prospective users. The PD approaches developed and applied are as varied as the things that can be designed, ranging from cities [4] to digital family calendars [21]. Even when restricting products to digital artefacts, the number of PD approaches that can be used is still high (e.g., [25]). But most of these approaches are paper-based (an overview is given in [31]).

Although the use of paper to gather feedback has a long history and is well established in PD research, there are several benefits of using a software tool to gather user feedback. They include relaxing from the constraint that a user and the researcher conducting the evaluation should be in the same location [10] and providing a more realistic experience, especially for prototypes of higher fidelity, as the user can evaluate a working prototype rather than a non-interactive paper print-out [28]. However, providing feedback using a software tool as opposed to using paper includes a learning curve for operating the functionality. This might result in different amount of feedback given. Another open question is whether the PD data elicited with a software tool are comparable with those elicited with its paper-based counterpart. This comparability has not yet been researched adequately. In a recent attempt, due to the lack of a dedicated PD tool, the comparison was done using a general-purpose online prototyping tool [12]. Inspired by the methodology and results of this attempt, we have created a dedicated PD online tool called PDotCapturer, and intended to compare it with a paper-based approach. Additionally, with the specific-purpose PD tool, more solid answers for the following research questions (RQs) can be obtained:

RQ1: *How does the amount of PD feedback gathered with a paper-based PD approach differ from that gathered with a PD software tool (i.e. a tool-based PD approach)?*

RQ2: *What are the qualitative differences in PD feedback gathered with a paper-based approach and that gathered with a tool-based approach?*

RQ3: *What are the effects of using a specific-purpose PD software tool instead of a generic prototyping software tool to perform PD activities?*

As part of a technology-enhanced learning research project, the PD activities reported in this paper have been conducted with prototypes for online lessons. They enable teachers and students to teach and learn scientific concepts with learning content being enhanced with digital artefacts such as online laboratories [5]. In the initial phase of the project, an off-the-shelf general-purpose prototyping tool, myBalsamiq [24], the online version of Balsamiq [3], was used to perform tool-based PD activities. Subsequently, we developed the dedicated PD online tool PDotCapturer based on user requirements, on insights gained from deploying myBalsamiq, and on interviews with developers [10].

To compare the PD feedback on interactive prototypes gathered with paper and PDotCapturer, we applied the coding scheme CAt+ (Categories plus Attributes). CAt+ was developed in [12] for analysing the quality of PD feedback on non-interactive prototypes captured with paper and myBalsamiq. By modifying the categories and attributes of CAt+ as required for coding the datasets presented in this paper, we improved its scope of applicability.

With the dedicated PD tool PDotCapturer, the enhanced coding scheme CAt+, the practical experience of applying the tool and the scheme, and the empirical evidence of their effectiveness, our work contributes to the HCI community by augmenting the repertoire of the PD toolbox.

2. RELATED WORK

Intrigued by the differences between paper and software tools and the possible advantages of tools for data gathering and data analysis in PD, we have been motivated to compare these two approaches. There have been several studies comparing paper-based methods with their tool-based counterparts in areas related to PD such as multimedia and software prototype design [2, 13, 17, 26], but to the best of our knowledge none in PD except [12]. Additionally, those studies mostly focused on quantitative data like the number of comments created and time expended. But quantitative results are not enough to adequately compare the outcome of paper-based and tool-based approaches. It is also important to consider the quality of PD comments. To quantify qualities and thus make them comparable, a coding scheme can be used to categorise and rate the comments. Several such schemes have been created in different research areas [15, 16, 18, 27], but they are usually domain-specific or even application-specific. As mentioned earlier, CAt+ is a recently developed coding scheme for PD feedback [12], be it captured with a paper-based or with a tool-based approach.

The study presented in [12] conducted PD activities on non-interactive mockups using the paper-based Layered Elaboration approach [30] and the software tool myBalsamiq [23]. Layered Elaboration is a PD approach that keeps the initial design and different iterations intact by overlaying them with acetates on which comments and re-design suggestions can be recorded instead of drawing on the prototype directly. Participants come up with an initial design and then hand it over to another user for further refinement. The refined prototype can then be further commented on by other participants by adding a new acetate layer each (Figure 1).

The results of comparing the paper-based and tool-based PD approaches reported in [12] were mixed and thus inconclusive regarding which of the two PD approaches was more effective. Hence, we aimed to contribute to the body of applied knowledge on this specific topic by conducting another study, with the goal of yielding a clearer picture and more conclusive findings when to use which approach.

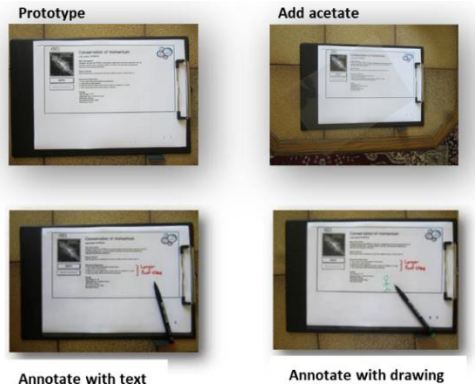

Figure 1: *Layered Elaboration.*

Additionally, [12] compared paper with a general purpose tool. Designers and developers can use myBalsamiq [23] to create mockups of the user interface of their software. Besides supporting the prototype creation process, myBalsamiq has some basic feedback functionalities for designers and developers to discuss alternative solutions. These can also be used by participants of PD activities to give feedback. However, myBalsamiq was not tailored for supporting PD. For instance, users are always presented with a full feature set of the tool instead of only a subset of features specific for the task at hand (giving feedback), and this can be overwhelming. We were thus motivated to develop PDotCapturer, assuming that this could improve the performance of the tool-based PD approach compared to paper. To verify our assumption, we conducted a comparison study similar to [12], and adapted its methodology to address more mature prototypes.

