

Investigating the Effects of Tutorials on Task Performance in Usability Evaluations

Jeff Sauro

1/2/2005

Stanford University

jeff@measuringusability.com

Abstract

Six participants attempted to complete common tasks with and without tutorials in an address-book application as part of an enterprise web-based software application. The results indicate that providing training materials to users prior to a usability test will likely have a significant effect on task time, number of errors committed and the usability score, but most likely not on perceived ease-of-use. Adding an interactive component to the tutorials will most likely not have any noticeable effect on task performance.

Introduction

A challenge to Human Factors Engineers is being able to assess the usability of a new product with actual users, when by definition, no users exist that have experience with the application. Training users prior to a usability assessment is regarded as a better measure of usability than testing users who have no experience (Rubin 1994). Without some controlled introduction to the User Interface, a usability test measures more “ease of learning” than “ease of use.” Providing brief training materials to participants before a usability test should equip participants with enough exposure to allow them to complete the tasks. This may simulate the type of behavior one might expect of a user with one week of system experience. There are many aspects of training materials that could affect task performance such as subtle variations in content, length, narration, number of steps, the presentation medium and whether the material is interactive. In addition to examining the affects of introducing training materials into a usability test, this study will also examine what affect interactivity in the training materials will have on task performance. This study will then seek to answer the following questions:

1. Is there a significant and practical significance in task performance during a usability evaluation when training materials are used?
2. Is there a significant difference in performance when training materials are presented in a passive versus interactive format?

Methods

Six participants performed seven common tasks on an Address Book application. Prior to executing a task, participants were given training materials (either interactive, or passive). The tasks and training materials were counterbalanced across participants. Measures of task time, number of errors, completion rates, subjective satisfaction ratings and a composite “usability score” were collected to reveal what effects training materials have on task performance in a usability test.

Application Tested

A web-based address-book application part of a large enterprise software suite was tested. The Address Book (as the name suggest) is used to store the names, addresses, phone numbers and a variety of other contact information for a business or individual.

Participants

Six participants (four male two female), all employees of the same software company in Denver, Colorado participated in the study. They were selected by convenience and the only screening requirement was that they don't use the application as part of their job. The participants varied in their exposure to the address-book application; three had never seen the application before and the remaining three hadn't had any exposure within the last year.

Materials

The two major goals of this study relied on presenting participants training materials (tutorials) prior to completing a task. The training materials consisted of 1 to 2 minute power-point presentations with voice narration and areas of the screen were highlighted using red-circles. The slides were automatically advanced during the presentation. At the end of each tutorial, two key points were reviewed. A screen shot of one of the slides with a section of transcribed narration is available in the Appendix. There were two types of tutorials created for three of the seven tasks—one "interactive" and one "passive." Both passive and interactive tutorials were identical except in the review section. The interactive review asked the participant to click on the area of the presentation where a certain function could be found that was integral in completing the task. The passive review contained the same review, except the area of the screen where the function was located was highlighted with a red circle. Both tutorials were identical in length and content.

Task Descriptions

The address book application has many features and functions, the participants were asked to complete seven total tasks reflecting basic functionality of the application. Three of the tasks were repeated with only minor details changed providing a repeated measure of the same task-type. In total, four task-types were tested.

Task Descriptions (Iterations)

Add a New Record to the Address Book (x2)

Add a contact to an existing record (x2)

Delete a record from the Address Book (x2)

Add an email address to an existing contact (x1)

Design.

The experiment was a within-subjects 3 x 3 x 2 mixed design. The factors included training type (none, passive, interactive) task type (new record, add contact, delete record) and task attempt (1st and 2nd). The tasks, training type and task order were counterbalanced across participants using a Greco-square: see Table 1 below.

Table 1: Counterbalanced Design of Training and Task by Participant

Participant	Training	Task	Sequence
1	None	Delete a Record	1
1	Passive	Add a Contact	2
1	Interactive	New Record	3
2	Passive	New Record	4
2	Interactive	Delete a Record	5
2	None	Add a Contact	6
3	Interactive	Add a Contact	7
3	None	New Record	8
3	Passive	Delete a Record	9
4	Interactive	New Record	10
4	Passive	Add a Contact	11
4	None	Delete a Record	12
5	None	Add a Contact	13
5	Interactive	Delete a Record	14
5	Passive	New Record	15
6	Passive	Delete a Record	16
6	None	New Record	17
6	Interactive	Add a Contact	18

Participants received training for two of the three tasks. After the counterbalanced sequence of tasks and training materials, all users attempted a fourth task (Add an email address) then attempted three more tasks that were slight variations on the first three tasks. These latter three tasks were different only in minor details (e.g. a different mailing address) and were presented in the same sequence as the first three tasks. For example, participant one's task order is displayed in Table 2 below.

