Getting the Right Design and the Design Right: Testing Many Is Better Than One

Maryam Tohidi
University of Toronto
Toronto, Canada
mtohidi@dgp.toronto.edu

William Buxton
Microsoft Research
Toronto, Canada
bill@billbuxton.com

Ronald Baecker
University of Toronto
Toronto, Canada
rmb@kmdi.utoronto.ca

Abigail Sellen
Microsoft Research
Cambridge, UK
asellen@microsoft.com

ABSTRACT
We present a study comparing usability testing of a single interface versus three functionally equivalent but stylistically distinct designs. We found that when presented with a single design, users give significantly higher ratings and were more reluctant to criticize than when presented with the same design in a group of three. Our results imply that by presenting users with alternative design solutions, subjective ratings are less prone to inflation and give rise to more and stronger criticisms when appropriate. Contrary to our expectations, our results also suggest that usability testing by itself, even when multiple designs are presented, is not an effective vehicle for soliciting constructive suggestions about how to improve the design from end users. It is a means to identify problems, not provide solutions.

Author Keywords
Design, Prototyping, Usability Testing, Evaluation, Methods, User Centered Design, Participatory Design.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The use of low-fidelity and paper prototypes is now well established in the design of commercial user interfaces [11, 12, 14]. This is largely due to their relatively low cost, coupled with the results of a number of researchers [2, 13, 15, 16] who have found that the usability data that they got from low and high fidelity prototypes were comparable. Hence, this type of instrument can provide a means to gain early insights into a design before the size of the investment prevents changes being made.

CHI 2006, April 22-27, 2006, Montréal, Québec, Canada.

Much of the often cited literature [11, 14] emphasizes the use of paper prototypes in usability testing [8]. The primary benefit in this case is to provide an inexpensive way to refine a design earlier in the process than would otherwise be possible. In this, they serve as an aid in *getting the design right.*

Another aspect of the relatively low cost of paper prototypes is their potential to enable the early exploration of more design alternatives than would otherwise be affordable (in time and money). Taking these two things together, an underlying question in our research is, "Can exposing users to multiple design alternatives also help us in *getting the right design?"* Besides helping us improve the usability of any particular design, can they also help us explore alternative designs?

Much of the often-cited literature on paper prototyping [11, 13, 14] focuses almost exclusively on the former. However, there is some literature on "parallel design" where different teams independently work on the same problem [7, 8, 9, 10], but this only touches on what we are interested in with the latter. Our experience in the traditional design arts, such as industrial design, graphic design and architecture, is that the simultaneous investigation of multiple alternatives by the same designer or team and the exploration of alternative designs pervades all stages of the process. The following quote from the VP of design for a major corporation captures this:

> *…a designer that pitched only one idea would probably be fired. I'd say 5 is an entry point for an early formal review (distilled from 100's). Oh, and if you are pushing one particular deign you will be found out, and also fired. By my standard it is about open mindedness, humility, discovery, and learning. If you aren't authentically dedicated to that approach you are just doing it wrong!* [1]

In this study, we investigate the impact of simultaneously evaluating three designs compared to just one during early usability testing.

[1] Alistair Hamilton, VP Design, Symbol Corp. Personal Communication.

The catalyst for this was a passage in [16]. Wiklund and his colleagues were investigating the impact that prototype fidelity had on user perceptions of usability. What is of interest for our purposes is the following excerpt from the discussion of their results:

> *In studies such as this one, we have found subjects reluctant to be critical of designs when they are asked to assign a rating to the design. In our usability tests, we see the same phenomenon even when we encourage subjects to be critical. We speculate that the test subjects feel that giving a low rating to a product gives the impression that they are "negative" people, that the ratings reflect negatively on their ability to use computer-based technology, that some of the blame for a product's poor performance falls on them, or that they don't want to hurt the feelings of the person conducting the test.*

Dicks has also observed a similar phenomenon; "Participants in a usability test are often in a strange environment. They may make many assumptions about what is going on that may not be accurate, including the possibility that they feel compelled to impress you and to under report errors." [3]

One possible implication of this is that if people are shown multiple prototypes, they could feel less pressured to impress the experimenters by praising a particular design. Being presented with multiple alternative designs may allow for a more accurate comparative evaluation.

We wanted to test this. We thought that having multiple designs to comment on would provide a better opportunity to balance negative with positive comments. Furthermore, our impression was that seeing multiple designs would let participants in a usability test know that there was not yet a commitment to a particular design, so they could be less hesitant to be critical. Finally, we believed that having experience with multiple design alternatives would provide a stronger foundation to inform suggestions: seeing the alternatives, they would know that some aspects could be better. The testing of these ideas is part of a larger agenda of gaining a better understanding of pursuing multiple designs throughout the entire design and development process.

RESEARCH QUESTIONS

Our approach was to design and conduct a simple experiment in which participants performed a usability test on a single interface design, and then evaluated it, or performed the same test on three alternative designs, one after the other, and then evaluated all three. We hypothesized that seeing one *vs.* three alternative designs would impact user feedback in the following ways:

H1: Participants will rate designs lower when all alternatives are seen, compared to when they see only one.

H2: Participants exposed to alternative designs will be less pressured to be positive, expressing fewer positive comments than those who only see one.

H3: Participants who see alternative designs will provide more suggestions for improvement compared to those who only see one.