This adaptation was necessary as the digital artefacts to be evaluated in this study were of higher fidelity (more interactive) than those in [12]. Specifically, the paper-based approach had to be changed. Instead of exploring as well as annotating printouts of the non-interactive prototypes directly

[12], in this study participants worked through the interactive prototypes on their computer screen while using the corresponding paper printouts to provide feedback.

Another rationale for developing PDotCapturer is that none of the existing annotation tools could meet the key requirements for being a dedicated PD tool [10, 11]. For instance, GABBEH [20] is a tool enabling users to give comments by mimicking paper prototyping electronically. But it only works together with the DENIM tool [22] and is thus restricted for use in PD activities. DisCo [29] is a tool supporting the digital version of the paper-based Layered Elaboration PD approach [30], but it is not available for public use and only supports designing from scratch, not providing feedback on existing interactive prototypes.

3. DESIGN OF PDOTCAPTURER

We developed the web-based PDotCapturer using the GWT (Google Web Toolkit) [8], which transforms Java source code into an HTML5 and JavaScript website. For data management and storage a MySQL database is used.

While being a specific-purpose tool, PDotCapture is flexible regarding the artefacts to which it can be applied; they are essentially any websites or online contents. Figure 2 shows the view of PDotCapturer where participants can give feedback. On the top of the screen (❶) is a box which displays the instructions for the current PD activity specified by a researcher. Underneath the Instructions box the software artefact under evaluation is integrated (❷). Participants can freely interact with it. When they want to comment on something, they click on a 'Give feedback' button in the tool box in the top left hand corner of the screen (not visible in the current view in Figure 2). This causes the artefact under evaluation to become non-interactive by being covered by a transparent layer on which the participant can give feedback. This feature aligns with the idea of the Layered Elaboration approach [30] mentioned earlier. The change from interactive to non-interactive mode is indicated through a sticky-note-like cursor instead of the default mouse cursor. By clicking on the position of interest, a numbered yellow sticky note is placed there (❸) to keep track of different comments and their positions. At the same time the mouse cursor is changed into a blue pen, indicating that the participant can now freely draw on the artefact (e.g. making the suggestion of adding a concept map as drawn at ❹). Undo and Redo buttons are provided for the user's control and freedom when drawing (❺) (cf. Nielsen's heuristic [23]). In addition to the visual feedback a textual comment with further information or more details can be given (❻) and the participant's current emotion (positive, negative, or neutral) can be indicated with one of the three smileys (❼). After saving a comment the participant can either proceed to give more feedback by clicking somewhere else on the screen, creating another sticky note there, or can continue to interact with and evaluate the software artefact by clicking on the 'Finish giving feedback' button in the tool box (not visible in the current view in Figure 2). This will hide all feedback given so far and enable interaction with the artefact again.

The main characteristics of the tool are:

- Online
- Presentation of instructions ❶
- Integration of live websites ❷
- Indication of feedback position ❸
- Support of freehand drawing ❹
- Specification of emotion ❼

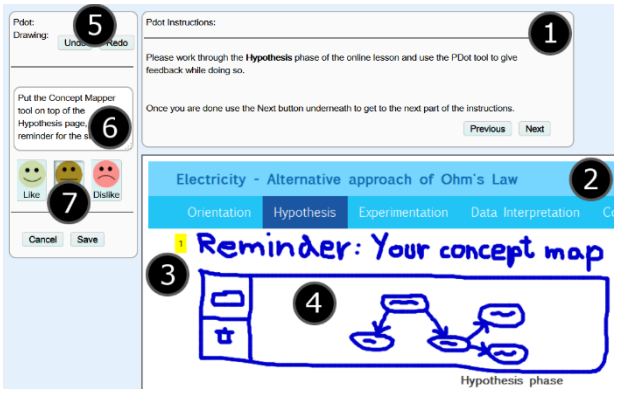

Figure 2: *Screenshot of PDotCapturer. Numbers added for references in text.*

4. DESIGN OF EMPIRICAL STUDY

The web-based interactive prototypes under evaluation were two online lessons. Those are websites to enable and support learning by presenting learning content (like text or YouTube videos), scaffolding apps (e.g., apps for creating hypotheses and for visualising data) and an online lab (which allows students to perform virtual experiments). The goal of the empirical study was to compare the outcome of PD activities using the paper-based and tool-based approach. The main difference to the existing work, especially to [12], is that we evaluated a dedicated PD tool in this study and compared it to a similar paper-based PD approach.

4.1. Paper-based and Tool-based PD Approach

For the paper-based evaluation of the interactive prototypes, 'evaluation booklets' have been created; they were printouts of individual webpages of the prototypes. As the online lessons are aimed at students rather than teachers the participating teachers were asked to adopt the role of their students when exploring the features of the

prototypes on their computers. For giving feedback, the participants had to annotate the corresponding printouts in their evaluation booklets.