Table 2: Participant One's Task Order (Treatment)

1. Delete a Record (no training)
2. Add a Contact (Passive)
3. New Record (Interactive)
4. Add an Email Address (no training)
5. Delete a Record 2 (no training)
6. Add a Contact 2 (no training)
7. New Record 2 (no training)

Dependent Measures

Time to complete the task, number of errors committed, task completion were used as the measures to look for differences in performance. Post-task subjective satisfaction was used to assess the participants perceived ease of use. A composite of these four variables

will also be used as a measure of overall “usability” (Sauro and Kindlund 2005). The composite measure is derived by combining the four measures in a Principal Components Analysis and storing the scores of the first Principal Component. See the Coding section for more details on how the dependent measures were recorded.

Procedure

Participants were told by the test administrator to imagine they worked for a company that would have a need for address book software to manage information such as company names, addresses and phone numbers. Participants were told to do the best they could on each task and not to speak to the test administrator while completing the task. The test administrator sat behind the participant in the same room.

Prior to completing a task, the participant received a task-instruction sheet (see appendix) that explained what they needed to do. After the participants said they understood the task, they received a power-point tutorial which explained how to accomplish the task. Participants only received a tutorial (interactive or passive) for two of the first three tasks (based on the counterbalanced design). The only difference in tutorials was whether the participant was asked to click in the tutorial. Each participant then completed the same “Add an Email Address” task which no one had been trained on. Tasks 5-7 were slight variations on the first three tasks and were presented in the same order so there were three other tasks between the first exposure to the task-type.

Once the participant verbally acknowledged they were done completing the task and the post-task questionnaire, the administrator answered any questions and followed up on any particular problems noticed during the task.

Coding

Task Time

Task time began when the participants verbally told the administrator they were ready to begin the task and ended when they verbally told the administrator they were done. Time was still running even if a participant had technically completed all aspects of the task but was still carrying out activities (such as verifying a record was deleted). Time was recorded in seconds.

Errors

Errors were defined as any unintentional activity, an activity that caused data loss or any activity not directly contributing to the successful completion of the task. Errors were recorded as counts.

Task Completion

Task completion was defined as a user successfully completing all required aspects of a task. If a participant left off any portion of the required aspects of a task the task was considered a failed task. Task completion was coded as a dummy variable 1 (success) or failure (0).

Satisfaction

A questionnaire was given to the participants immediately following each task (available in the appendix). A participant's satisfaction was derived by taking the average of the first two questions in the questionnaire (ease and time). For example, if a participant answered 6 and 7 for one task, then satisfaction for that task was coded as a 6.5.

Usability Score

Once the four measures described above were coded, a Principal Components Analysis was run on all observations. Principal Components Analysis (PCA) is a statistical technique that linearly transforms an original set of variables into a smaller set of uncorrelated variables that represents most of the information in the original set of variables. Its goal is to reduce the dimensionality of the original data set (Jolliffe 2002). The scores of the first principal component, containing the most variance, were stored and used to represent the best combination of the other four variables—a measure of usability (Sauro and Kindlund 2005). The values are standardized similar to a z-score and higher values indicate better usability.

An example of the values for participants 1 and 2 are presented in Table 3 below.

Table 3: Participants One and Two Dependent Measures by Treatment (In Task Order Sequence)

Participant	Training	Task	Time	Completion	Errors	Satisfaction	Usability
1	None	Delete a Record	155	1	2	1.5	-0.995
1	Passive	Add a Contact	182	1	2	1.5	-1.209
1	Interactive	New Record	245	1	0	1.5	-0.772
1	None	Add Email	63	1	0	2.0	0.730
1	None	Delete a Record 2	34	1	0	1.5	0.903
1	None	Add a Contact 2	98	1	1	2.0	-0.016
1	None	New Record 2	165	1	0	3.0	0.035
2	Passive	New Record	180	1	1	1.5	-0.725
2	Interactive	Delete a Record	52	1	0	1.0	0.702
2	None	Add a Contact	265	1	4	4.5	-2.460
2	None	Add Email	158	1	2	1.5	-1.018
2	None	New Record 2	124	0	1	2.5	-0.805
2	None	Delete a Record 2	54	1	1	1.5	0.275
2	None	Add a Contact 2	111	1	1	3.0	-0.004