METHOD

The study was a between-subjects design: different groups of participants saw one of three prototypes in isolation or were presented with all three prototypes. Here, our goal was to assess the impact of these different conditions on user ratings, number of positive and negative comments, as well as the number of substantial and superficial suggestions for improvement made in response to the prototypes they saw.

The system we chose to design was a House Climate Control System (HCCS) that regulates the temperature of a house by controlling the underlying heating and cooling mechanisms. Such systems typically work based on pre-programmed settings allowing for a number of daily temperature adjustments based on user-selected time intervals. The user is in charge to create, modify and activate settings, and can also override the current program temporarily. Similar systems have been used in other design practice studies and were found to be both practical and generalizable [6]. The HCCS system in this study was designed for a touch-sensitive screen.

All three interfaces were designed by the same team. In addition, the experimenters tried to ensure that the three alternative designs were consistent in terms of fidelity, functionality and quality.

Two rounds of pilot testing involving twelve pilot participants were conducted to test the prototypes, experimental set-up and procedure. The main benefit of these pilot tests was to point out the need for an extra person to play "the computer" in order to better manage and structure the tests. In addition, the pilot testing helped to refine the post-experiment questionnaire and interview questions.

Participants

We ran a total of 48 participants. They consisted of 26 female and 22 male students or recent graduates from a wide range of disciplines from the University of Toronto including Architecture, Computer Science, Engineering, Economics, Fine Arts, Life Sciences, Linguistics, Political Science, Psychology, and Women's Studies. The participants were intentionally taken from a wide range of backgrounds because we speculated that there might be differences in the way an Art student, for example, might criticize a design than the way a Computer Science student would. Participants were randomly assigned to conditions to avoid any biases. All participants received $10 Canadian as compensation.

Paper Prototypes

Three paper prototypes of a HCCS were developed. No attempt was made to make any one of them better or worse than the others. Each design had a distinctive stylistic approach: one was based on round dials and analog clocks (the Circular variation – Figure 1), another based on drop-down lists (the Tabular variation – Figure 2), and a third one based on horizontal timelines and sliders (the Linear variation – Figure 3). All interface components were hand-drawn using coloured markers on 5"x8" white index-card sheets. The cards were then laminated for durability as well as ease of handling and manipulation.

All three paper interfaces possessed identical functionalities incorporating the essential features of a House Climate Control System: controlling the current house temperature, creating and managing pre-programmed settings (for Spring, Summer, Fall, Winter, as well as Weekdays, Weekends, On Vacation and Special occasions), overriding the active program, and resetting the time and date of the house climate control system.

The prototypes were quick and inexpensive to make, with the bulk of the time being spent on their design.

Simulating Interaction

A number of simple, yet effective, methods were used to 'operate' the paper interfaces. Water-based markers were used to write on the laminated index-cards. These recordings could be easily erased using a wet napkin. Scotch® Removable Tape was used to quickly stick and remove movable parts. Pre-populated paper-based menus, drop-down-lists and pop-up-screens were created that manually "popped" in and out of the main screen. All of these operations were carried out by a person (other than the main experimenter) who simulated the behaviour of the system. We will refer to this person as the "*Computer*" [11, 14]. Having a dedicated person to play the role of the computer helped greatly in managing the tests, and reduced any time-delays typical of testing paper prototypes. The *Computer* became quite efficient in operating the paper interfaces after a short period of training.

Data Capture

Observations were recorded by note-taking throughout the experiments. Furthermore, two digital video cameras captured audio and video data, one focusing on the paper prototype and another capturing a wider angle view of the participant and the experimenter.

Experimental Conditions

Participants were randomly assigned either to one of three *single design* conditions or to the *multiple design* condition. In the single design conditions, participants were presented with only one of the three paper prototypes (either the "Circular", "Tabular" or "Linear" conditions), with no mention of other existing designs. In the other condition (the "Multiple" condition), each participant was presented with all three of these different prototypes in counterbalanced order. Forty-eight participants took part in all: twelve in the Multiple condition and twelve in each of the single design conditions.