For the tool-based evaluation, participating teachers were demonstrated how to use PDotCapturer with an example unrelated to online learning in order not to bias feedback. They were then asked to assume the role of students when working through the online lesson. They used PDotCapturer to give their comments (cf. the evaluation booklets in the case of the paper-based evaluation) while exploring the prototypes on their computers.

4.2. Participants and Procedure

The PD study was conducted in the context of a week-long teacher training event where the resources and instructions were in English. 33 science teachers from different European countries were involved. The study was implemented as two sessions, Workshop-A and Workshop-B, on two consecutive days. In Workshop-A participants worked with an online lesson on electricity whereas in Workshop-B they worked with an online lesson on buoyancy. These two specific online lessons were selected as artefacts for the study as they shared the same basic pedagogical structure and user interface design, differing mainly in the domain-specific content. The one day time gap helped mitigate the issue of memory bias.

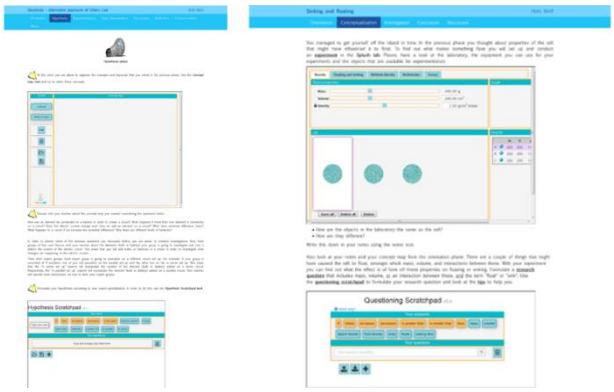

Figure 3: *Screenshots of Electricity (left) and Buoyancy (right) online lesson highlighting their similarity.*

To control the order effect, one-third of the participants used the paper-based approach and the rest used the tool-based approach in Workshop-A, and vice-versa in Workshop-B (see Table 1). The uneven split was due to the fact that the participants were free to choose one of the two approaches in Workshop-A; the free choice was aimed to motivate them. Probably due to curiosity, the majority opted for using PDotCapturer although they were informed that they would have to use the other medium in Workshop-B. Furthermore, owing to the constraints of the infrastructure (i.e. number of computers available in the computer lab in which the study was performed and slow internet connectivity), the participants were asked to work in self-selected pairs (except one single person because of the odd number of participants). On average, the participants spent 47.5 minutes in interacting with the respective online lesson and giving feedback. The tight time schedule for the workshop slots of the teacher training event did not allow us to conduct any post-study interviews or questionnaires to gather information on their preference for the approaches. Due to the privacy concern, the demographic data of the participants were not collected.

Table 1: *Distribution of participants (P1-P33) in the two approaches in the two workshops.*

	Workshop-A: Electricity Lesson	**Workshop-B: Buoyancy Lesson**
Paper-based	*5 pairs (P1-P10)*	*11 pairs + 1 single (P11-P33)*
Tool-based	*11 pairs + 1 single (P11-P33)*	*5 pairs (P1-P10)*

This setup enabled us to perform two within-subject comparisons, taking into account the fact that the design of the Electricity and Buoyancy online lessons were highly similar (Figure 3):

- Comparison 1 involved the 5 groups (P1-P10) who used paper in Workshop-A and then used the tool in Workshop-B.
- Comparison 2 involved the 12 groups (P11-P33) who used the tool in Workshop-A and then paper in Workshop-B.

5. DATA ANALYSIS

To facilitate the process of data analysis, raw data have been digitized from the paper-based approach or exported from the tool-based approach to spreadsheets. The former took about 8.5 hours whereas the latter took only a few minutes. This showed clearly one of the advantages of the tool-based approach. The data have then been coded in a fully crossed design [9] by two HCI researchers with about one and seven years of experience in usability research.

5.1. CAt+ and adaptations of the coding scheme

5.1.1. Categories

The Categories coding of CAt+ consists of a combination of a main category with a sub-category. The following list gives definitions and examples for the six main categories:

- **Content:** Comment on the information presented, e.g. “Replace this by a cartoon explaining the text!”.
- **Design:** Comment on the visual representation of the prototype as a whole or parts of it; e.g. “Should be bigger”.

- **Functionality:** Comment on the features of the prototype as a whole or parts of it; e.g. “I would like to have more […] elements like transistors, cepecitors, motors, inductance”.
- **Picture:** Comment on images; e.g. “centre pictures”.
- **Unknown:** Obscure comment, e.g. “ciareainobele”.
- **Irrelevant:** Comment which does not address the prototype or parts of it; e.g. “This type of experiment needs to be done first really with real objects and real liquids!”.

As the main category only gives a general classification of the comment, it is combined with a more specific sub-category to define the nature of a comment with finer details. For example, two Content comments: “insert a short definition of this term here” and “delete this sentence” are coded as Content-Add or Content-Remove, respectively. Due to space restrictions and to avoid repetition, we refer to [12] for definitions, examples, and additional information on the sub-categories.

Results of coding the current datasets support the applicability of the CAt+ Categories and Sub-Categories compiled and described in [12] to datasets other than the ones coded there. As already discussed in [12], some new combinations of main and sub-categories would probably be required when CAt+ was to be applied in other contexts. We did indeed need several such modifications for our datasets. In addition, we created one new subcategory: Fix. This was used in combination with the main category Functionality as ‘Functionality-Fix’ to code comments where the participant reported an issue and requested to resolve it.

