Changing Paradigm – Changing Experience?

Comparative Usability Evaluation of *Windows 7* and *Windows 8*

Tim Schneidermeier, Franziska Hertlein, and Christian Wolff

Media Informatics Group, University of Regensburg,
Universitätsstraße 31, 93051 Regensburg, Germany
`{Tim.Schneidermeier,Christian.Wolff}@ur.de,`
`Franziska.Hertlein@stud.uni-regensburg.de`

Abstract. With the introduction of MS *Windows 8*, *Modern UI / Metro Design* was established as a new design paradigm for interaction. In this paper, we evaluate the usability of *Windows 8* in comparison with *Windows 7* with respect to effectiveness, efficiency and satisfaction. Our test was conducted on three sample rates with differing experience on MS *Windows* systems. The findings concerning each of the three dimensions are presented as well as results for the overall usability.

Keywords: comparative usability evaluation, flat design, *Windows 8*.

1 Introduction

In October 2012 Microsoft introduced its latest update for its operating system (OS), *Windows 8*. It was released with a major change in its interaction paradigm, differing widely from its predecessors. *Windows 8* is designed as a multi-platform OS, being operable on PCs as well as touch-based tablet devices. Microsoft`s approach is aiming at an seamless experience on different devices. Yet the main challenge for the success of *Windows 8* will be to add additional benefits – whether by functionality or usability in a workspace environment. In this paper, we present a usability study comparing *Windows 7* as a representative of the traditional WIMP (windows, icons, menu, pointers) interaction paradigm and *Windows 8* in terms of efficiency, effectiveness and user satisfaction. The study was conducted based on typical operating system tasks which were collected by asking users about typical tasks they have to complete on a regular basis. Both operating systems were tested by 24 participants, divided in three groups according to their experience with *Windows*. We also wanted to see if there is a connection between the prior experience with *Windows* and the way the participant interacts with *Windows 8*.

2 Object of Investigation: *Windows 7* and *Windows 8*

The *look and feel* as well as the interaction style of the new *Windows* user interface at the same time differ in many ways from its predecessors and yet are quite similar in

other respects: Instead of the familiar desktop surface known since the beginnings of *Windows* (*Windows 1.0* 1985), the newly introduced start screen in *Metro Design / Modern UI* (figure 1) gives an overview of selected apps. These are represented by tiles and vary from colored shapes with a logo to dynamic content showing a preview of the app data (photos, weather data etc.).

Fig. 1. *Windows 8*'s start screen in *Metro Design*

2.1   Flat Design vs. Skeuomorphism

The skeuomorphic design language in *Windows 7* (and older versions) is based on a realistic representation of real world objects, using shadows, light reflection, color gradients etc. as design tools (Grossman 2013). The objects in *Windows 8* (e.g. tiles, buttons etc.) on the contrary are designed with simple shapes, no color gradients and no kittenish details. The *flat design* should provide a clearly structured two-dimensional *look and feel* (see figure 2).

Fig. 2. Icons for "Video"; left image: skeuomorphism; right image[1]: *flat design*

---

[1] The *Windows 8*'s tile for "Video" is now styled according to (*ISO 18035:2003*, 2008).

2.2   Full Screen vs. *Windows*

Based on the WIMP paradigm starting a program or double-clicking a folder in *Windows 7* will open a new window showing the selected content. The users can manipulate these windows in different ways (e.g. resizing, dragging); to terminate a software the user can just close the window the application is running in. The interaction in *Windows 8* uses slightly different rules: Starting an app by clicking on a tile in the start screen will open the app in full screen mode, closing an app won't work by clicking in the right top corner, instead you have to click at the border above and drag it down (alternatively one can use the new *Windows 8* task manager). Working with multiple windows simultaneously in *Windows 7* allows you to customize the desktop to your needs: you can just drag the windows and position them according to your preferences. Operating in full screen mode in *Windows 8* is not that customizable. Yet there is still a possibility to show at least a limited number of apps at the same time: An app can be snapped at the right or left border, so that it won`t use the whole space. If another app is opened, it uses the remaining space.[2]

2.3   Desktop View

Even though *Windows 8* is introducing a new interaction style, the familiar desktop view can still be found "behind" the start screen. It can be opened by selecting the corresponding tile on the start screen. Like any other app in *Windows 8* it will run in full screen mode. The surface is almost the same as known from previous *Windows* releases, but is missing the start button to trigger the well known start menu[3] in the lower left corner, which is replaced by the newly introduced vertical toolbar *charms*. The *charms* bar can be activated throughout shortcuts or the OS hovering above one of the right corners (on touch-based devices it is reached with a swipe gesture from right to left). The *charms* menu provides five items to the user: *Search*, *Share*, *Start*, *Devices* and *Settings*. While some features provide the same functionality throughout the system, others are app-sensitive, i.e. the provided functions vary whether they are activated on the start screen or within an app. For example, clicking on the Start item within an application will lead to the *Windows 8* start screen, using the same element on the start screen the application last used will be displayed (Microsoft, 2014a).