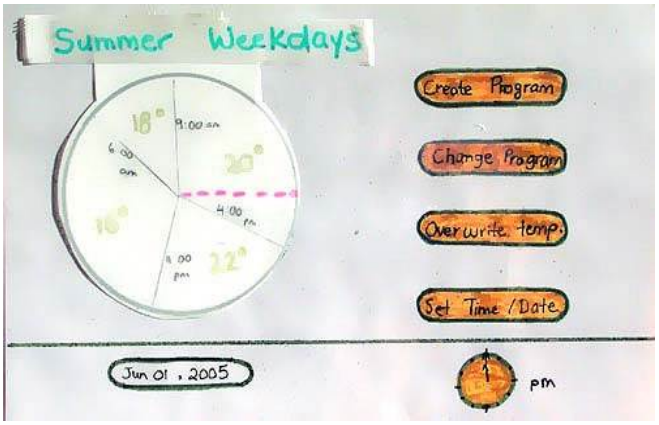

Figure 1. The "Circular" paper prototype

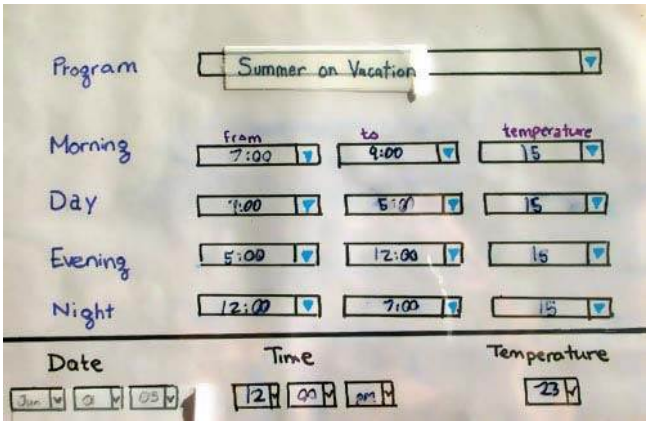

Figure 2. The "Tabular" paper prototype

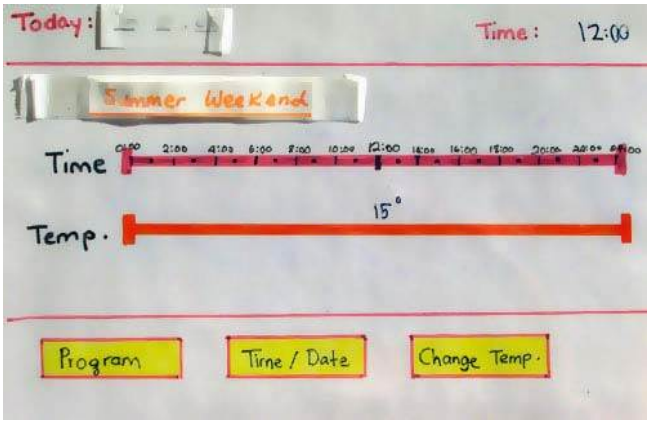

Figure 3. The "Linear" paper prototype

Procedure

Each experimental session started with a short introduction phase in which the experimenter briefed the participant about the study, the main functionalities of a HCCS were reviewed, and images of 4 common HCCS's in the market were shown to ensure the participant was comfortable with the conceptual model of the system. Participants were

asked to use their fingers to interact with the 'touch-sensitive' interface(s), and think-aloud while performing the tasks.

Four tasks were designed to test the main functionalities of the HCCS: creating two pre-programmed settings (one for weekends, and one for vacation), resetting the time and date of the thermostat, and overriding the current temperature. In cases where participants saw multiple designs, they performed the same set of tasks on all prototypes, one prototype at a time. Upon completion of the tasks, (that is, after seeing all three alternatives for those in the Multiple design condition), each participant filled out a questionnaire to rate the design and usability of the prototype(s) viewed, and participated in a short semi-structured interview in which they expressed their likes and dislikes of the system as well as suggestions for design improvement. At this point, participants could view and refer back to each of the paper prototypes they saw.

Post-Experiment Questionnaire

The questionnaire extracted opinions about the usability and design of the paper prototype(s) based on ratings on ten-point Likert scales. The first section focused on ease-of-use and consisted of four questions to rate the subjective difficulty of each of the four tasks. For example,

"In task 1 you were asked to create the 'Summer Weekend' program. How easy or difficult was it to perform this task using the prototype?"

Extremely difficult									Extremely Easy
1	2	3	4	5	6	7	8	9	10

The next section solicited feedback about the prototypes based on commonly used design dimensions that had been refined through the pilot testing. Seven dimensions were examined, again using a ten-point rating scale to indicate how well or poorly each dimension was implemented in the design. These had to do with: aesthetics, readability, consistency, match to real-world metaphors, navigation, explorability, and learnability. Finally an overall rating of the usability of the prototype was provided.

Two separate formats of the same questionnaire were prepared corresponding to the experimental condition: a single version in which only the individual design was rated, and a multiple version to rate all three variations viewed. In this latter version, each question contained three ten-point Likert scales, one for each design viewed.

No time limit was enforced, but most participants completed the questionnaire in five to ten minutes.

Post-Experiment Interview

Upon completion of the questionnaire a short semi-structured interview was conducted by the experimenter. Participants were first asked to list all the features that they liked or disliked about the design of the interface(s) - any positive or negative aspects that came to their minds. The next question was as follows: "If we were to use this design as our final design would you change anything, and if so what?" This allowed the participants to provide suggestions for change in order to improve the design of the interface(s) reviewed. Lastly they were asked for any other suggestions or comments on the general usability and usefulness of the HCCS.

RESULTS

In order to assess the impact of showing multiple design variations as opposed to a single one on user feedback, we observed the effects on user ratings for the ease-of-use, design and usability of the prototypes (H1), users' willingness to criticize the design (H2), as well as the number and type of suggestions for design change (H3).

Impact of Showing Alternative Designs on Participants' Design Ratings (H1)

Within each questionnaire, each prototype was rated on twelve measures: four ease-of-use ratings for each of the four tasks, seven design dimensions and one overall rating of the usability of the prototype. In order to arrive at a final "score" for each prototype, we averaged each participant's ratings for the first eleven questions, thus excluding the overall rating given by the participant.

The rationale for this was that when we examined the cross correlation between ratings given for individual questions 1 through 11, we found that most questions were significantly positively correlated with the overall rating given by the user in the final question ($p < .05$). Furthermore an *average* of these first eleven questions was also positively and significantly correlated with users' overall ratings (at a $p < 0.01$ significance level). On this basis we reasoned that the average represented both a robust and reliable measure of overall evaluation which was redundant with the rating given in the final question. The remainder of the analysis is therefore based on this averaged final score for each paper prototype.

Ratings for Single versus Multiple Interfaces

For each different prototype, we compared the score assigned to the design when it was seen individually with the score assigned when the same interface was seen in a group of three alternatives. Mann-Whitney U tests were conducted to test for differences across these conditions (see Table 1). Differences of $p < .05$ are considered to be significant.