The addition of the combination Functionality-Fix to the coding scheme can be justified based on the evaluated artefacts: With more actual prototypic functionality in the software artefacts evaluated, the participants were now able to voice their opinions where something did not work in the way they would have expected it or preferred. To reflect this change in the fidelity of the artefacts evaluated, the CAt+ coding scheme was enhanced accordingly.

Although six combinations of main and subcategory present in the initial CAt+ coding scheme did not occur while coding the current datasets, they were not removed. Their non-usage can again be explained by the nature and fidelity of the artefacts evaluated. Thus to be applicable for coding feedback on prototypes of different or increasing fidelity, CAt+ was augmented by adding new sub-categories without removing the existing ones.

Only 4 of the 339 comments (0.01%) in our datasets were initially coded as ‘Picture’. As pictures are either used as design elements or learning content, we argue that the CAt+ coding scheme can be simplified and thus improved by removing the main category Picture. Those comments can be put either in the Content or Design category. Accordingly, we re-coded the 4 Picture comments.

The word ‘Unknown’ has the implication that something is yet to be discovered, whereas the meaning of this Category is more in the direction of ‘Unintelligible’ or ‘Obscure’. We therefore propose to rename this main category to ‘Incomprehensible’. This use of a clearer term improves the usability of the CAt+ coding scheme, especially for novice coders.

The Categories and Sub-Categories of the CAt+ coding scheme after adaptation while coding the datasets presented in this paper can be seen in Figure 4.

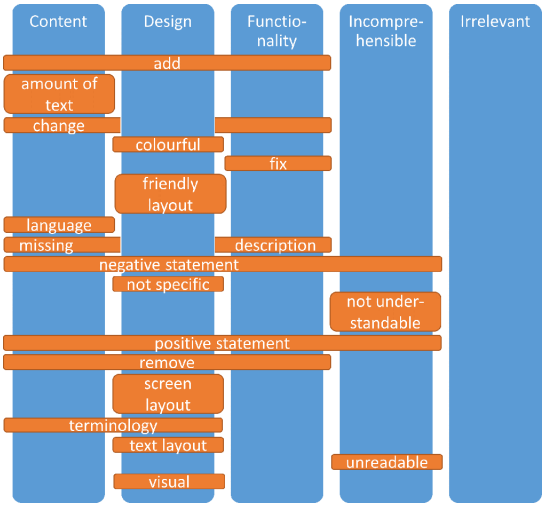

Figure 4: *CAt+ categories and sub-categories after adaptation based on the datasets analysed for this paper.*

5.1.2. Attributes

CAt+ comprises three attributes: Impact (five-level), Specificity (three four-level sub-attributes), and Uniqueness (dichotomous values). Each attribute will be elaborated in the following.

The Impact attribute of CAt+ is concerned with the question: how much of the user interface and interaction design is going to change if this comment is addressed by designers and developers. The effect of the user’s comment on the prototype is measured as one of the following 5 levels:

- **0:** As the comment does not suggest a change, there would be no effect of considering it or not.
- **1:** Addressing this comment would result in a localised change of a single element on the current screen, without affecting the page surrounding this element.
- **2:** Addressing this comment would result in several localised changes of several elements on the current screen, without changing the page outside of those elements affected.

- **3:** Addressing this comment would result in a change of the whole page or at least a big portion of its layout.
- **4:** Addressing this comment would result in a change of several existing pages or the need to create new ones. The changes visible for the user would thus go beyond the current page.

As the existing levels were sufficient to code the current datasets, no adaptation of the CAt+ attribute Impact was needed.

The *Specificity* attribute of CAt+ is concerned with the question how clear a comment is regarding three dimensions, resulting in the following three sub-attributes:

- **Specificity-Target:** The first dimension is used to rate how clearly the comment defines the element with which it is concerned.
- **Specificity-Problem / -Reasoning:** The second dimension is used to rate how clearly the issue encountered is expressed or, for positive comments, how clear the explanation, why this is good.
- **Specificity-Solution:** The third dimension is used to rate how clearly the change proposed by the participant is expressed.

The initial Specificity ratings were considered as being unnecessarily complicated and cumbersome as each of the three dimensions had two sub-ratings: 'stated' and 'guessability'. The rating 'stated' evaluates if the comment text contains the information whereas 'guessability' evaluates if the coder can deduce the information from comment text and context. To improve CAt+ by simplifying it, the two sub-ratings were combined to a single Specificity rating for each dimension. It has the following four levels:

- **0:** not stated and not even guessable
- **1:** not stated, but guessable
- **2:** somewhat stated
- **3:** clearly stated

The *Uniqueness* attribute of CAt+ is concerned with the question how many distinctive comments a data gathering approach evoked. Every comment that solely repeats the idea of a previously coded one is rated as 0 (= duplicate). The rating of the initial comment is kept as 1 (= unique comment), because otherwise encountering a duplicate would remove this idea completely from the data pool. But as the goal is to count distinctive ideas, only the ones after the initial mentioning are 'filtered out' by coding them as duplicates of an existing comment. The existing CAt+ codes were again sufficient and therefore there was no need to adapt the coding scheme.

6. RESULTS AND DISCUSSIONS

6.1. RQ1: Quantity of Comments

Regarding the quantity of comments both comparisons show a similar trend: Using the paper-based approach resulted in more comments being gathered compared to the tool-based approach. This is true for total number of comments (see Table 2) as well as for comments per group (see Table 3). For Comparison 1 the total number of comments dropped from 68 to 25, in Comparison 2 from 125 to 121.