UI Manipulation Speed

At the beginning of the usability test and prior to completing any tasks, participants completed a Graphical User Interface pretest which measured the amount of time it took to answer simple questions using basic web-based user interface objects. This time was used as a controlled covariate to mitigate the effects of unwanted individual differences in task completion times. That is, if a participant happens to complete a simple task requiring little cognitive effort and no domain knowledge slowly, we would also expect their domain specific tasks to be completed slowly and we wouldn't want that to negatively affect the task time. The assessment is available here:

http://www.measuringusability.com/survey/task_main.htm under Warm-Up

Results

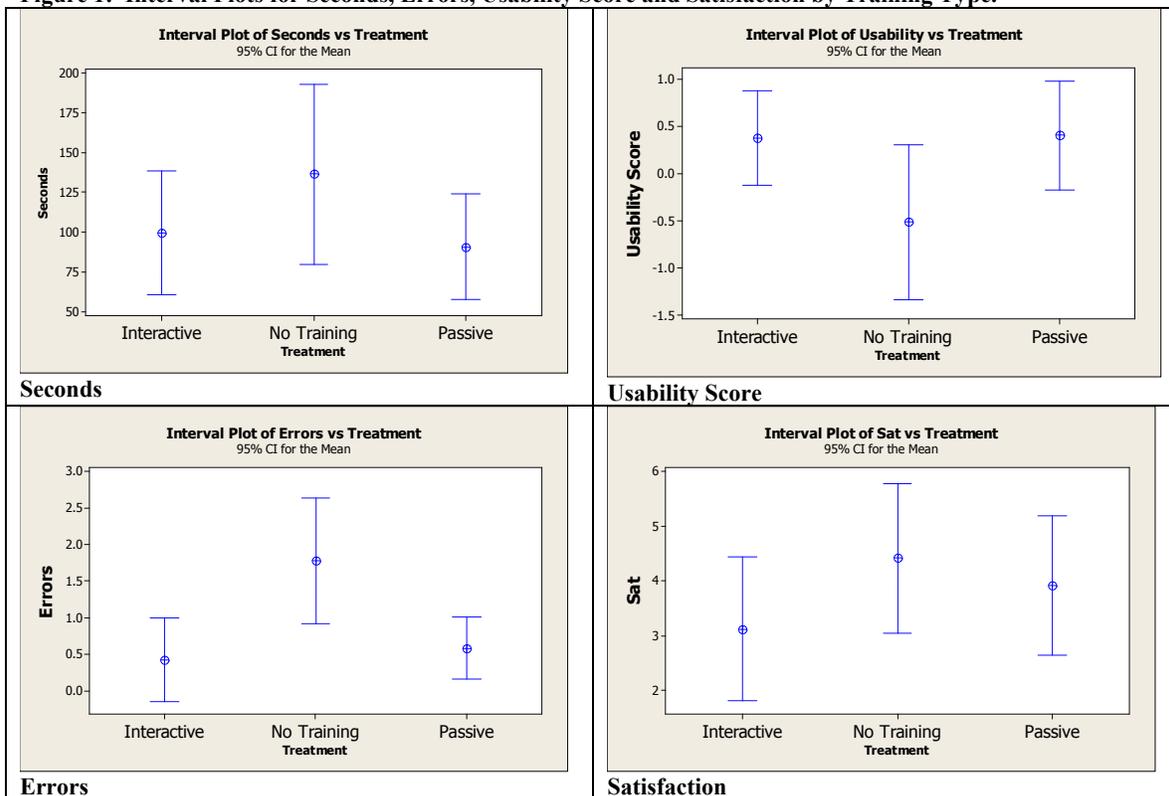
The means and standard deviations for each dependent measure for the three levels of the training factor are displayed in Table 4. Figure 1 displays the means and 95% Confidence Intervals by Training Type. Task four was excluded from the table and figure to keep cell counts equal (12 observations per cell).

Table 4: Mean and Standard Deviation for all Dependent Variables by Treatment

Training	Mean (SD)				
	Time	Completion	Errors	Satisfaction	Usability
None	164 (130)	83 %	2.1 (.5)	4.4 (2.2)	-.877 (1.88)
Passive	91 (53)	92 %	.58 (.2)	3.9 (2.0)	.404 (.911)
Interactive	100 (62)	100 %	.42 (.9)	3.1 (2.1)	.373 (.791)

N = 12 per cell

Figure 1: Interval Plots for Seconds, Errors, Usability Score and Satisfaction by Training Type.