Table 1 shows that the average score for each design was higher when seen individually, compared to when seen in a group of three. While only the results for the *Circular* and *Linear* designs were significant, those for the Tabular variation did approach the $p < .05$ level. Overall, the results show support for hypothesis H1. As we shall see, the post-experiment interviews provide some insights into why the results for the Tabular design may not have been as strong as for the other two.

Analysis of Comments and Suggestions

In order to test hypotheses H2 and H3, we analyzed participants' comments, criticisms, and suggestions. To do this, we reviewed video recordings of experimental sessions including verbal protocols and actions while performing tasks, as well as comments and suggestions made in the post-experiment interviews. Here, we chose a selection of half of the 48 tapes for in-depth analysis: six tapes were chosen at random from the Multiple condition, and six from each of the single design conditions for a total of 24 tapes.

Paper Prototype	Seen individually N = 12	Seen in group of 3 N = 12	p value *=sig.
Circular	9.08 *(SD = 0.53)*	8.13 *(SD = 1.10)*	0.004*
Tabular	8.83 *(SD = 0.56)*	8.39 *(SD = 0.86)*	0.064
Linear	7.92 *(SD = 0.85)*	6.89 *(SD = 0.91)*	0.014*

Table 1. Mean and (standard deviation) for each interface, and results of the Mann-Whitney U test (one-tailed).

Through preliminary analysis of the video data, we generated a way of classifying participants' statements. We identified two broad classes of statements of interest: participants either made "comments" (facts or personal opinions) about the design, or provided "suggestions" for change to improve the current design. In turn, we found the comments could be classified as either "positive" (e.g. "I like the way it guides you through each step of the process"), or as "negative" (e.g. "this is too cluttered and hard to read"). Suggestions, on the other hand, were found to be either "substantial" (e.g. "it would be nice to allow for more than 4 intervals per day") or "superficial" (e.g. "the colours are dull; I would have more colours."). In terms of the substantial suggestions, we further classified them as ideas for improvement which were original or "new" (e.g. "it would be nice to have both Celsius and Fahrenheit") or as "borrowed" from ideas they had seen in other interfaces (e.g. "show the current temperature on the main screen, like in the Tabular interface").

The tree structure shown in Figure 4 graphically illustrates this categorization of participants' statements. The "leaves" of this tree constitute the categories according to which user feedback was analyzed, as shown in Table 2.

It is important to note that the category of "borrowed" ideas only applies to prototypes seen in the Multiple design condition, since in the single design conditions, participants could not have borrowed the idea from another interface, as none was shown.

An external assessor reviewed the 24 selected video tapes and extracted a total of 337 statements that were either comments or suggestions for change. Two independent judges were given the list of extracted statements to be categorized based on the above taxonomy. They were provided with definitions and examples of each category, and viewed the transcribed versions of all 24 videos in order to better understand the context of each statement.

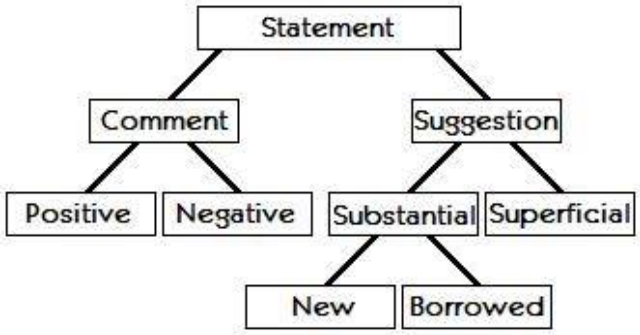

Figure 4. Categorization of User Feedback

The inter-rater reliability for categorization of user statements was measured using Cohen's Kappa test with a measure of agreement of 85%, which was highly significant ($p < .001$). Table 2 summarizes the findings from the video analysis based on the mean of the judges' scores.

In Table 2, the column labeled *Circular seen alone* shows the data for participants' comments and suggestions regarding the Circular variation when the prototype was seen in a single design condition. The column labeled, *Circular seen in a group of 3* refers to the number of comments and suggestions that were made with respect to the Circular design when the interface was shown in the Multiple design condition. The latter therefore, is a subset of all comments and suggestions generated in the Multiple design condition that were directed towards the Circular variation only. The same holds true for the Tabular and Linear prototype data shown in Table 2.

A t-test was carried out on each category to test for differences between the number of statements of each type generated when the prototype was seen alone, and when it was seen in a group of three. These results are also shown in Table 2.

Impact of Showing Alternative Designs on User Criticism (H2)

In order to assess the effects of showing multiple prototypes on participants' willingness to be critical of each design, we looked at the number of positive and negative comments generated in each condition. As shown in Table 2, the number of positive comments generated towards a design was significantly lower when the prototype was seen in a group of three, as opposed to individually ($p < 0.05$). This effect was consistent across all three prototypes.