Table 2: *Total number of comments.*

	Workshop-A: Electricity Lesson	**Workshop-B: Buoyancy Lesson**
Paper-based	*68 (5 pairs)*	*125 (11 pairs + 1 single)*
Tool-based	*121 (11 pairs + 1 single)*	*25 (5 pairs)*

Table 3: *Number of comments per group.*

	Paper-based	**Tool-based**
Comparison 1	*68 / 5 = 13.60*	*25 / 5 = 5.00*
Comparison 2	*125 / 12 = 10.42*	*121 / 12 = 10.08*

One explanation for the big drop in Comparison 1 might be attributed to the fact that participants could freely choose the approach they would like to use in Workshop-A, but were asked to use the other one in Workshop-B. Participants who opted for the tool-based approach in the first workshop were apparently more motivated to use technology to provide feedback than those who opted for the paper-based approach in Workshop-A. Subsequently the number of comments dropped substantially when the participants who did not choose the tool-based approach in Workshop-A had to use PDotCapturer in Workshop-B. The observation that the participants provided less feedback with the tool might be caused by their unease or low confidence in using technology. This supposition, however, cannot be verified because no demographic data of participants were collected.

An explanation why paper generally resulted in more comments could be that even when a tool has good usability, paper is still more intuitive to use. With PDotCapturer the participants needed to learn how to use it and get familiar with giving feedback through it. On the other hand, how to write and draw on paper is already familiar to the participants so they can start right away with providing feedback and they might also be quicker in providing written feedback on paper, depending on their speed of handwriting as compared to their typing speed.

These are general issues of tool over paper use, but they can be mitigated by making a tool easier to use.

Hence, quantitatively the paper-based approach seems to be more powerful than the tool, but it looks like the motivation of the participants to use the tool-based approach has played a role in the difference in the amount of feedback gathered with the two approaches.

6.2. RQ2: Quality of Comments

6.2.1. Data processing to compare paper- and tool-based feedback

To account for the differences in number of comments gathered using the two different approaches we normalised the data for the qualitative analysis. We did so by dividing the number of comments assigned to the individual categories and attributes levels by the number of comments gathered in the respective workshop and with the respective approach. The following tables show the percentages of the respective sub-totals.

To identify whether one of the two approaches resulted in significantly more comments in one of the CAt+ Categories or Attributes a Pearson's χ^2 analysis was performed after the coding was finished.

6.2.2. Categories

As shown in Table 4, both comparisons show consistent trends for some categories and are inconsistent for others. The numbers of Content and Design comments gathered were higher when the tool-based approach was used, while the number of Functionality comments was higher when the paper-based approach was used. The results for the two categories Incomprehensible and Irrelevant were inconclusive, because the trends went in opposite directions. In addition, as neither of these two categories would be useful for further analysis, they were grouped and referred to as "Inapplicable".

Table 4: *Percentage of comments per main category.*

	Comparison 1		**Comparison 2**	
	Paper	*Tool*	*Paper*	*Tool*
Content	*23.53*	*40.00*	*27.20*	*35.54*
Design	*4.41*	*24.00*	*8.00*	*13.22*
Functionality	*54.41*	*20.00*	*44.00*	*36.36*
Inapplicable	*17.64*	*16.00*	*20.80*	*14.88*

The consistent trend of having a bigger share of Content comments with the tool compared to paper in both comparisons supports our assumption that the two online lessons selected for the evaluation and comparison are highly similar despite being on different topics. If the topic would have influenced the likeliness to provide a comment on the content of the online lesson the data would be show opposing trends.

When looking at the distribution of comments in the three main categories that showed consistent trends (Content, Design, and Functionality) it can be seen that the tool created a somewhat more equal distribution of comments to categories than paper (see Table 4).

A possible explanation for the differences in the distribution of Content, Design, and Functionality comments might lie in the prototype presentation. The non-interactiveness of the printouts on paper might have triggered the participants to question how interaction elements work, why they worked this way, and how they expected them to work. With the interactive presentation of the prototype in PDotCapturer and the option to give feedback without a switch of the medium, the participants might have been enabled to follow the online lesson more smoothly, therefore focusing more on its content and noticing design issues. Additionally, the separation of the online lesson onto different screenshot pages in the evaluation booklet might have caused the participants to focus more on details and separate interaction elements rather than the design and content of the complete page.

Although the differences between paper and tool regarding the CAt+ categories are interesting and can be explained, no significant association was found between the approach and CAt+ category ($\chi^2(4) = 8.053$, $p > 0.05$).

As it is not meaningful to do a CAt+ Attributes rating for Inapplicable comments, the subsequent results only include comments that were coded as Content, Design, or Functionality.

6.2.3. Impact

The fact that Impact level 2 was not present in three of the four datasets coded for this paper implies that it might be feasible to merge it either with level 1 or 3. Thus we went back to the comments coded as having an impact of 2. As they all affected the prototype on a local level of individual elements without affecting the whole screen, we decided that for our current datasets Impact 1 and 2 can be merged. When comparing this finding with the results presented in [12], we see a similar trend. But the trend in the datasets presented there is not as strong as in our current datasets. Hence, we need more evidence to decide on retaining or removing this Impact level from the CAt+ coding scheme. Nevertheless, we would suggest considering the possible merge of Impact level 1 and level 2 when coding PD results with CAt+ in the future.

