CPSC 444

Team χ^2 - Milestone IV

Collocated, Collaborative Diagramming





Alex Totolici 69229045

André Malan 58353061

Gabe Silk 93911014

Jré Sarenac 17191081

Piam Kiarostami 94358991

March 31, 2010

[A1] Pilot Test

1. Overview

We conducted the pilot experiment with three subjects, in the same room that we conducted the real experiment in. We chose classmates as participants in order to get more constructive feedback on the experimental design. We had the subjects execute a task on each of the three interfaces from the experimental design (whiteboard, computer, prototype), just as in the regular experiment to come.

We initially estimated that 10 minutes should be the allotted time for the tasks. During the pilot, the participants finished after approximately 7 minutes on the multi-touch prototype interface. After this, we continued monitoring their interaction to identify more weaknesses in the prototype. We noticed during this time, that aside from suggestions, participants also posed questions and made corrections to each others' works. We then conducted a discussion where we talked about the different tasks with the participants in order to get suggestions for improvements.

2. Adjustments

a) We decided that we needed to record three types of collaboration instead of one, because we felt that there could be some significance in the *type* of speech acts that were made during the trials (see above section), and that this could help us measure and compare collaboration along multiple axes
b) In lieu of the early completion time, we restricted the allotted time per task to 5 minutes
c) We decided not to measure the time spent talking as it was almost impossible to do with the number of experimenters and the equipment that we had available
d) We created a 2 minute warm-up task for participants to complete before each condition, utilizing the Power Law of Practice to reduce problems stemming from familiarity with UML and with the conditions
e) We realized that it was necessary to have a pen for each participant so that they could fill out their consent forms and questionnaires in parallel
f) We realized that we needed a more detailed coding sheet including a table for tallying and place for a drawing of how the group was arranged
g) Several bugs were uncovered in our prototype and improvements were suggested by the participants. Modifications included: smaller keyboards, object rotation, and a color coding of keyboards to match the section being edited as an easy visual cue

[A2] Experiment Report

[Note: **bolded items** indicate a divergence from the Milestone III report]

Abstract

An assessment of collaboration, user satisfaction, and performance using collocated, collaborative multi-touch software on a table interface, consisting of an experiment comparing three tools: computer software, whiteboard, and a novel multi-touch prototype. Collaboration was measured using speech-act counts, performance was measured by comparing work products, and user satisfaction was compared via questionnaire using multiple Likert scales. The multi-touch prototype was preferred to the computer interface, and the whiteboard was not significantly preferred to the prototype. Further research is recommended to discover novel interaction methods for multi-touch, collaborative diagramming software.

1. Introduction

Collocated, collaborative diagramming is an essential task for many software engineers working on medium to large-scale projects. This experiment focused specifically on UML diagramming tasks for upper-level, computer science students. Current diagramming methods vary from software tools such as Microsoft's Visio, to primitive tools such as the whiteboard or the classic pen/paper approach.

Each diagramming method has its relative advantages and disadvantages. Software tools are generally harder to collaborate with effectively as only one member of the group can have control of the system at any given time. Primitive tools are generally better with respect to collaboration but require the additional task of having users transfer the diagram form the whiteboard (or pen/paper, etc.) to a more professional computer version.

The goal of our prototype was to merge the advantages of the primitive and software tools in order to allow users to efficiently and enjoyably create collocated, collaborative diagrams. This paper describes the experiment methods in detail, provides recommendations for future research and discusses the impact and relation to other works.

2. Description of Experiment

2.1 Introduction

Our goal for the experiment was to assess whether our novel multi-touch system:

a) Allowed users to collaborate effectively
b) Gave users a performance benefit over traditional tools
c) Provided a degree of user satisfaction at least as high as with traditional tools

The key challenge in evaluating our system was to establish solid metrics for each of these evaluation goals. In particular, measuring collaboration proved to be difficult, given the somewhat complex and subjective nature of the concept. We opted to use a hybrid of subjective and objective methods to triangulate on, with certain methods and metrics drawn from the literature.

In order to establish that our design approach was viable, we hoped to see an increase in user collaboration over conventional tools, while also showing that in the worst case, our system was not significantly worse with regards to functionality and enjoyment.