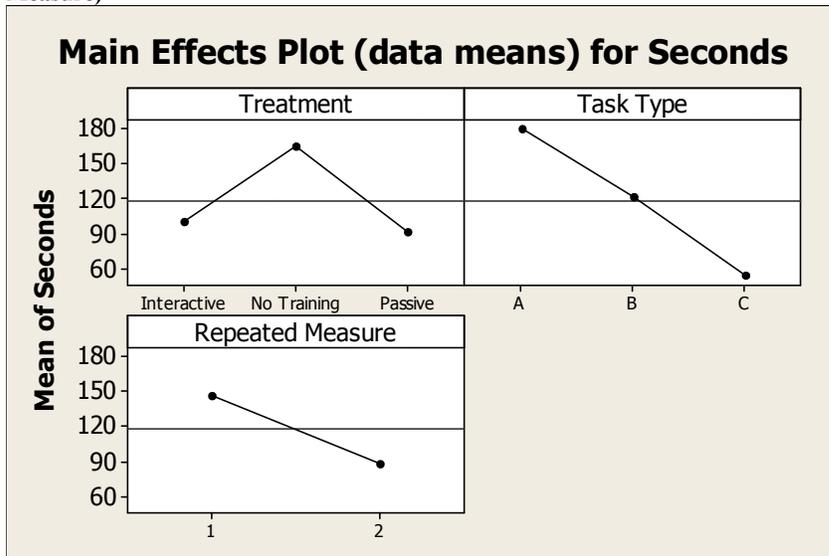
Training Effect on Performance

The first research question asked if there is a significant difference in performance when participants receive training in a usability test. Main effects and interaction effects between Task Type (A: *New Record*, B: *Add Contact*, C: *Delete Record*), Training Type (None, Passive, Interactive) and Repeated Measure (1st Exposure, 2nd Exposure) were examined using a General Linear Model for all dependent variable except task completion (which was coded using a dummy 1/0 variable). The participants “UI Manipulation Speed” was used as a covariate to control for the individual differences in using the mouse, typing, clicking and locating objects on the screen.

Task Time

For the dependent variable Task Time, there were significant differences between different training types [$F(2, 35) = 4.80$ $p < .05$] task types [$F(2, 35) = 11.84$ $p < .01$] and the second attempt at the task [$F(1, 35) = 7.73$ $p < .05$]. Figure 2 displays the main effects for task time. There were no interactions between any of the factors and seconds.

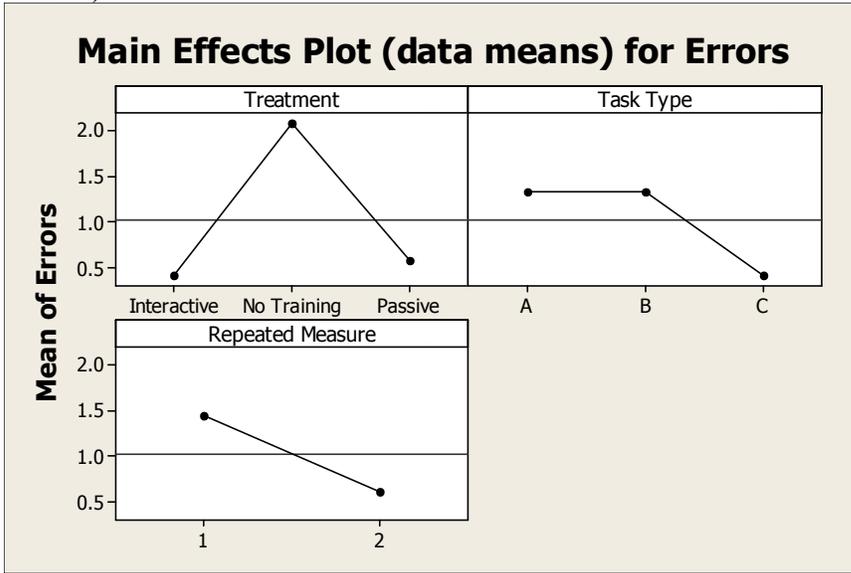
Figure 2: Task Time Main Effects Plot for Training Type (Treatment), Task Type and Task Attempt (Repeated Measure)



Errors

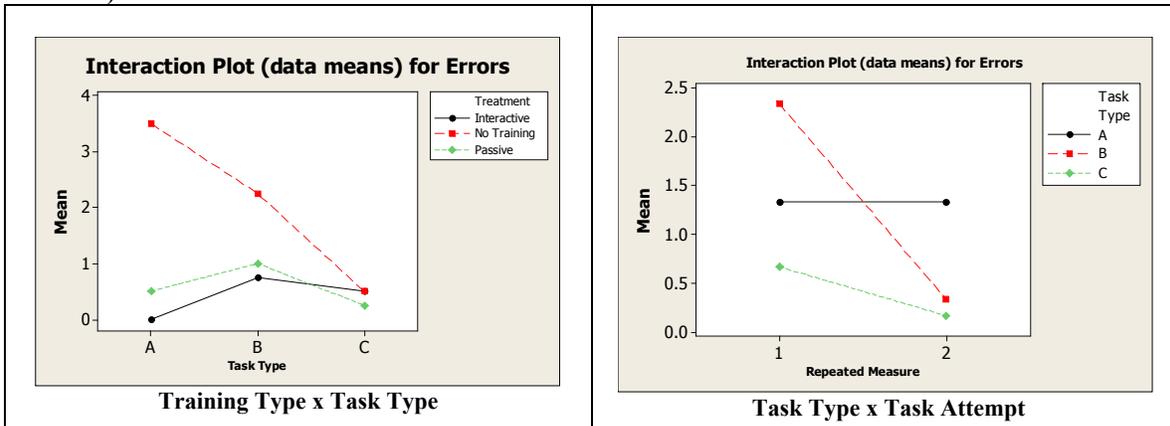
For the dependent variable Errors, there were significant differences between different training types [$F(2, 35) = 14.71$ $p < .01$] task types [$F(2, 35) = 4.89$ $p < .05$] and the second attempt at the task [$F(1, 35) = 9.09$ $p < .01$]. Figure 3 displays the main effects for errors.