3   Study Design

The goal of this study was to compare the usability for standard tasks using *Windows 7* and *Windows 8*. We have chosen a user-based over an expert-based

---

[2] The version of *Windows 8* tested in this study only provides the possibility to snap two apps in the proportion 1:3. The latest update *Windows 8.1* allows for displaying up to four apps at once (Microsoft, 2014a).

[3] In *Windows 8.1* a start button is re-introduced, which offers e.g. power options or a device manager (Microsoft, 2014b).

evaluation to get more realistic results. To measure the usability each participant had to complete six tasks on both systems. In addition to measuring the overall usability we were interested in whether prior experience in using *Windows 7* had any influence on using *Windows 8*. As the latter introduces a new interaction paradigm, we wanted to find out if there is a correlation between using *Windows 8* with and without any prior experience with older versions of the OS in terms of efficiency, effectiveness and user satisfaction. The test participants therefore were divided in groups with different levels concerning their expertise with *Windows*.

3.1   Measuring the Usability

Usability is measured by effectiveness, efficiency and user satisfaction (*ISO 9241*, 1998). Effectiveness was measured by whether a user was able to successfully complete a task (task success score / task completion rate, Sauro & Lewis 2012, p. 12), efficiency by the number of interaction steps (mouse clicks) and time per task. The task success score also allowed us to distinguish whether a task was completed without help or with an additional advice from the test supervisor. We used Morae[4] (TechSmith, 2013) to record the data during the test sessions.

To measure the individual user satisfaction we used the *AttrakDiff2,* a questionnaire developed to measure the user experience (Hassenzahl, Burmester, & Koller, 2008). It is based on 28 bipolar, seven-stage items. The antonyms of the *AttrakDiff2* questionnaire can be mapped onto four scales: pragmatic quality, hedonic quality (stimulation), hedonic quality (identity) and attractiveness. The tool provides a result visualization, where the rated products are positioned according to a pragmatic and hedonic axis. A product with a position in the top right corner can be seen as most pragmatic and hedonic and therefore offers a good user satisfaction. (User Interface Design GmbH, 2013)

For interpreting the results, we used a paired *t*-test. This test is suited for evaluating data of one control sample but two products. The paired *t*-test is a one-sample *t*-test executed with the difference of two measured values by person (Sauro & Lewis, 2012, p. 63ff). This approach should show whether the arithmetical means are significantly different. The null hypothesis is accordingly that no difference of the means exists. Although we had less than 30 participants, the paired *t*-test can nevertheless produce correct *p*-values (Sauro & Lewis 2012, p. 68). All tests are based on a significance level α=0.05.

3.2   Conducting the Study

We conducted the usability study with three groups of participants differing in their experience with the *Windows* OS (see table 1).

---

[4] Software to conduct, record and analyze data for usability studies.

Table 1. Test persons and their *Windows* experience

experience-level	number of users
up to *Windows 8*	10
up to *Windows 7*	10
no *Windows* experience	4

Even though we took a lot of effort, we weren't able to find a sufficient number of participants with no prior *Windows* experience. The ones who took part in the study were all Linux users with little or no expertise in *Windows*, which probably means that they had significant general ICT expertise. All participates were between 18 and 28 years old and students.

Table 2. Test tasks

Task number	Description
1	Checking emails using a browser
2	Text processing
3	Organizing files in folders
4	Changing default values
5	Installation of new software
6	Shutting down the device

To identify typical tasks for the study we conducted a survey asking about frequent tasks users have to accomplish on a regular basis. Based on the results we identified six tasks for our study design (see table 2). The user tests were conducted within one month at the usability lab of the *Media Informatics Group*, University of Regensburg (see figure 3). The participants were asked to accomplish all tasks on both OS, *Windows 7* and *Windows 8* using the same 13" *Samsung* ultrabook.

Fig. 3. Test environment

After the initial welcome, the test procedure was explained to each participant. Additionally, all participants were introduced to the *Thinking aloud* method and kindly asked to communicate their thoughts and feelings during the test session (Lewis 1982, Nielsen Norman Group, 2014). In order to prevent any influence on the test results based on sequence of the test, the order in which the OS had to be used was randomized. Each test session followed the same sequence: After editing the six tasks on OS A the participants were asked to fill out the *AttrakDiff2* questionnaire and continue with OS B.

