## Results

We first present the results of the evaluation questionnaires, followed by the comparison questionnaire. The evaluation questionnaire was administered immediately after each 15-minute session and is modeled on the NASA-TLX workload scale. Subjects rank five statements on a 7-point Likert scale, with 1 as “Very Low” and 7 as “Very High”.

The questions were:

1. How mentally demanding was the task?
2. How hurried or rushed was the pace of the task?
3. How successful do you think you were in accomplishing what you were asked to do?
4. How effectively did the browser support what you were asked to do?
5. How insecure, discouraged, irritated, stressed, and annoyed were you?

To analyze these responses we used the Wilcoxon Matched-Pairs Signed-Ranks, a non-parametric test, because the data is ordinal. The matched-pairs signed-ranks are shown in Table 1.

|  |  | N | Mean Rank | Sum of Ranks |
| --- | --- | --- | --- | --- |
| Q1T4 - Q1FF | Negative Ranks | 1a | 1.50 | 1.50 |
| Positive Ranks | 2b | 2.25 | 4.50 |
| Ties | 5c |  |  |
| Total | 8 |  |  |
| Q2T4 - Q2FF | Negative Ranks | 1d | 3.50 | 3.50 |
| Positive Ranks | 4e | 2.88 | 11.50 |
| Ties | 3f |  |  |
| Total | 8 |  |  |
| Q3T4 - Q3FF | Negative Ranks | 2g | 3.00 | 6.00 |
| Positive Ranks | 2h | 2.00 | 4.00 |
| Ties | 4i |  |  |
| Total | 8 |  |  |
| Q4T4 - Q4FF | Negative Ranks | 2j | 3.75 | 7.50 |
| Positive Ranks | 6k | 4.75 | 28.50 |
| Ties | 0l |  |  |
| Total | 8 |  |  |
| Q5T4 - Q5FF | Negative Ranks | 3m | 2.33 | 7.00 |
| Positive Ranks | 2n | 4.00 | 8.00 |
| Ties | 3o |  |  |
| Total | 8 |  |  |

Table : Table of signed ranks for the five questions from the evaluation questionnaire.

The results of the Wilconox test are shown in Table 2. All results were not statistically significant, p > 0.10. While the null hypothesis, which states that the workload while using each browser is identical, is tenable, we believe that this is almost certainly due to insufficient power rather than a true equality of workloads. Boxplots of the data are given in Figure 1; in particular, the responses Q4 (p=0.135) appears to be different between browsers but no statistically significant result was found.

|  | Q1T4 - Q1FF | Q2T4 - Q2FF | Q3T4 - Q3FF | Q4T4 - Q4FF | Q5T4 - Q5FF |
| --- | --- | --- | --- | --- | --- |
| Z | -.816a | -1.089a | -.378b | -1.496a | -.135a |
| Asymp. Sig. (2-tailed) | .414 | .276 | .705 | .135 | .892 |
| a. Based on negative ranks. | | | | | |
| b. Based on positive ranks. | | | | | |

Table : Results of the Wilcoxon Matched-Pairs Signed ranks test. No results were statistically significant.



Figure : Boxplots of data from the evaluation questionnaire. In particular, Q4 seems to show a difference, although it was not significantly different.

At the end of the experimental session, each subject completed another questionnaire which contained questions comparing the two conditions, as well as general questions about task-based browsing. Questions 2 and 3 were forced-choice yes/no questions, with room for comments. In questions 1, 6, and 7, subjects choose between the TabFour browser and Mozilla Firefox; two subjects selected both browsers for Q6 and noted that they perceived no difference. Questions 4, 5, and 8 were free-response. For precise wording of the questions, refer to Appendix X.

All subjects reported that it was easier to complete the tasks using TabFour. All subjects reported that task-based browsing was useful, and that they would use task-based browsing if it was available in their regular web browser. Also, all subjects reported that it was easier to share webpages with other people in the TabFour browser.

Results were more mixed when subjects were asked in which browser it was easier to return to previously used webpages. The TabFour browser was chosen by 5 subjects, but 2 reported no preference and 1 preferred Firefox, stating that “Firefox lets you see all things at once.” In TabFour a user may have to first select the desired task, and then select the desired tab within that task; if only a moderate number of tabs are opened the same result may be achieved by directly selecting the desired tab. We omit analysing this question with a statistical test since they have insufficient power to effectively deal with a data set this small; only a unanimous result is significant.

## Qualitative Results

We received a moderate amount of feedback in the comment areas of the questionnaires and from the free-response questions, which is presented here.

Users frequently reported that they found things more organized in TabFour than in Firefox. For example, P1 states that “[TabFour]’s a lot more organized with the subcategories.” P4 reports, “I like the nested tabs, more organized,“ and according to P7, “The organization is great to keep everything in order.”

Users also frequently commented that it was much easier to get the shared links from their “friend” through the online shared tasks in TabFour, as opposed to receive a list of links in Windows Live Messenger. Subject P8 “Liked the TabFour browser better than having to open all the new tabs myself in Firefox.” P7 reports that “It was a lot easier to use the tasks [Jane] made me” in TabFour. According to P4, the “Context file makes it quicker,” and P5 says that “It was easier than trying to separately click and open each link.”

User practices were mixed when it came to how they would complete the tasks. The first interruption was given a sense of urgency, to encourage subjects to switch tasks. At least one user, however, continued working on the main task until completion, and only switched to the secondary tasks at the end: he stated that “I always stuck to one task. I ignored Mark.” Most users did switch tasks throughout the experiment, and reported that switching was easier in TabFour. For example, P5 reports that in TabFour “It was easier to jump in to a new task whenever I wanted and to leave the first task temporarily unfinished.”

## Limitations

The chief limitation of our evaluation is that it is extremely targeted on only a few features of the TabFour browser. Participants made heavy use of switching between tasks, and of loading previously shared tasks from the internet. Other features were optional and used only lightly if at all. For example, users saw no value in creating bookmarks since they knew they would only be using the system for a short time. The number of tasks used was still reasonably small and all tasks could fit in the task bar, so users had no need to suspend a task in order to move it to the “Inactive” sidebar. A longer field study could reveal if this feature really does aid resumption of tasks after longer periods such as days or weeks, as hypothesized. Finally, there was no format specified for the users’ response during the study; they almost all extracted the relevant information from a page and typed it into an email or instant message manually rather than using the drag-and-drop feature or sharing a task over the internet. One user took a screenshot of the information on the webpage, cropped it in Microsoft Paint, and emailed the image to the experimenter.

Another limitation is that we only used extremely novice users. Each subject was shown how the various features worked by the experimenter, but did not actually interact with the system except when actively working on the tasks assigned to them. Users may behave differently as they develop strategies over time.

All tasks in the experiment were externally driven, and under fairly severe time pressure. It is not clear whether users would be willing to accept the small extra overhead of organizing tasks for their own less structured web browsing needs. Again, this would require observing real world usage over a longer timeframe than was practical for a laboratory experiment.