Figure 3: Errors Main Effects Plot for Training Type (Treatment), Task Type and Task Attempt (Repeated Measure)



There were two significant interaction effects between task type and training type [$F(4, 35) = 4.63$ $p < .01$] and between task attempt and task type [$F(2, 35) = 4.73$ $p < .05$]. The interaction plots are displayed in Figure 4.

Figure 4: Errors Interaction Plots for Training Type (Treatment) and Task Type by Task Attempt (Repeated Measure)



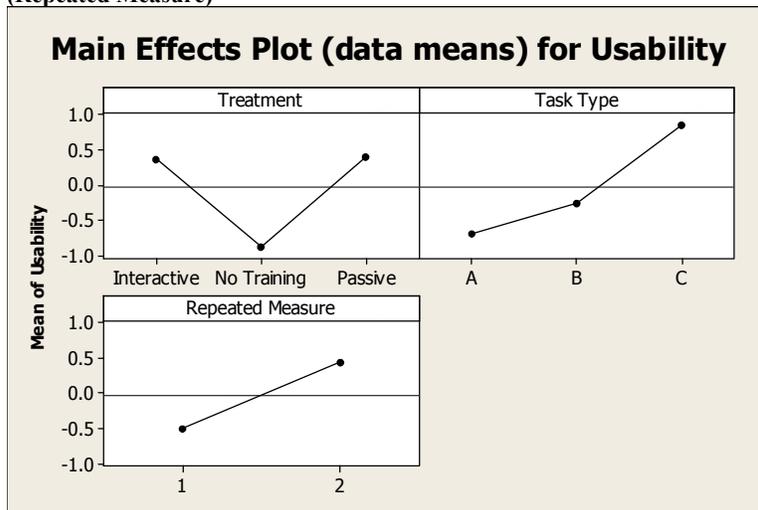
Satisfaction

While satisfaction appeared to increase between task attempts (see Figure 6 below) the differences weren't significant. There were no other significant main effects or interactions in satisfaction observed.

Usability Score

For the dependent variable Usability Score, there were significant differences between different training types [$F(2, 35) = 7.60$ $p < .01$] task types [$F(2, 35) = 9.14$ $p < .01$] and the second attempt at the task [$F(1, 35) = 9.28$ $p < .01$]. Figure 6 displays the main effects for the usability score. There were no significant interactions between the usability score and any of the factors.

Figure 5: Usability Score Main Effects Plot for Training Type (Treatment), Task Type and Task Attempt (Repeated Measure)



Completion Rates

There were only three total task failures out of the 36 observations (2 in the no training group and 1 in the Passive Training group). Since cell counts were so low, both Passive and Interactive treatments were collapsed into one cell—"Training". To compare completion rates, a Fisher Exact test was run on the consolidated training group against the no training group. The results are displayed in Table 6 below.

Table 6: Fisher’s Exact Test on the 2 x 2 Contingency Table for Training by No Training

Treatment	Failed Task	Completed Task	Total
No Training	2	10	12
Training	1	23	24
All	3	33	36
$p = .252$			

While there is not a significant difference in task completion between the training and no training groups ($p > .25$), the odds that a participant will fail a task without training is 4 times greater than a participant failing with training (odds ratio $\theta = .166 / .042 = 4$). When the fourth task that no participants received training on (Add an Email Address) is included in the contingency table, the odds of task failure decrease (odds ratio $\theta = .111 / .042 = 2.6$). Table 7 displays the cell values.

Table 7: Fisher’s Exact Test on the 2 x 2 Contingency Table for Training by No Training Includes 4th Task Type

Treatment	Failed Task	Completed Task	Total
No Training	2	16	18
Training	1	23	24
All	3	39	42
$p = .567$			