When checking the CAt+ Impact rating results (Table 5) for trends, it can again be seen that the two comparisons show some consistent trends and some inconsistent ones. The number of comments

with an Impact rating of 0 or 3 is higher with the tool whereas the number of comments with an Impact rating of 1 & 2 is higher when using the paper-based approach. The results for Impact rating of 4 are inconclusive.

Table 5: *Percentage of comments per Impact level.*

	Comparison 1		**Comparison 2**	
	Paper	*Tool*	*Paper*	*Tool*
0	*41.07*	*57.14*	*24.24*	*30.10*
1 & 2	*44.64*	*19.05*	*49.49*	*22.33*
3	*10.71*	*23.81*	*20.20*	*39.81*
4	*3.57*	*0.00*	*6.06*	*7.77*

By actively asking participants about their mood (see ❼ in Figure 2), the tool might elicit more comments stating generic positive or negative feelings towards screen elements. As such comments are helpful to get a general idea what participants like or dislike, but do not have any immediate influence on the prototype re-design, they would be rated as having an impact of 0.

The especially high amount of 0 Impact comments for the tool in Comparison 1 might be explained by the motivational issue described earlier. When the participants who initially chose the paper-based approach in Workshop-A had to use the tool in Workshop-B they might have been less motivated to give feedback and therefore have produced less well-thought and impactful comments.

A possible explanation why the paper-based approach gathered more localised (1 & 2) comments addressing single elements, where the tool-based approach gathered more global (3) comments addressing the whole page could again be the presentation of the online lesson. Splitting it up as different screenshot pages in the evaluation booklet might have led to the participants focusing more on details rather than the complete page. Although in cases covering multiple printout pages they would also have to scroll in the prototype, this might have been perceived as one continuous page (with scrolling) rather than separated sections of a page (navigated through by scrolling). The triggering of comments targeting several pages might be independent of the approach used as neither of the approaches can give an overview over several pages at the same time.

In general, the percentages are inversely proportional to the level of Impact, suggesting that participants tended to consider the design issues locally. This can be explained by the fact that at any given time only a part of the prototype is visible on the screen. Thus the attention of the participants is focussed on sections rather than a holistic overview. This observation can further be explained by the effort it takes to come up with feedback of different Impact levels. Noticing an issue resulting in an improvement suggestion that affects only a single screen element (Impact 1) is most of the time quicker and more straightforward than thinking of and expressing an idea affecting several pages (Impact 4). For example compare the comment 'this tool needs an undo button [undo button drawn]' with the comment 'teachers should be able to put a quiz at the end of each page, making sure that the student has understood the learning content presented on this page before moving on' (actual feedback of participants slightly rephrased to make it more comprehensible).

The chi-square results for Impact are significant ($\chi^2(3) = 17.723$, $p < 0.05$), thus the presented changes can be attributed to the approach used to gather the feedback.

6.2.4. Specificity

The results for the Specificity rating (Table 6) are based on the three sub-ratings. For instance, it is less helpful to have a clear target but unspecific problem description and solution compared to a clear target together with a clear problem description and a well described proposed solution. Therefore three possible levels of Specificity for a comment are differentiated:

- '**unspecific**': All sub-ratings have a value of less than 3.
- '**specific**': Only one sub-rating has a value of 3.
- '**very specific**': Two or more sub-ratings have a value of 3.

Table 6: *Percentage of comments per Specificity level.*

	Comparison 1		**Comparison 2**	
	Paper	*Tool*	*Paper*	*Tool*
unspecific	*1.79*	*13.64*	*15.15*	*4.85*
specific	*71.43*	*50.00*	*45.45*	*57.28*
very specific	*26.79*	*36.36*	*39.39*	*37.86*

The result of chi-square for Specificity is significant ($\chi^2(2) = 6.808$, $p < 0.05$). But as the CAt+ rating results (Table 6) are inconclusive, no point can be made regarding the relation between approach and Specificity of the comments gathered.

6.2.5. Uniqueness

When comparing the CAt+ Uniqueness rating results (Table 7) it can be derived that the tool-based approach resulted in fewer duplicates than the paper-based approach.

This can partly be explained by the higher number of comments gathered with the paper-based approach over the tool-based approach. The higher the number of comments, the higher the chance of getting duplicates (assuming the pool of possible

ideas is limited and even more so for ideas that are obvious).

Table 7: *Percentage of comments per Uniqueness value.*

	Comparison 1		**Comparison 2**	
	Paper	*Tool*	*Paper*	*Tool*
0	*7.14*	*0.00*	*11.11*	*6.80*
1	*92.86*	*100.00*	*88.89*	*93.20*

Another possible explanation for the paper-based approach gathering more duplicates might be that the repetition of similar looking pages on the printouts might have elicited the same feedback several times.

Table 8: *Number of unique comments.*

Comparison 1		**Comparison 2**	
Paper	*Tool*	*Paper*	*Tool*
52	*21*	*88*	*96*

As the paper-based approach resulted in more duplicates than the tool-based approach, the quantitative performance of the tool compared to paper improves when comparing how many unique comments were gathered (see Table 8). Paper is still more than twice (2.48) as effective than the tool in Comparison 1, but the tool-based approach exceeds the paper-based approach for Comparison 2.

For Uniqueness the chi-square result ($\chi^2(1) = 1.158$, $p > 0.05$) was not significant. Thus although there were differences in the number of unique comments, they were independent of the approach used to gather them.