2.2 Method

2.2.1 Participants

Our participant pool consisted of **fifteen** upper-level, university-level computer science students. The participants were divided randomly into five groups of three. **Participants were recruited simply by word of mouth and were not required to have filled out our questionnaire from MS II.** These were the ideal participants for this experiment.

2.2.2 Conditions

We are comparing three different levels or conditions of the independent variable, namely interface type. The three conditions were the whiteboard, a computer software (Creately), and our prototype.

2.2.3 Tasks

Each group was given a UML task to complete and a diagramming tool to use to complete the given task. Each task required simply connection high-level architecture. The tasks were designed to be too large for the participants to complete in the amount of time given **(5 minutes)**, thus (hopefully) forcing participants to work collaboratively. Task 1 was completed on the whiteboard, task 2 on the computer software and task 3 on the prototype. All tasks can be found in appendix A1.

2.2.4 Design

The experiment followed a 1x3 design, with 1 level of expertise and 3 interfaces that were tested. **Satisfaction and collaboration was measured between subjects, while performance was measured within subjects**.

Trials were conducted on groups of three people each. For each diagramming method, we looked at three metrics of evaluation (collaboration, performance and satisfaction), where **the collaboration metric was further broken down into a simple count of the number of suggestions, questions, and corrections made.** Performance was measured by using a predetermined marking scheme that was designed for the problems we assigned. The problems themselves were selected to be of equivalent difficulty by using old exam questions of the same worth. Satisfaction was measured by self reporting[1] [2], using a Likert-scale questionnaire.

[1] Kevin Baker , Saul Greenberg , Carl Gutwin, Empirical development of a heuristic evaluation methodology for shared workspace groupware, *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, November 16-20, 2002, New Orleans, Louisiana, USA

Our subjects were randomly assigned different questions on each diagramming method, so as to minimize *learning* effects, and a short practice question was given before each round to minimize *learning curve* effects.

2.2.5 Procedure

1. Each group of users was given an instruction sheet with a single UML diagramming problem on it (see appendix A1 for instruction sheet and UML problems)
2. **Each group was given two minutes to complete the warm-up task**
3. Each group was directed to one of the three stations – whiteboard, external software, prototype – and given a short warm-up practice question to be done at that station
4. Each group was given one of the UML diagramming tasks, and told to use the tool(s) at their station to attempt the assigned task
    a. **Task 1 was completed on the whiteboard**
    b. **Task 2 was completed on the computer**
    c. **Task 3 was completed on the prototype**
5. Users were given **5 minutes** to complete each task
6. After the allotted time was up, we took pictures of the users' end products for later scoring, and erase whatever they have created
7. Groups were then rotated onto the next station in round-robin fashion

After all three trials were completed, we conducted a brief unstructured interview to determine the users' overall impression of each of the diagramming tools, **preceded by** a questionnaire involving a self-assessment of their perceived level of collaboration and overall satisfaction.

2.2.6 Apparatus

The experiment was conducted in a laboratory in the ICICS building. The room had enough space for all three stations to be set up. The experiment had three stations, as well as **a central table containing a consent form and questionnaire for each participant and a working pen**. Each station had its own set of equipment along with the following common equipment:

- 3 sheets of paper, each one with 1 of the 3 tasks to be performed
- 1 sheet of paper with a practice task to be performed
- 1 sheet of paper with instructions for how to complete the tasks
- 1 sheet of paper with suggestions for how to use the medium
- Timing device
- Digital camera

The individual stations and their equipment were as follows:

- Whiteboard with three whiteboard markers, consisting of three different colors
- Computer with browser open and *Creately*[2] loaded (see appendix B2 for screenshots)
    - *Creately* set up with only the tools needed to complete the task
- *SMART Table*[3] with our prototype loaded (see part C for screenshots of our prototype)

[2] A web-based diagramming and design application service operated by Cinergix, Pty Ltd. (http://www.creately.com/)
[3] A multi-touch, multi-user touchscreen developed specifically for primary education (http://www2.smarttech.com/st/en-US/Products/SMART+Table/

2.2.7 Independent and Dependent Variables

Independent Variables

- Diagramming Method
    - three levels: whiteboard, software, prototype

Dependent Variables

1. Collaboration level
    a. Measured by amount of time a group spends talking, coupled with a numeric count of the number of suggestions made, followed by self reporting via questionnaire and interview
2. Performance
    a. measured using a predetermined marking scheme taken from the exams in which the problem(s) appeared
3. Satisfaction
    a. measured using a 5-point Likert scale (see Appendix B1.b)