Statement type	**Circular seen alone** N = 6	**Circular seen in a group of 3** N = 6	p value	**Tabular seen alone** N = 6	**Tabular seen in a group of 3** N = 6	p value	**Linear seen alone** N = 6	**Linear seen in a group of 3** N = 6	p value
Positive Comments	6.42 *(2.87)*	2.67 *(2.80)*	0.045*	4.75 *(1.17)*	2.17 *(1.47)*	0.008*	5.83 *(2.66)*	2.33 *(1.66)*	0.025*
Negative Comments	1.83 *(1.03)*	2.50 *(1.38)*	0.365	1.50 *(1.61)*	4.08 *(4.10)*	0.182	1.92 *(2.01)*	5.08 *(2.46)*	0.036*
Superficial Suggestions	0.92 *(0.92)*	0.67 *(0.88)*	0.640	0.50 *(0.45)*	0.17 *(0.41)*	0.208	1.75 *(0.82)*	0.33 *(0.61)*	0.008*
New Substantial Suggestions	2.17 *(1.81)*	0.75 *(1.17)*	0.138	1.08 *(1.11)*	1.58 *(1.39)*	0.508	2.00 *(1.52)*	0.92 *(1.32)*	0.217
Borrowed Substantial Suggestions	N/A	0.92 *(0.74)*	N/A	N/A	0.50 *(0.63)*	N/A	N/A	0.50 *(0.63)*	N/A
Total Substantial Suggestions	2.17 *(1.81)*	1.67 *(1.60)*	0.623	1.08 *(1.11)*	2.08 *(1.50)*	0.219	2.00 *(1.52)*	1.42 *(1.66)*	0.539

Table 2. Mean and *(standard deviation)* of number of user comments and suggestions. p-values (*= sig at p < .05) are the result of two-tailed t-tests.

Furthermore, for the Linear prototype (the one rated lowest among the three) the number of negative comments generated was significantly higher ($p < 0.05$) when it was seen in a group of 3 compared to when it was seen on its own. All of this confirms a more positive attitude when single designs are seen, as is typical of usability testing. This effect appears to be minimized when designs are seen against other alternatives, allowing participants to be more critical of designs, supporting hypothesis H2.

Further evidence of this difference for negative and positive comments is shown in Figure 5 to Figure 7. These figures illustrate the frequency of statements extracted from the 24 selected videos. Three values are shown for each of the statement categories: the left-most bar shows the total number of statements of that type generated when the prototype was seen alone; the middle bar shows the *overall* number of statements of that type generated in the Multiple design condition (with respect to all three designs); the right-most bar, on the other hand, shows the subset of statements generated in the Multiple design condition that were related to the prototype in question.

With respect to the Linear prototype, for example, we can see that the number of positive comments for the single design condition is less than the total number of positive comments for all three designs in the Multiple condition, however, this is not significant ($p > 0.5$). This is the case even though participants have seen three different designs and potentially could have made three times the number in the single design condition. The negative comments tell a different story, however. Here the number of negative comments for all the designs in the Multiple condition is over *six times* more than that for the single Linear design (70 compared to 11.5 negative comments). This difference, unsurprisingly, is significant ($p < 0.005$). All of this is indicative of an interaction effect in which we found that, when participants saw the Linear design in isolation, they made a preponderance of positive comments compared to relatively few negative ones. When they saw that same design next to two alternatives, the balance of positive and negative comments changed, this time with more negative comments than positive. This is confirmed by a significant interaction for positive and negative comments directed at the Linear condition, across single and multiple design conditions ($p < .001$), shown in Figure 8.

Impact of Showing Alternative Designs on the Number of Ideas and Suggestions for Design (H3)

Finally, we predicted that presenting participants with multiple designs would increase the number of ideas and suggestions for design improvements. This was based on the belief that being exposed to a variety of alternative solutions would stimulate new ideas and suggestions. Contrary to our prediction, we found that there was no significant difference in the number of *substantial* suggestions generated by the participants in the Multiple condition compared to the single design condition (Table 2). We did however see that the number of *superficial* suggestions generated towards the Linear prototype decreased significantly from the Single to Multiple design condition. We speculate in the Discussion on why these effects could have happened.