Repeated Measure Analysis

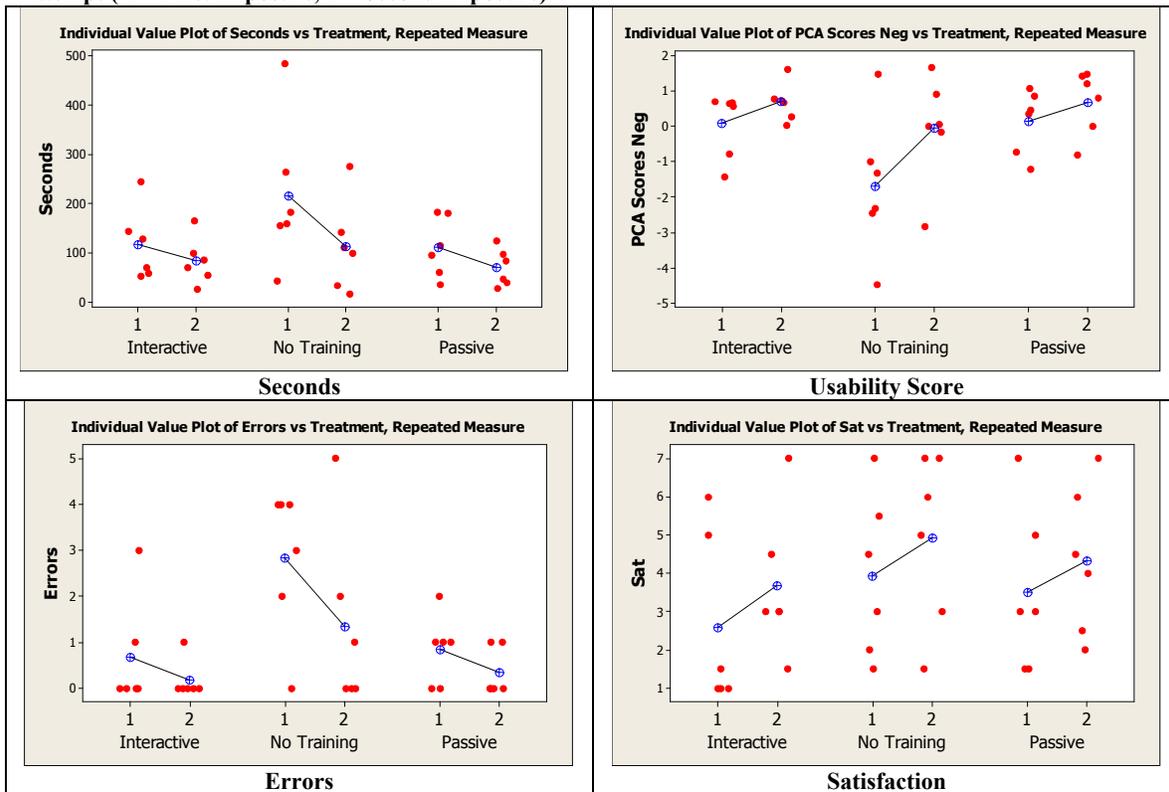
Participants attempted each task twice to provide a way to see if the effects of different training types might affect performance on later task attempts differently. As shown above in the main effects plots (Figures 4-6), the second task attempt shows a significant difference in performance from the 1st attempt for task time, errors and the usability score. The mean and standard deviations are displayed in Table 8 and a visual representation of the differences is also displayed in Figure 6.

Table 8: Mean and Standard Deviation for Dependent Variables by Training Type and Task Attempt (1st and 2nd) (Completion Rates Excluded)

Training	Mean (SD)							
	Time		Errors		Sat		Usability	
	1 st	2 nd						
None	215 (150)	113 (93)	2.8 (1.6)	1.3 (1.9)	3.9 (2.1)	4.9 (2.2)	-1.7 (1.9)	-.06 (1.5)
Passive	111 (60)	70 (38)	.83 (.75)	.33 (.52)	3.5 (2.1)	4.3 (1.9)	.13 (.90)	.68 (.91)
Interactive	116 (74)	83 (48)	.67 (1.2)	.16 (.41)	2.6 (2.3)	3.7 (1.9)	.06 (.93)	.68 (.53)

N = 6 per cell

Figure 6: Individual Value Plots for Seconds, Errors, Usability Score and Satisfaction by Treatment and Task Attempt (1 = First Exposure, 2 = Second Exposure)



A visual inspection of the interval plots suggests that tasks with no training showed larger differences. Participants’ second task attempt was subtracted from their first task attempt providing “difference-scores” for the dependent measures. Taking advantage of the within-subjects design, two paired t-tests were run (between interactive and no training and between passive and no training). The results are displayed in Table 9 and confirm that tasks without training show a significantly larger ($p < .10$) difference between measures for task time and the usability score.

Table 9: Differences in Second Task Attempt from First Task Attempt by Dependent Measure per Participant (Int = Interactive Training, Pass = Passive Training)

	Task Time			Errors			Satisfaction			Usability Score		
	Int.	None	Pass.	Int.	None	Pass.	Int.	None	Pass	Int.	None	Pass
1	80	121	84	0	-2	-1	1.5	0.0	0.5	0.81	1.90	1.19
2	-2	154	56	1	-3	0	0.5	-1.5	1.0	-0.43	2.46	-0.08
3	58	208	68	-3	1	-1	2.0	4.0	1.0	2.10	1.64	1.12
4	29	26	13	0	0	-1	-1.5	0.0	0.0	0.06	0.21	0.57
5	45	41	31	-1	-4	0	2.0	1.5	1.0	1.06	2.37	0.36
6	-12	59	-3	0	-1	0	2.0	2.0	1.5	0.13	1.17	0.15
Mean	33	101.5	41.5	-0.5	-1.5	-0.5	-0.5	1	0.83	0.83	1.63	0.55
SD	35.3	71.6	33.5	1.4	1.9	0.5	0.5	1.9	0.50	0.50	0.80	0.50
<i>p</i>*	.065	.035		.426	.332		.903	.828		.084	.054	