2.2.8 Hypotheses

1. **Null:** User collaboration is equal regardless of the diagramming tool used
   Alternative: Collaboration level is not equal among all three methods of diagramming
2. **Null:** User performance is equal regardless of diagramming tool used
   Alternative: User performance level is not equal among all three methods of diagramming
3. **Null:** User satisfaction level is equal among all three methods of diagramming
   Alternative: User satisfaction level is not equal among all three methods of diagramming

2.3 Problems/Limitations

An important problem was discovered in the group setting of the post-experiment discussion, which was conducted with all of our participants in the same room. The group discussion was informative, but limited in that any self-reporting of satisfaction, enjoyment, or general opinions could be influenced by the group setting. The discussion tended to be steered by the most vocal and/or opinionated members, while several participants were reluctant to speak up, possibly due to shyness or a desire to not go against the group consensus. This scenario could be avoided in the future by conducting individual interviews, which we feel would give more representative results.

Another problem was the fluctuation of the SMART Table's touch sensitivity under varying illumination levels. This is an artifact of the SMART Table’s use of total-internal-reflection technology to sense touch contact points. During one round of our experiments, the room had a higher illumination level, forcing our users to press harder to get an equivalent response. This effect wasn't noticed until our users had finished their task, and we could not repeat the experiment at an adjusted illumination level (on the same participants) because our results would be skewed by learning effects. This problem could be avoided either by conducting the entire experiment in one round, or by recording exactly which lights were switched on between rounds.

A third limitation was encountered in our measure for performance. Before the experiment, we believed that the tasks we assigned were impossible to complete within the allotted time. Indeed, we had already reduced the allotted time based on the pilot test (see section A1.1 and item A1.2b for details). This would have allowed us to make comparisons across groups in terms of *completion level per unit time*. However, since almost every group completed their diagram within the allotted time (or less time), our measure became meaningless. If this study were to be repeated, we suggest that tasks of sufficient difficulty are chosen, such that it is inconceivable to finish them within the allotted time.

3. Results

Task Observations

To evaluate collaboration, we used one-way ANOVA with alpha=0.05 to test for differences between interfaces in terms of speech-act counts. Our speech acts were classified into three categories: suggestions, questions, and corrections. A one-way ANOVA was used for each of these categories, for a total of three one-way ANOVAs. When a critical F-value was found, multiple t-tests were used with an adjusted alpha value to assess specific differences between interfaces in terms of speech-act counts.

The same approach was used to evaluate user satisfaction: four one-way ANOVAs – one for each questionnaire question (see appendix A1, or appendix B1.c from the Milestone III report) – were used with alpha=0.05 to test for differences between interfaces in terms of overall user satisfaction. Further t-tests were used to find specific differences between groups.

The following table provides summary statistics for the three interfaces:

	Computer			**Prototype**			**Whiteboard**		
	Q	S	C	Q	S	C	Q	S	C
Sum	24	90	26	51	90	29	67	139	25
Average	4.8	18	5.2	10.2	18	5.8	13.4	27.8	5
Variance	3.7	29.5	13.7	9.7	36.5	0.7	9.3	35.2	6.5

Table 3.0: Summary statistics for observed speech acts on three different interfaces
[Q=questions, S=suggestions, C=corrections]

This is a summary of the ANOVA results on observations:

	F-Statistic	F-Critical (one-tail)
Questions	12.5	3.89
Suggestions	4.75	3.89
Corrections	0.124	3.89

Table 3.1: ANOVA results for questions, suggestions, and corrections

Based on this data, we did not pursue further comparisons in terms of corrections (notice the insignificant F-value), but we did perform multiple t-tests on suggestion counts and correction counts. We obtained a significant result when comparing number of questions posed on the computer vs. prototype, with a t-value of 3.30 vs. a critical value of 2.60. The following chart shows the difference between *all three groups* in terms of questions generated by users:

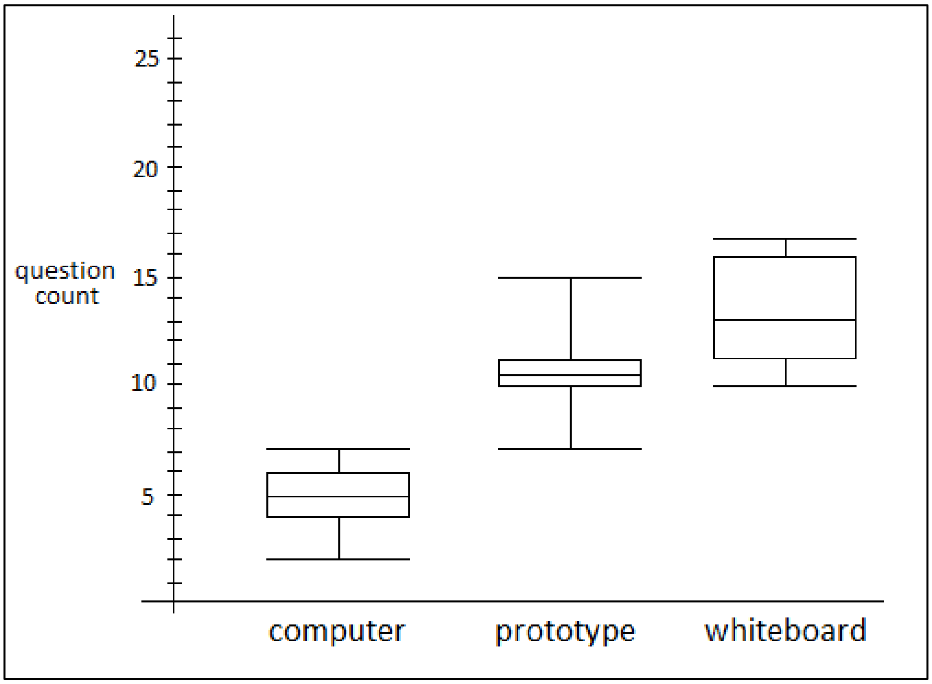

Figure 3.2: Box-and-whisker plot for question count on three interfaces

In terms of suggestions, the t-tests revealed that the prototype did not significantly differ from either the computer or the whiteboard. For more detailed summary statistics, see appendix A3.

Questionnaire

The questionnaire revealed three key results:

- The prototype had a significantly higher mean than the computer in terms of enjoyability (see question 2 from questionnaire in appendix A1), with t-stat = 2.88 vs. t-critical = 2.24.
- The prototype had a significantly higher mean than the computer in terms of whether users would potentially use the system again (see question 4 from questionnaire in appendix A1), with t-stat = 2.59 vs. t-critical = 2.24.

- The whiteboard had a significantly higher mean than the prototype in terms of perceived usability (see question 1 from questionnaire in appendix A1), with a t-stat = 3.13 vs. t-critical = 2.24.

4. Discussion

4.1 Interpretation of results

We found that significantly more questions were generated by users working on our prototype than by users working on the computer. We believe that this may be due to an increase in situational awareness which comes about from increased engagement, as well as an increase in the visual information available to the users. In particular, the positioning of users on a table based interface forces a more open configuration of users, allowing for more open visual cues, including gestures and body language, between users. As discussed by Kraut[4], visual information is a strong conversational resource that can maintain task awareness in collaborative tasks. We feel that our prototype's design enhanced the amount of information visible by inherently providing an open setting for collaborative work, increasing our users focus on the task and thus leading to more questions.

The whiteboard performed significantly better than both the computer and our prototype. We believe this is due to the task difficulty: the tasks chosen were not complex enough for the users to encounter the limitations of the whiteboard. We observed that groups were finishing the UML tasks in the allotted time or sometimes faster on every interface. Similar work by Tolosaa[5] used highly complex task with hundreds of objects, where "completion" was not a reality. We had intended to ask questions that where harder to accomplish and consisted of enough classes so that the manipulation and positioning of those objects became important. In our experiment, this was not the case: most tasks could be completed successfully without ever having to make large structural changes, *ie* manipulating, moving, and/or deleting large numbers of objects at once. Therefore, with increased task complexity, we believe that our prototype could perform at least as well as the whiteboard, if not better.