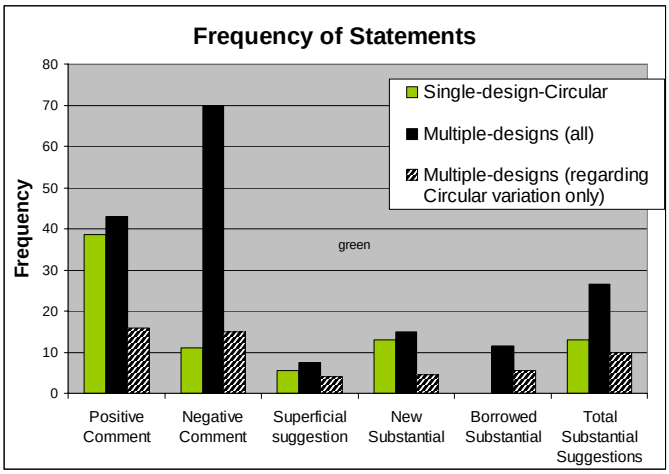

Figure 5. Frequency of statements for Circular prototype

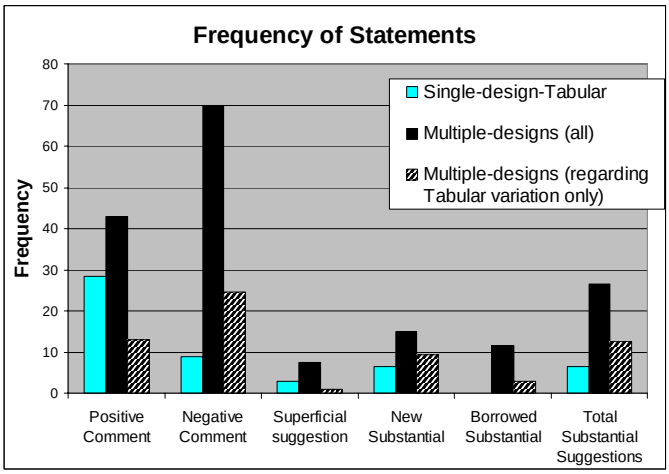

Figure 6. Frequency of statements for Tabular prototype

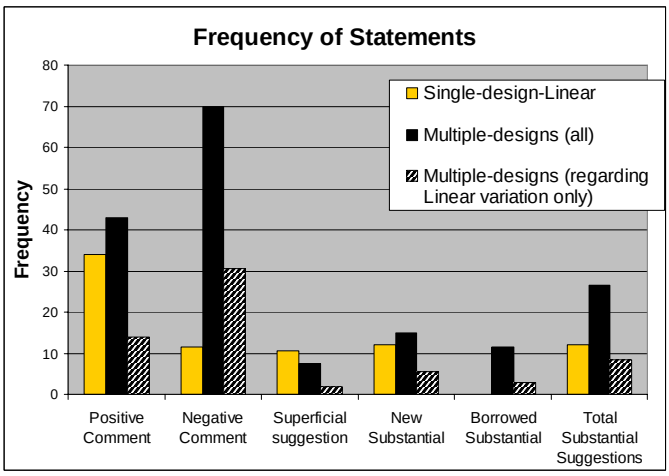

Figure 7. Frequency of statements for Linear prototype

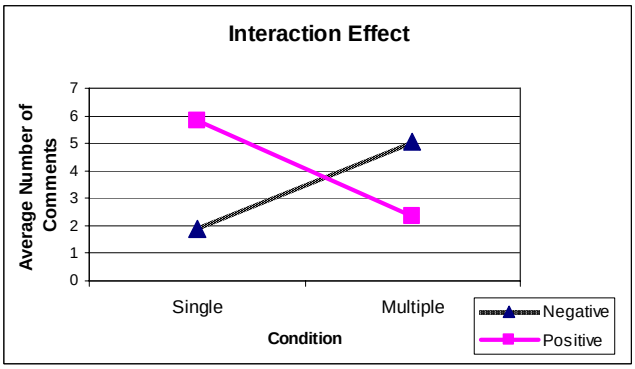

Figure 8. Interaction between positive and negative comments made in response to the Linear design in both conditions.

DISCUSSION

Faint Praise (H1)

While it is usually nice to be told how good our work is, one also wants to be confident that praise is deserved. Our study shows that overall ratings and positive comments are affected by the number of variations of the design that a participant is exposed to. When only one prototype is seen it provokes higher ratings and more positive comments. While without further study one cannot be sure how valid the positive (or negative) comments are, our working assumption is that the ratings and reactions in the multiple condition are more considered, and therefore more relevant to the designer, since they are based on a broader – and therefore more informed – base of experience.

Hence, we do not view the fact that participants' evaluations were higher in the Single conditions as desriable. As we have pointed out, the fact that usability tests were vulnerable to overly positive results has been suggested in the literature in the past. Our study confirms that it does happen, provides some indication of the degree to which it happens, and suggests an approach to avoiding the problem in the future.

We believe that Wiklund and his colleagues' [16] analysis as to why participants this occurs was reasonable. Showing participants more than one prototype provides a clear message that we have not yet made up our mind as to what path to follow. Since we have not made a commitment, and present ourselves as neutral with respect to the alternatives presented, they cannot disappoint us (unless they have an equally negative reaction to all of the alternatives shown – something that did not occur in our test). One way of saying this is that presenting multiple alternatives provides them a means to criticize without being negative. We also believed that this would be all the more true since the approach provided more opportunities to balance negative comments with positive ones, which we will discuss more fully below.

One aspect of the data deserves a bit more discussion: why we did not get significance with the Tabular design. It was close, but not there.

When we looked in more depth at the questionnaire data in the Multiple condition, they indicated that the Tabular variation was the easiest one to use, and rated highest in terms of learnability. The post-experiment interview data also supported this finding; "so familiar and simple", "very easy for a first-time user", and "I am used to menus like this from operating systems" are examples of user feedback in response to the Tabular version which was based on commonly used drop-down-menus. All interactions in the Tabular version were carried out by touching the down arrows of the drop-down-menus and selecting appropriate values from the lists. This was a very familiar interaction for the majority of our participants. While most familiar, this design, however, was arguably the least appropriate of all three.

The use of drop-down lists was referred to as inappropriate and "machine like" by some participants. "Very easy to use and very straight forward; that's the only thing I like; this one is not friendly at all! Very machine like!" explained one of our participants about this design in the Multiple condition. Another participant in this condition commented "well even though it's consistent to have the time be [presented] in little pull down menus as well, to have it more looking like a clock would be better because then you… just whip it around to the right time".

It seems that people have different priorities for measuring how 'friendly' or usable a system is; while some participants appreciated the simplicity, and ease of learning of the Tabular version, others voted against it in favor of the more "visual", and "iconic " nature of the Circular and the Linear variation, even if they required some initial training.

In short, users had more familiarity with the interaction style of this design, regardless of its intrinsic appropriateness, and this differentiated it from the other two. This may be why there was a weaker differentiation between the two conditions in this case. What this might mean in terms of which of the three one might pursue for productization, if any, is not a question for this current study.