6.2.6. Inter-rater reliability

To evaluate the inter-rater reliability of the two coders who coded the data, weighted Cohen's kappa [9] was calculated.

Table 9: *Weighted Cohen's kappa to determine inter-rater reliability.*

	Weighted Cohen's kappa
Categories	*0.76*
Impact	*0.85*
Specificity-Target	*0.60*
Specificity-Reasoning	*0.70*
Specificity-Solution	*0.79*
Uniqueness	*0.79*

For the Categories a weight of 2 was applied when main and subcategory differed and a weight of 1 if there was at least agreement on the main category but differing sub-categories assigned. For Impact and the three Specificity sub-ratings the difference between higher and lower value assigned was used as the weight. For Uniqueness the standard weight was applied. The results are shown in Table 9.

The guidelines on kappa rating magnitudes are not consistent in the literature (e.g. [1, 6, 7]). But with a result of 0.6 or higher (mostly 0.7 or higher) for each rating dimension, it is reasonable to assume that our inter-rater reliability is good.

6.3. RQ3: Comparison of tools

Although there is still a noticeable difference in the number of comments gathered with the different approaches in Comparison 1 (i.e., 2.72 times between paper and tool), it is a vast improvement over the difference reported in the previous study [12] (cf. 3.69 times between paper and tool). For Comparison 2 the number of comments gathered is nearly equal with paper-based and tool-based approach. With regard to RQ3, this implies that using a more appropriate tool for the task results in a noticeable improvement in terms of the number of comments gathered. As no consistent trend between myBalsamiq and PDotCapturer regarding the CAt+ Category coding can be seen, it can be assumed that switching the tool did not have an influence on the type of comments gathered.

A caveat is that the participants of the study and artefacts evaluated were different from their counterparts in [12]. Thus there might be factors other than the change of the tool (from myBalsamiq to PDotCapturer) influencing the number of comments gathered. Nevertheless, it is reasonable to argue that deploying a dedicated PD tool can contribute at least partially to the improvements observed.

Although the numbers cannot be compared directly when comparing the qualitative results for myBalsamiq reported in [12] with those for PDotCapturer reported here, some consistent trends can be seen for the different CAt+ attributes.

Using PDotCapturer resulted in more comments with an Impact rating of 0. As myBalsamiq, like paper, does not encourage specifying emotions, this difference can be explained in the same way as above. The trends for Impact ratings of 1(& 2) and 3 are much stronger in the datasets presented here. This can be explained by the dedication of the tool leading to PD-specific functionality, which emphasizes the differences between paper and tool. As the findings are inconclusive regarding Impact of 4 for both tools, this supports the assumption that the amount of comments with this level might be independent of the approach used to gather them.

On the contrary to the findings reported for myBalsamiq, where the paper consistently resulted in a higher number of very specific Specificity-Target comments, the results for PDotCapturer are inconclusive (and thus because of space restrictions not presented in detail). The increase in target

specification when using PDotCapturer can be explained by the improved way to indicate the position of a comment in PDotCapturer over myBalsamiq.

Regarding the Uniqueness the difference in duplicates between paper and PDotCapturer is lower than reported for paper and myBalsamiq. This can be explained by the comparable total number of comments gathered between paper and tool with PDotCapturer.

7. LIMITATIONS

One limitation of the work presented here is the fact that the findings are based on only two workshops. But from the given number of comments gathered we conclude that the results are still reliable. Nevertheless more studies should be conducted to further support the findings presented here.

Another limitation is caused by the lack of demographic data about the participants or their preference for the two approaches, which could have been capture with a post-study survey. Such data could have helped us substantiate some of the explanations given in the discussion of the results.

8. CONCLUSION

From the quantitative results presented in this paper it can be concluded that paper still results in more feedback than a tool. Regarding the question if the paper- or tool-based approach should be used to perform PD activities, one should say that paper still outperforms the tool.

From the qualitative results presented it can be concluded that paper and tool differ regarding two of the three qualitative attributes. How large the impact of a feedback is and how many duplicates are created is influenced by the approach used. The results regarding the specificity of comments is inconclusive. Although not significant the data presented in this paper shows that there seems to be a trend for paper to elicit more comments regarding functionality, whereas the tool results in proportionately more comments on the content and design.

Regarding the robustness of the CAt+ coding scheme it can be said that it is pretty complete, as only one new sub-category had to be added when coding the comments in the current datasets. And given that CAt+ was initially developed to code feedback on mock-ups with limited functionality, it makes sense that fixing functionality did not appear back then, but it is now when actual prototypes are evaluated. That some of the sub-categories did not occur but therefore different combinations of main and sub-categories were applied supports the initial idea of having all those sub-categories and allow them to be freely combined with any of the main categories as needed. Nonetheless, the granularity of the categories and attributes should not be too fine as to undermine the usability and utility of the scheme.

We presented several improvement suggestions for CAt+ to make the coding scheme easier to understand and apply in the future. However, from our experience of using it to code the feedback on different software artefacts it can be concluded that CAt+ is suitable to be generally used to rate PD feedback, be it on paper or gathered using a tool.

9. FUTURE WORK

Participants, who had used PDotCapturer to give feedback, also provided suggestions on how to improve it. These will be considered for future iterations of the tool. The findings presented in this paper are currently mostly based on the datasets of comments gathered in the study presented here. To substantiate these findings additional studies evaluating different software artefacts and collecting additional data (e.g. demographical information and subjective opinions) should be conducted. Furthermore, theoretical frameworks (e.g., naturalness of interaction [14]) need to be identified to enhance the understanding of differences between paper and tool.