4   Results

In the following, we present the main results for the three dimensions of our test, effectiveness (task success score), efficiency (mouse clicks / completion time) as well as overall user satisfaction (*AttrakDiff2*).

4.1   Effectiveness

Figure 4 shows the number of people needed help to complete a task. One can see at first glance that more help was needed while using *Windows 8*. In general, only in five cases help had to be provided to the participants completing the tasks 1, 2, 4 and 5 using *Windows 7*. Considerably more help had to be given when the participants were using *Windows 8*. Task 3 could only be completed by eight out of 24 people – five of them were actual *Windows 8* users. Changing the wallpaper (task 4) seemed to be very easy at all. For the other four tasks, between three and eight test participants needed help. The test conductor helped during the tests of *Windows 8* altogether 40 times.

It is obvious that the editing of task 3 in *Windows 8* is the most difficult task: The test subjects had to select five favourite pictures out of a folder and copy it in a new one, which they had to create and rename. The problems here may be due to the use of the *Windows 8's Photo app*. Most of the test participants used this application to have a look at the pictures. The behaviour of the application is similar to the *File Explorer* combined with the *Windows Photo Viewer*, which are known from *Windows 7*. That is why the functions of these applications, for example creating a new folder and copying pictures, were expected in the new *Windows 8*-application, but the application doesn't offer these. To do so, users have to change to the desktop. Several participants realized this only with clues given by the test supervisor. Before they could finish the task in the desktop environment, they had to close the *Photo app* or at least minimize it. This was because of the use of the full screen the second huge problem for solving task 3.

The help, which had to be provided for solving task 4, has similar reasons: Changing the wallpaper is a function offered amongst others in the *File Explorer*, so the test persons looked for it at this position. But to solve this task, they had to use the option offered in the desktop environment.

Eight people needed help during the editing of task 6 (shutting down the device). To find this option in *Windows 8*, you have to call the *charms* bar. The option "shut down" can be found in the category *Settings*.

Because we wanted to know whether there is a significant difference in the quantity of providing help, we executed a paired t-test. The null hypothesis was: While working on *Windows 8* help has to be provided as often as while working on *Windows 7*. The test results show that this null-thesis has to be refused (t=-5.7171, df=23, p=0.0000) and more help had been provided during testing *Windows 8*.

All tasks could be solved, no matter in which system they were edited. But it is clear that more *Windows 8* users needed help during the test in *Windows 8* than *Windows 7* users during testing *Windows 7*. Therefore, *Windows 8* can be judged to be less effective than *Windows 7* in that respect.
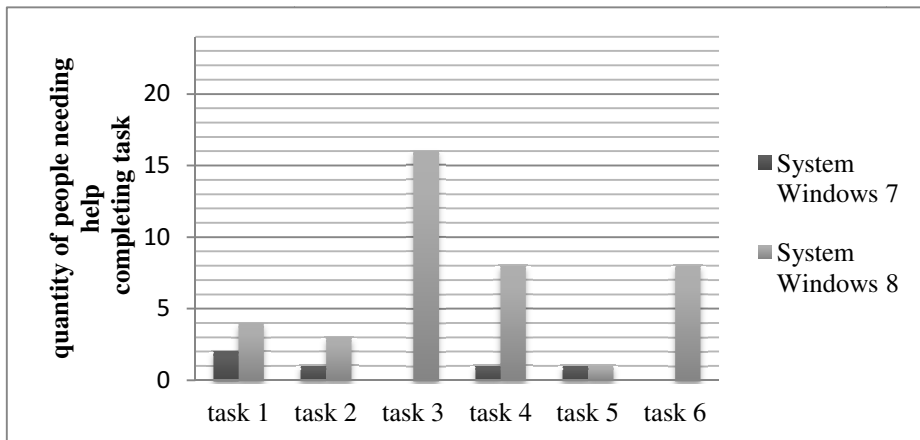

Fig. 4. Task success score for both systems

4.2   Efficiency

Summarizing the time which was needed for solving all six tasks, it can be seen that all groups testing *Windows 7* are very close to each other in comparison with the testing of *Windows 8*. The results of the average total time testing *Windows 7* are: *Windows 7* users 7:25 minutes, the group of *Windows 8* users 7:46 minutes and the persons with least *Windows* experience 8:34 minutes.