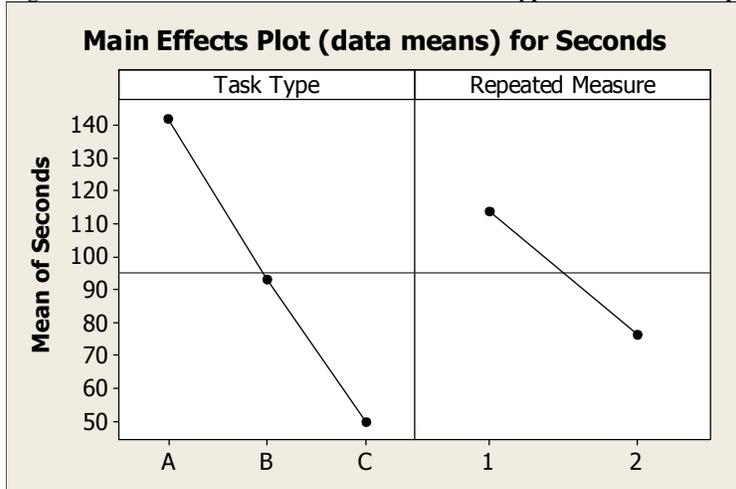
* The bottom row of the table represents the p values for the paired t-tests. The first value for each dependent measure is the result of the paired t-test between Interactive and No Training and the second value is between Passive and No Training. Bold values represent significant differences ($p < .10$)

Differences between Passive & Interactive Training

The second research question asked if there is a significant difference in performance when participants receive different types of training (interactive vs. passive).

To assess the differences between just the two types of training, the “No Training” level was removed from the Training factor. All significant main effects and interactions reported above were no longer significant with the exception of task type [$F(2, 23) = 10.67$, $p < .01$] and task attempt [$F(1, 23) = 4.96$, $p < .05$] for only the dependent variable task time. These two main effects are displayed in Figure 7.

Figure 7: Task Time Main Effects Plot for Task Type and Task Attempt (Repeated Measure)



Repeated Measure Analysis

When the “No Training” group was removed from analysis the differences between the first and second attempts disappeared except for in task time as seen in Figure 7 above.

Discussion

The evidence suggests training will affect task performance in a usability test. The results show that even with this small sample size that performance as measured by task time, errors and the composite usability score are significantly affected when training is presented to subjects prior to task completion in a usability test. While a participant's performance may be affected by training, perceived ease (as measured by the post-task satisfaction questions) only had a moderate but not significant increase (see Figure 6 above).

The data from this study do not suggest that there is a significant difference between passive and interactive training. When the "No Training" condition was removed from the analysis, the only main effects observed were differences in time between task types and between the first and second attempts at the task—not between the training levels (passive and interactive).

There were no significant differences due to the different training types for any of the dependent measures or any interactions or were there any differences in the second exposure to the task.

There was however a significant difference in the difference-scores between participants' first and second task attempt for task time and usability. While task performance improves on subsequent attempts, the degree of change is greater for participants without training. This suggests that the first task exposure is used by participants to get acquainted with the application and in a sense, make up for the orientation they are not receiving from training.

General Discussion

The differences between the active and interactive training programs were very subtle. Perhaps a larger sample size in future studies would provide more power to notice smaller effects if they exist. More power may also provide more data to discern whether satisfaction is significantly affected by training.

Task selection has an effect on task performance and perception (most of them significant). Future analyses should include more tasks that vary in difficulty and length to better understand how usability is affected by task complexity. With only three task failures, task completion rates were not a good way to distinguish between differing performances. Future analyses might include more difficult tasks to improve the value of this often quoted and easily understood metric.

Performance increases between task attempts slightly more for participants who didn't receive any training prior to a task. This suggests that participants learn more from first attempting the task and then transfer that experience to the next exposure. Not having any training however still increases the odds that a participant will make the same mistake repeatedly (transferring incorrect steps). For example, one participant who didn't receive training on task type B: *Add a Contact to a Record*, made the same critical mistake (added a "Related Person" instead of a Contact) on his first attempt and second attempt of the task. Future analysis may also examine to what degree exposure to a task provides a similar degree of performance as training (assuming the user is performing the task correctly).