4.2 Relation to other works

The notion of work as "gaming" is a new and exciting field of research, and it has applications in many areas of inquiry. Work done by Ahn[6] has shown that making tasks "fun" is an effective method of increasing user productivity. Our study showed that people *enjoyed* collaborating together while working on problems on our prototype. The overall look and feel of our interface – including animation,

[4] Robert E. Kraut , Susan R. Fussell , Jane Siegel, Visual information as a conversational resource in collaborative physical tasks, Human-Computer Interaction, v.18 n.1, p.13-49, June 2003

[5] José Barranquero Tolosaa, , Jose E. Labra Gayoa, , Ana B. Martínez Prietoa, , Sheila Méndez Núñeza, and Patricia Ordóñez de Pablos, Interactive web environment for collaborative and extensible diagram based learning, Computers in Human Behavior v. 26, i. 2, P. 210-217, March 2010

[6] Ahn, L. v., Liu, R., Blum, M.: Peekaboom: a game for locating objects in images. Proceedings of the SIGCHI conference on Human Factors in computing systems, Montreal, Quebec, Canada (2006)

color-codings, and multi-touch input – provides a fun way to collaboratively work on diagrams. Thus, our study provides support to research undertaken by Ahn and others in the relationship between work and play.

4.3 Impact for practitioners

In all of our performance metrics we found that the table was no worse than using a computer and in some cases significantly better. Thus, it makes sense to continue developing table-top collaborative diagramming interfaces in an effort to replace existing software ones. Although tabletop interfaces are not common at this time, our data indicates that users found the tabletop interface enjoyable, and could see themselves using such interfaces in the future. We believe that new multi-touch interfaces should be developed which capitalize on the strengths of software solutions (persistence, easy manipulation, ease of sharing) without sacrificing the flexibility of conventional solutions, such as whiteboards and pen-and-paper.

4.4 Critical Reflection

We were not able to keep some variables consistent that we should have. Notably, we discovered that the responsiveness of our prototype changed dramatically depending on the lighting of the room and the lighting was not consistent across tests (see section 2.3 for details). We suspect that this may have lowered the ability of some users to complete the task on the prototype and may have decreased their satisfaction with the interface.

We did not realize how important task design would be in this experiment. The miscalculation in the time that it took to complete the tasks denied us of potentially rich data. We also overlooked the fact that the complexity of a task may have a significant impact on the levels of collaboration required, as well as the role that interface usability has on the ability to complete a task.

We acknowledge potential flaws in our metrics of collaboration, satisfaction, and performance. Our metric of collaboration – comparing counts of different categories of speech-acts – does not take into account meta-effects such as parallelization and delegation of tasks, as well as gestures and body language. Our metric of performance could be confounded by the innate ability of our participants to answer exam-style questions, without giving useful information as to how our users would perform under normal conditions. Our measure of satisfaction using self reporting could be positively skewed due to prior acquaintance with our subjects, and could also be influenced by the group context (see section 2.4 for details).

4.5 Research Agenda

Our research suggests that exploring touch table interfaces could reveal interesting insights into how people can enjoy the collocated collaborative experience. Our results in regards to enjoyment showed that the touch-table scored higher than the computer or the whiteboard despite our prototype being rough around the edges. One interesting avenue of further research would be to determine whether or not this was simply the byproduct of a novel design or the result of something deeper. Many questions could be explored. Were users intrigued by the novel interface, or did they genuinely enjoy working together in the open setting provided by a table interface? Did users feel that the multi-touch input modality increased their productivity? Does this sort of interface have value in the workplace? How can learning curves be further reduced on collocated, multi-touch diagramming tools?

5. Conclusions

The prototype ranks significantly higher than computer software when it came to user satisfaction. The prototype also ranked significantly higher than the computer when users were asked if they would use that diagramming tool again. Our prototype was only outperformed by the whiteboard in terms of one metric: whether users could effectively accomplish their task. We hope that by adding more functionality and increasing usability, this result would change in future experiments. Since our tasks were not as difficult as expected, we were unable to measure the performance outcomes on the three interfaces.

Acknowledgements

We would like to acknowledge Joanna McGrenere and Matthew Brehmer for providing us with a framework under which to conduct this project and for their excellent feedback as to the strengths and weaknesses of our designs. We would also like to acknowledge Jen Fernquist for providing us with access to a Smart Table to run our prototype, instructions on how to develop for it, and a laboratory in which to run our study. We would also like to thank Shan Naziripour for bringing life, humor, talent and beauty to our video.