Increased Criticism (H2)

One of the more interesting aspects of our results is in how the increased freedom to be critical, which we expected, was accompanied by a decrease in positive comments when seeing multiple alternatives. In the case of the Linear prototype, which was rated lowest among the three, and therefore was arguably the least successful design, we also saw a significant increase in the negative comments in the Multiple condition. What appears to be the case here is that being exposed to three alternatives seems to have removed inhibitions about being negative that appear to be present when the Linear design was seen on its own. One explanation for this may be that the experimenter's lack of commitment to any one design – as conveyed by showing three alternatives – removed any concern users might have had in causing disappointment. Hence they felt more open to point out the areas where this design fell short of the other alternatives.

The one thing to keep in mind in this is that negative comments are good (although, see our discussion concerning [5], below). Hence, the increase in negative comments that occurred in the Multiple condition for the poorly designed interface potentially provides more and hopefully better data to the usability engineer in terms of zeroing in on problems with the design.

Are We All Designers? (H3)

The other thing that surprised us with our results was the fact that we did not get more suggestions for design improvement in the Multiple condition. In this, we appear to have fallen into the trap of confusing usability engineering with participatory design.

Our assumption was that showing users multiple design solutions would release some creative juices, and that the juxtaposition of alternative designs might both suggest to the user that there might be other possibilities, as well as provide a wider base of experience that might serve as a catalyst to suggestions from them.

This was not the case. On the one hand, what our data suggests, and is in retrospect obvious, is that the procedures for participatory design are very different than usability testing – even if the same types of instruments (such as paper prototypes) may be employed by both.

We did however find that superficial suggestions decreased from the single to Multiple condition for the Linear prototype. Superficial suggestions, as understood by the judges, had to do with minor aspects of the visual appearance of the prototypes, without indication of how they might improve the design. For example the statement "I would change the colours" was considered superficial, where as "I would use colours with higher contrast, to improve readability" was considered substantial. The significant decrease in superficial suggestions with respect to the Linear prototype in the Multiple design condition could imply that seeing other alternative designs helped the participants focus on more critical aspects of the design, rather than minor details. Unfortunately this didn't lead to them producing more substantial suggestions.

The second thought that we had on considering these results was, again, that we shouldn't have been surprised. Yes, exposure to three different designs did provide a broader base of experience to the participants in that condition. But there is no reason to believe that that experience would suddenly inject some creativity into the participants.

Their negative comments were based on actual experience, and in the Multiple condition, their confidence due to their increased experience, was probably higher. Hence, they had yet another reason to be comfortable making negative comments.

But as for making suggestions? This would involve speculation, and stepping out on a limb for which they had no training, experience, or language. Furthermore, in making such suggestions, they – novices in design – would be stepping into the domain of experts – those who designed the interface in the first place. It seems reasonable that users would be hesitant to expose themselves as potentially naïve in terms of their suggestions, just as Wiklund *et al* speculated on their potential concern with appearing not to be competent using computer-based technology.

In short, design and creativity are specialized skills. There is no reason to expect them to be spontaneously manifest in those not trained in the field. Unless we can find a methodology that changes this, perhaps the focus in usability testing should remain in detecting errors, not soliciting ideas. If we want to engage users in that activity, then participatory design or some other appropriate techniques are required.

These findings are especially important in light of [5], who found that "…redesign proposals were assessed by developers as being of higher utility than problem descriptions." If this is true, our data suggest that usability testing in either of the conditions that we tested is not going to help generate such proposals. And, we believe that our study makes some contribution towards rectifying the situation pointed out by the authors that, "…no studies have investigated redesign proposals as a distinct and systematic outcome of usability evaluation."

How Our Approach Helps in Getting the Right Design

One reason that conventional usability testing mainly helps designers get the design right (rather than the right design) is this. Once a design is prototyped and tested, it hardly ever gets rejected by the users. Rather, it typically leads to an iterative improvement of the same design, rather than a return to the drawing board (which might lead to an alternative *right* design). As stated by Bowers and Pycock [1] "explicit issuing of requests for redesign of DNP [Designer's Note Pad, a tool that was prototyped and evaluated in their study] is very rare in our materials. Equally rare are explicit disagreements on the part of the user to suggestions made by designers".

This was the case in our study as well. Not a single participant from the 36 who saw an individual prototype issued a "request for redesign", showed an "explicit disagreement", or explicitly rejected the design viewed. On the other hand, three out of the twelve participants in the Multiple design condition explicitly rejected one of the prototypes viewed. Concerning one prototype, one participant said "I won't buy this product because even though it's very easy to use it's hard to read". Another participant reasoned, "When I look at this... it's too complex, I think aesthetically it's the worst; I wouldn't want to choose it". A third participant not only stated which prototype she would eliminate, she also named the most successful one in her opinion: "If I were to choose which one to use it would probably be the drop-down. This one, because it's more quick I think. The first one I wouldn't choose at all". Two other participants in the Multiple design condition explicitly named the most successful prototype of the three. All of this information was volunteered by the participants in the Multiple design condition, but nothing like it in the Single condition.

We believe that when testing multiple alternatives, statements such as above, combined with comparative user ratings, and comments for each of the prototypes viewed, can help the designers in selecting the right design, before proceeding with getting the design right.

CONCLUSION

Perhaps the most important contribution of this research is its implications regarding usability engineering. One of the standard texts [8] teaches that multiple alternatives are to be considered only at the very beginning of the process. From then on, one is taught to work through successive iterations of the one design chosen from the many. What our findings suggest is that low-cost techniques, such as paper-prototyping enable multiple alternatives to be explored beyond the initial ideation phase. More to the point, they suggest that doing so can enable us to obtain a less inflated subjective appraisal of our designs, as well as obtain more critical comments that help identify problems.