Errors as a dependent variable for this analysis turned out to be sensitive to the subtle differences in task-types. It is the only dependent variable that showed a significant interaction (task-types and repeated measure, see Figure 4 above) as such it should be included in future analyses. The usability score also provided value in its ability to summarize the correlations between errors, task time, satisfaction and completion rates into one dependent variable. What's more, the usability score includes task completion rates (coded as a dummy binary variable) which allows this aspect of usability to be included into statistical analyses such as the ANOVA that prove crucial in discerning experimental effects.

Finally as a general observation, while participants were watching the tutorials they frequently would move the mouse at certain points during the tutorial when no mouse movements were required. This was especially noticeable in participants who were exposed to the interactive tutorials first. When these participants were exposed to the passive tutorial, they were anticipating mouse interaction and moved their mouse-pointer across the tutorial highlighting the salient features and points. This may suggest that participants, even when not explicitly cued to use their mouse to interact with the tutorial, were doing so as a way of getting to know the application using proxy-spatial manipulations. This behavior continued during the task scenarios as participants would gyrate their mouse pointer, sometimes quite erratically as a way for them to receive feedback from the application. While many of the actions were pragmatic (e.g. moving

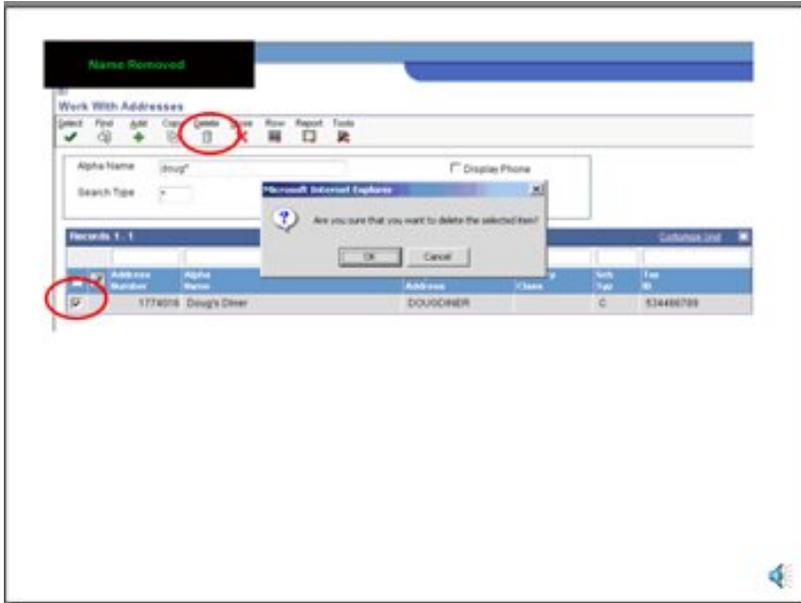
the mouse to more quickly see if an hour-glass returned to a cursor after a search process has ended) many actions appeared epistemic (Kirsh, & Maglio, P. 1994). The actions didn't appear to move the participant closer to task completion; rather they appeared to serve as a way of orienting the participant to the architecture of the interface. For example, while looking for an icon, participants would often encircle the icons with their mouse as a way of eliminating where they've been. If possible, future analyses should measure these actions and see if the interactive tutorials increase them and in some way lead to better task performance.

References

1. Jolliffe, Ian T.(2002). Principal Component Analysis. Secaucus, NJ, USA: Springer-Verlag.
2. Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513-549.
3. Rubin, J.(1994) “Handbook of Usability Testing”
4. Sauro, J. & Kindlund E. (2005) "[A Method to Standardize Usability Metrics into a Single Score.](#)" in *Proceedings of the Conference in Human Factors in Computing Systems (CHI 2005)* Portland, OR.

Appendix

Screen Shot from Delete Address Book Record Task Tutorial.



Transcribed Narration

“To delete the record from the address book, first select the check-box next to the item in the grid. Next, Select the “Delete” Icon from the took-bar. You will be presented with a confirmation dialogue. Click “OK” and the address book record will be deleted from the system.”

Task Instruction Sheet for the Create a Customer Record Task

Task 1 – Create a Customer Address Book Record

Bob's Book Emporium is one of your new customers and therefore you need to add them to your address book.

- Add the customer record using the information listed below.

Name	Bob's Books Emporium
Long Address Number	BobsBooks
Tax ID	865530900
Search Type	C
Address	4200 E 9 th Ave Kansas City, MO 64101 USA
Phone Number – Business	402-654-9511

- Save your changes
- Inform the facilitator when finished.

Post-Task Questionnaire

I was satisfied with the ease of completing this task

Strongly Disagree Strongly Agree

1 2 3 4 5 6 7

I was satisfied with the amount of time it took to complete the task.

Strongly Disagree Strongly Agree

1 2 3 4 5 6 7

In comparison to the product I currently use to perform this task, this application was

Much Harder to Use Much Easier to Use

1 2 3 4 5 6 7 N/A

How often do you complete this task in your job?

Never Once a Year Monthly Weekly Daily Multiple Times a Day

Comments