Appendix A

[A1] Evaluation Instruments

We chose classmates as participants The changes made to these since Milestone 3 are the following:

- We reformatted each piece so that it fit onto an entire page by itself.
- Warm up tasks were added.

The order of the evaluation instruments is as follows:

1. **Questionnaire**
2. **Coding Sheet**
3. **Instructions**
4. **Warm Up Tasks**
5. **Tasks**

[This page left intentionally blank]

Please Put the name of the tool you used next to your answer. Use this sheet for all three tools.

Overall, how well did tool X perform, in terms of letting you do what you needed to accomplish the assigned task(s)?

- Very well
- Well
- Neutral
- Poor
- Very poor

Overall, how enjoyable did you find using tool X to be?

1. Very enjoyable
2. Somewhat enjoyable
3. Neither enjoyable nor unpleasant
4. Somewhat not enjoyable
5. Not enjoyable at all

Overall, how satisfied were you with tool X?

6. Very satisfied
7. Satisfied
8. Neither satisfied nor dissatisfied
9. Dissatisfied
10. Very dissatisfied

Would you use tool X again?

11. Definitely
12. Possibly
13. Not sure
14. Likely not
15. Definitely not

Coding Sheet:

Below tally the number of times participants interact:

Group #

Type of interface: ____________________________________

Draw the participants positioning here:

Question	Suggestions	Corrections

Experiment Instructions:

Your group will be performing a UML-like design task. You will have 5 minutes for each task, and you should focus on completing as much as possible, but you are not expected to finish.

Do not worry about all the specifics of UML diagramming, especially when it comes to differences between arrow, box and line types. Please do try and fill in class names, fields and methods as appropriate. We will be assessing how much you complete and the quality of your team's design.

Whiteboard

At this station you will be using a whiteboard to complete the design. Each of you will receive a different color pen and everyone should use their own pen throughout the process.

Experiment Instructions:

Your group will be performing a UML-like design task. You will have 5 minutes for each task, and you should focus on completing as much as possible, but you are not expected to finish.

Do not worry about all the specifics of UML diagramming, especially when it comes to differences between arrow, box and line types. Please do try and fill in class names, fields and methods as appropriate. We will be assessing how much you complete and the quality of your team's design.

Creately

At this station you will be using a diagramming tool called *Creately*. You will only need to use the box tools and the connectors.

Experiment Instructions:

Your group will be performing a UML-like design task. You will have 5 minutes for each task, and you should focus on completing as much as possible, but you are not expected to finish.

Do not worry about all the specifics of UML diagramming, especially when it comes to differences between arrow, box and line types. Please do try and fill in class names, fields and methods as appropriate. We will be assessing how much you complete and the quality of your team's design.

Prototype

At this station you will be using our prototype.

Please take a minute to familiarize yourself with the tools at your disposal. Inform the facilitator when you are ready to receive your tasks and begin the experiment.

Warm Up Tasks:

Have each member complete one of these tasks:

TASK 1

- create a parent class
  - give it a name
  - add 2 methods to it

TASK 2

- create a subclass
  - give it a name
  - add 2 methods to it
  - create an arrow that points to the parent class.

TASK 3

- create another subclass
  - give it a name
  - add 2 methods to it
  - create an arrow that points to the parent class made in task 1.

1. Draw a UML class diagram for the following software system for modeling a bank.

Each of the bank’s customers can access their account(s) through withdrawals, deposits, or balance inquiries at a bank machine. Each transaction (ie, withdrawal, deposit or balance inquiry) must store the date and time that the transaction occurred. Once a month, a statement that contains a list of all of the transactions that were completed over the last month is generated for each account and mailed to the customer. The bank must be able to produce a list of all of its customers as well as a list of transactions that were completed by a particular bank machine.

2. Draw a UML class diagram for the following software system for modeling a restaurant.

At a restaurant, groups of customers are seated at tables and each table is served by one server. Each customer orders from a menu. The menu contains sub-menus as well as individual items. At the end of the meal, each table is given a bill, which lists all of the items that were ordered by the customers at that table, along with the prices of the items and the total amount owing by the table.