The hardware available for the current study constrained PDotCapturer users to mouse-based interaction. However, the naturalness of touch-based interaction, which would bring giving feedback using PDotCapturer closer to the paper-based experience, might elicit very different results. It would thus be interesting to compare PDotCapturer on a touch-screen device with a paper-based approach.

Finally, CAt+ should be applied to other datasets to be collected in different contexts so as to further validate its robustness and generalizability.

10. 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) and the Next-Lab project (Grant Agreement no. 731685) under the Horizon 2020 research and innovation programme. 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. Altman, D. G. Practical statistics for medical research. Chapman and Hall, London, 1991.
2. Bailey, B. P. and Konstan, J. A. Are informal tools better?: comparing DEMAIS, pencil and paper, and authorware for early multimedia design. In Proc. of SIGCHI conference on human factors in computing systems, ACM Press (2003), 313-320.
3. Balsamiq. Rapid, effective and fun wireframing software. | Balsamiq. https://balsamiq.com/
4. Crewe, K. The Quality of Participatory Design: The Effects of Citizen Input on the Design of the Boston Southwest Corridor. Journal of the American Planning Association 67, 4 (2001), 437-455.
5. de Jong, T., Sotiriou, S. and Gillet, D. Innovations in STEM education: the Go-Lab federation of online labs. Smart Learning Environments 1, 1 (2014), 1-16.
6. 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.
7. Fleiss, J. L., Levin, B. and Paik, M. C. Statistical methods for rates and proportions, 3rd ed. Hoboken, John Wiley & Sons, 2003.
8. GWT Project. http://www.gwtproject.org/
9. Hallgren. K. A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in quantitative methods for psychology 8, 1 (2012), 23-34.
10. Heintz, M., Law, E. L.-C., Govaerts, S., Holzer, A. and Gillet, D. Pdot: participatory design online tool. In CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14), ACM Press (2014), 2581-2586.
11. Heintz, M., Law, E. L.-C. and Heintz, S. Review of Online Tools for Asynchronous Distribute Online Participatory Design. In: Dominik Renzel, Ralf Klamma (eds.), "Large-Scale Social Requirements Engineering", IEEE Computer Society Special Technical Community on Social Networking E-Letter, vol. 2, no. 3, September 2014, 2014.
12. Heintz, M., Law, E. L.-C. and Soleimani, S. Paper or Pixel? Comparing Paper- and Tool-Based Participatory Design Approaches. In Human-Computer Interaction – INTERACT 2015, Springer International Publishing (2015), 501-517.
13. 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. In Visual Languages and Human-Centric Computing, 2008. VL/HCC 2008. IEEE Symposium on, IEEE (2008), 86-90.
14. Jacob, R. J. K., Girouard, A., Hirshfield, L. M., Horn, M. S., Shaer, O., Solovey, E. T. and Zigelbaum, J. Reality-based interaction: a framework for post-WIMP interfaces. In Proc. CHI 2008, ACM Press (2008), 201-210.
15. 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, 3 (2005).
16. Könings, K. D., Brand-Gruwel, S. and van Merriënboer, J. J. G. An approach to participatory instructional design in secondary education: an exploratory study. Educational Research 52, 1 (2010), 45-59.
17. Macdonald, F. and Miller, J. A comparison of tool-based and paper-based software inspection. Empirical Software Engineering 3, 3 (1998), 233-253.
18. Madden, A., Ruthven I. and McMenemy, D. A classification scheme for content analyses of YouTube video comments. Journal of documentation 69, 5 (2013), 693-714.
19. Maltby, J. and Day, L. Early success in statistics. Pearson Education, 2002.
20. Naghsh, A. M. and Dearden, A. GABBEH: A tool to support collaboration in electronic paper prototyping. CSCW 2004, 2004.
21. Neustaedter C. and Bernheim Brush, A. J. "LINC-ing" the family: the participatory design of an inkable family calendar. In Proc. CHI 2006, ACM Press (2006), 141-150.
22. 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.
23. Nielsen, J. (1994a). Heuristic evaluation. Usability inspection methods, 17(1):25–62.
24. Painless Remote UX | myBalsamiq. https://www.mybalsamiq.com/
25. Sanders, E. B.-N., Brandt, E. and Binder, T. A framework for organizing the tools and techniques of participatory design. In Proc. PDC 2010, ACM Press (2010), 195-198.
26. 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. In Proc. of the International

Working Conference on Advanced Visual Interfaces, ACM Press (2012), 18-25.

27. 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, ACM Press (2007), 82-91.
28. 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.
29. Walsh, G., Druin, A., Guha, M. L., Bonsignore, E., Foss, E., Yip, J. C., Golub, E., Clegg, T., Brown, Q., Brewer, R., Joshi, A., and Brown, R. DisCo: a co-design online tool for asynchronous distributed child and adult design partners. In Proc. of the 11th International Conference on Interaction Design and Children, ACM Press (2012), 11-19.
30. Walsh, G., Druin, A., Guha, M. L., Foss, E., Golub, E., Hatley, L., Bonsignore, E. and Franckel, S. Layered elaboration: a new technique for co-design with children. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press (2010), 1237-1240.
31. 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. CHI 2013, ACM Press (2013), 2893-2902.