Solving the tasks in *Windows 8*, *Windows 7* users needed 14:35 minutes, the user of other systems 14:59 minutes and the active *Windows 8* users only 9:04. It is interesting that the time of *Windows 7* users and not *Windows*-users are very close together and that the *Windows 8* users have a clear advance. Thus we have to reason that prior experience in handling *Windows* systems is no immediate advantage for working with *Windows 8*. Task 5 required most time in both systems. During this task, software needed to be downloaded and installed. This caused a waiting time

which is part of the time displayed in fig. 5. An examination of the total time needed for all tasks with the paired t-test shows that the time the *Windows 8* users needed differs significantly from the time of the *Windows 7* users (t = 3.5906, df = 9, p = 0.005834).

The *Windows 8* users needed almost 6 minutes less than the rest for finishing the tasks in *Windows 8*. During the test in *Windows 7*, the *Windows 7* users couldn't reach a similar result. Therefore it is assumed that experience in handling *Windows 8* can raise the efficiency considerably.
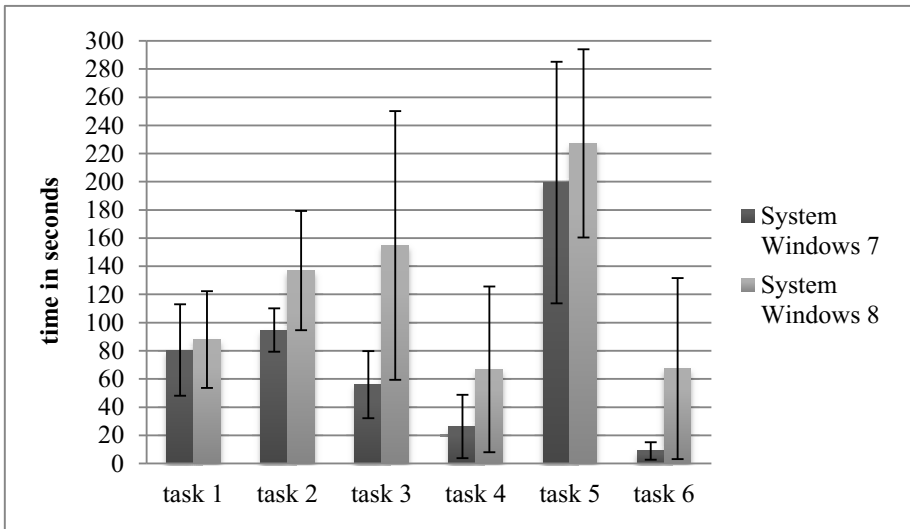
Fig. 5. Average time of all test participants

Figure 5 shows the average time of all test results. It is apparent that every task in *Windows 8* on average took more time than in *Windows 7*. Only one out of ten *Windows 8* users could beat his own *Windows 7* task time using *Windows 8* needing less time in every task. But the paired t-test shows that the null hypothesis, editing the tasks in *Windows 8* needs more time than in *Windows 7*, could not be rejected because of the inequality of the variance (t = -1.2855, df = 9, p = 0.2307).

As a second variable for measuring the efficiency, the sum of clicks was taken. Unfortunately, results logged by Morae were erroneous: In nine tests, zero clicks per task were counted, mainly during the tests in *Windows 8*. Fig. 6 gives the results for the remaining tests. These results support the findings of the time analysis: Task 3 and task 6 show the most explicit difference between both systems. And in *Windows 8*, every edited task on average needs more clicks to finish than in *Windows 7*.
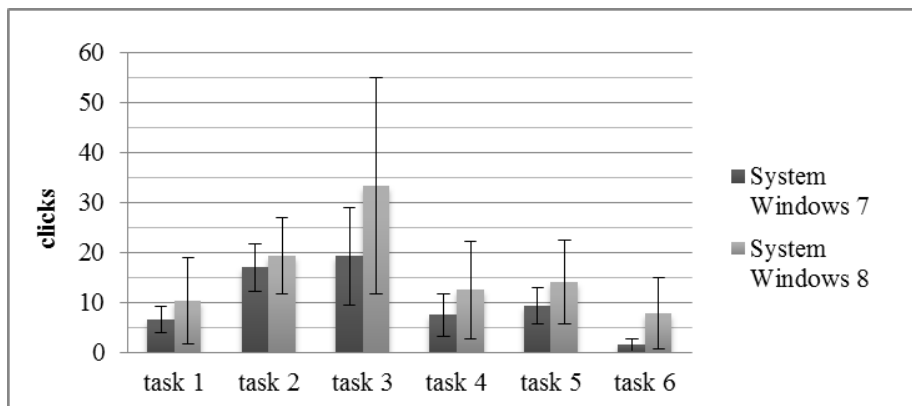
Fig. 6. Average mouse clicks of all test participants

4.3 User Satisfaction

The results of the AttrakDiff2 questionnaire (see figure 7) show that the users see clear difference in handling *Windows 7* and *Windows 8*. Especially the benchmarking of the pragmatic quality is very different. The automatically derived analysis of AttrakDiff2 states that *Windows 7* "assists its users optimally", while the "hedonic value is only average" and has "clearly room for improvement". In general, *Windows 7* can said to be "practice oriented".