3. Draw a UML class diagram for the following system.

You are writing a retail store management system. The store sells clothing including jeans, tops, jackets, and shoes. The system needs to track the inventory and the sales. For each sale, the system must know the items that were sold, the method of payment and the cashier who completed the transaction. Each cashier has a name, contact information, and a unique id. The inventory includes all of the items that the store sells. Each item has a specific cost associated with it.

[A3] Supplementary Analysis

The following tables provide detailed summary statistics for the *total count of suggestions + corrections + questions* generated during the experiment sessions on each of the three interfaces:

computer	
Mean	28
Standard Error	4.370354677
Median	29
Stdev	9.772410143
Sample Variance	95.5
Kurtosis	0.228118747
Skewness	-0.42592396
Range	26
Minimum	14
Maximum	40
Sum	140
Count	5

prototype	
Mean	34
Standard Error	3.728270376
Median	36
Stdev	8.336666
Sample Variance	69.5
Kurtosis	-2.037058123
Skewness	-0.151018741
Range	20
Minimum	24
Maximum	44
Sum	170
Count	5

whiteboard	
Mean	46.2
Standard Error	3.307567082
Median	43
Stdev	7.395944835
Sample Variance	54.7
Kurtosis	3.693638895
Skewness	1.905041187
Range	18
Minimum	41
Maximum	59
Sum	231
Count	5

[B1] Final Conclusions and Recommendations

Conclusions

Our overall design approach was validated by two key results: the prototype ranked significantly higher than computer software when it came to user satisfaction, and the prototype ranked significantly higher than the computer when users were asked if they would use that diagramming tool again. Our prototype was only outperformed by the whiteboard in terms of one metric: whether users could effectively accomplish their task. In all other respects, it performed no worse than the whiteboard. Of course, adjustments could and should be made in the future to improve upon our basic design. Adding more functionality and increasing smoothness of touch gestures would be good ideas.

Recommendations

Based on our positive findings, we believe our design direction to be a worthwhile and viable direction for collaborative, collocated diagramming. Further research should be conducted with an improved prototype, with additional functionality that replaces and improves on some of the other main negative aspects of incumbent interfaces.

[B2] Reflection

Design Process

An interesting aspect of the user-centered design process was the opportunity to use the process itself to generate ideas. Due to the fact that we were *collaborating as a team* to build a project centered on collaboration, we had a lot of opportunities to reflect on our own behavior and use that reflection to provide direction to our work. For example, when we initially began the project we started using some commercial diagramming software to sketch out some rough ideas. We immediately realized the limitations: a group of five of us we were huddled around a single computer, trying to hash out ideas in a creative and freeform way. It was near impossible. We took inspiration from this and decided to work on creating a better collaborative tool.

Another interesting aspect was the pilot study. We found it very useful and informative. After piloting our experiment we gained an immense amount of insight into the workings of our experiment and ways in which it could be improved. However, in an experiment of this complexity, one pilot was not enough, as there were still major problems in our experiment design that it did not uncover (such as task completion time).

Despite the advantages of our design process, aspects of it were still inadequate. One aspect of our design that could be vastly improved was in regards to our encoding technique. Individual variability in encoding technique was not properly accounted for in any way, as the same person was an encoder for the same application. It would have been better if we had rotated our encoders around to different applications in order to minimize the amount of potential bias. Encoding is a difficult task and natural differences in style and accounting could have resulted in data that was not indicative of our interface.

Another problem was that we did not perform enough iterations to fully iron out the design issues in our prototype. This was primarily a time constraint issue (3 months to design and test). However, it is still worth mentioning that many iterations with constant feedback from users is the best method to create effective software.

Methods

One of our most valuable methods of driving the direction of our project was interviewing users at every opportunity possible. All members of the group tried to gather information informally through friends and family members regarding collocated, collaborative diagramming, in order to make our actual questionnaires, interviews and experiment as solid as possible.

We observed an important limitation of questionnaires as experimental tools: you need to think of every possible question beforehand. An example was our questionnaire following the experiment, which focused on user satisfaction. Post experiment, we realized that it was lacking questions regarding how subjects felt the tools helped them in *collaboratively* completing the task at hand. But we did not think of that beforehand, and when the experiment was in progress, it was too late to change the questionnaires. We feel that this problem limits the usefulness of questionnaires.