By exposing users to multiple designs, we give them the opportunity to express which is their favourite, as well as reject designs that they would not consider, in light of the alternatives. A quote from one of our participants illustrates this point, "If I were to choose which one to use it would probably be this one; because it's more quick I think; the first one I wouldn't choose at all". This did not happen in a single design condition. Yet this type of feedback is a crucial part of "getting the right design."

Another important outcome is recognition of the need to reconsider the narrow definition of "parallel design", again as taught by [8]. While the notion of having different teams or designers working simultaneously but independently on the same problem is sometimes a useful technique, it is too narrow a perspective to capture the value of pursuing alternative designs. The designs in our study were not constructed independently by different designers. In our experience, neither are the vast majority of alternative designs that are produced daily by conventional industrial designers, architects and graphics designers, for example. Generating meaningfully distinct design alternatives is standard practice for them, and the ability to do so as an individual or as a team is one of the fundamental skills of the design profession. It is the norm, not the exception. This is an aspect of design practice from which the HCI community may well be able to benefit. We believe that our results suggest that is a direction that warrants further exploration.

This study has some implications for participatory design, as well. Erickson [4] among others has written about the potential value of paper prototypes to help engage end users in the design process. However, our data caution against assuming that any such benefit will result for their use in the context of usability testing, even when multiple designs are used. Consequently, if one agrees with the findings of [5], then our study emphasizes the need to adopt approaches beyond standard usability testing techniques in order to generate redesign proposals.

Pursuing this constitutes the next phase of our research. At the end of each session, we had participants sketch their ideal interface. What we learn from those drawings will form the basis for our next study.

Next, the research reported in this paper looked at comparing prototypes that reflected fundamentally different design languages, or styles. What about the details of any one of designs however? What if, based on a study like this, one made the choice to pursue the Circular style for the final product? How might our results impact what comes next, as we pursue that particular design?

Finally, we believe that the issues that we are starting to explore may become ever more relevant as we move away from software-only designs running on PCs, to products that are embedded in hybrid software/hardware appliances. We hope that this study will lead to new approaches that might help us face these challenges. Doing so would bring our own practice more into line with that of industrial designers, for example, with whom we are inevitably going to be working with more closely in the future.

ACKNOWLEDGMENTS

We would like to thank Aha Blume, Ian Chan, Irina Ivanova, Keri May, and Andrew Warr for their contributions to this research. This research has been partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Microsoft Research. This support is gratefully acknowledged.

REFERENCES

1. Bowers, J., and Pycock, J. Talking Through Design: Requirements and Resistance in Cooperative Prototyping, Proc. CHI '94, *ACM Press*, (1994), 299-305.
2. Catani, M.B. and Biers, D.W. Usability evaluation and prototype fidelity: Users and usability professionals, *Proc. Human Factors and Ergonomics Society 42nd Annual Meeting*, (1998), 1331-1336.
3. Dicks, R.S. Mis-usability: on the uses and misuses of usability testing. Proc. SIGDOC 2002, *ACM Press*, (2002), 26-30.
4. Erickson, T. Notes on Design Practice: Stories and Prototypes as Catalysts for Communication. In J. Carroll (Ed.). *Scenario-Based Design: Envisioning Work and Technology in System Development.* New York: Wiley & Sons, (1995), 37-58.
5. Hornbæk, K., and Frøkjær, E. Comparing usability problems and redesign proposals as input to practical systems development, *Proc. of CHI 2005*, ACM Press (2005), 391-400.
6. Jensen, S. The Simplicity Shift: Innovative Design Tactics in a Corporate World, *Cambridge University Press*, (2002), 133-148.
7. Nielsen, J. and Desurvire, H. Comparative Design Review: An Exercise in Parallel Design (Panel). *Proc. INTERCHI 1993*, 414-417.
8. Nielsen, J. Usability Engineering. *Academic Press* (1993).
9. Nielsen, J. Diversified Parallel Design: Contrasting Design Approaches (Panel). *Proc. CHI 199*4, ACM Press (1994), 179-180.
10. Ovaska, S. and Raiha, K.J. Parallel design in the classroom, *Proc. CHI 1995*, ACM Press (1995), 264-265.
11. Rettig, M., Prototyping for tiny fingers. *Communications of the ACM (CACM)*, ACM Press (1994), 37(4), 21-27.
12. Rudd, J. Stern, K. and Isensee, S. Low Vs. High-Fidelity Prototyping Debate, *Interactions*, (1996), 76-85.
13. Sefelin, R., Tscheligi, M., and Giller, V. Paper prototyping – what is it good for? A comparison of paper-and computer-based prototyping, *Proc. CHI 2003*, ACM Press (2003), 778-779.
14. Snyder C. *Paper Prototyping - The Fast and Easy Way to Design and Refine User Interfaces*, Morgan Kaufmann (2003).
15. Walker , M., Takayama, L. and Landay, J.A. High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes, *Proc. Human Factors and Ergonomics Society 46th Annual Meeting*, (2002), 661-665.
16. Wiklund, M., Thurrott, C., and Dumas, J. (1992). Does the Fidelity of Software Prototypes Affect the Perception of Usability? *Proc. Human Factors Society 36th Annual Meeting*, (1992), 399-403.