Windows 8's user interface is rated as "neutral". It is more hedonic than *Windows 7* but in return less pragmatic. The *AttrakDiff2* results show that there is room for improvement in both dimensions, but definitely more in the pragmatic one. The confidence interval of *Windows 8* on the pragmatic axis has the broadest range. This could be explained with the participants' different experience in operating *Windows 8*: The groups didn't agree about the pragmatics of the system. The detected results in pragmatic and hedonic quality differ significantly. As *Windows 8* still is a quite novel product, less possible overall experience might be one explanation for this.
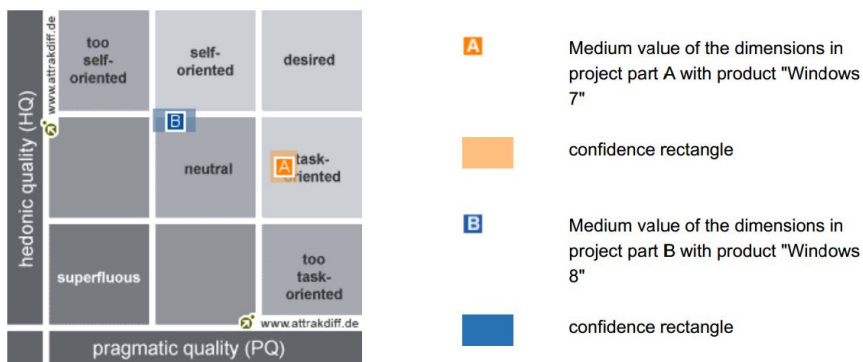

Fig. 7. Results of *AttrakDiff2*

To get a deeper understanding we had a closer look at some strongly rated adjective pairs in AttrakDiff2. Table 3 shows the results per group and per system. Items had to be rated on a scale ranging from 0 to 7, depending on whether they were closer to the left or the right item in the adjectives pairs. We converted the displayed result scales so that now seven is the best possible rating for every item. Especially the very negative review of *Windows 8* by the users without *Windows* experience is conspicuous. It seems as they were very displeased with the system and the results disagree strongly with the remaining results. There could be a strong bias not explained by missing experience, but possibly rooted in general convictions concerning specific products. We did not study this dimension, though.

The new direction in designing the user interface of *Windows 8* is mirrored by the adjectives "innovative" and "novel", but for not experienced *Windows* user also by "impractical" and "confusing". In general, the ratings of *Windows 8* are very different for the three groups. In comparison, *Windows 7* is rated as equally "pleasant" and "good" by both *Windows*-experienced groups. In the opinion of the last group, it is also "straight forwards" and "clearly structured" – according to them the clear opposite of *Windows 8*. In general, there was a stronger agreement among the three groups in rating *Windows 7* than in *Windows 8*. According to AttrakDiff2, both good results, but *Windows 7* is closer to an ideal user satisfaction.

Table 3. Highest rated items in the *AttrakDiff2* adjective pairs

user group	tested system	
	Windows 7	*Windows 8*
***Windows 7*-user**	practical (6.1)	innovative (5.8)
	pleasant (6)	challenging (5.5)
	good (6)	attractive (5.4)
***Windows 8*-user**	pleasant (6.2)	novel (5.9)
	manageable (6.1)	creative (5.9)
	good (6.1)	presentable (5.8)
No *Windows*-user	straight forwards (6.25)	confusing (6.25)
	clearly structured (6)	unpleasant (5.75)
	predictable (6)	impractical (5.5)

5   Conclusion

Summing up it is clear that the results for *Windows 7* are better than those for *Windows 8* for the three dimensions tested. Especially in the field of efficiency clear differences could be exposed. But in spite of the good results, one of the control samples has to be reconsidered critically: It was hard to find people completely lacking *Windows* experience, but probably the differentiating criterion is too weak: The paradigm used in *Windows 7* and further versions of *Windows* is WIMP which is the current state of the art for graphical user interfaces on PCs and can e.g. also be found in Linux Ubuntu or the Mac OS series.

Even if the persons in the control sample never have used or even seen a *Windows* product before the test (which is unlikely given the strong market position of the *Windows* family of products), they also should not have used any other operation system based on the WIMP paradigm. If they had, they know how to handle windows, interact with icons and menus by clicking with a mouse from a similar system to *Windows 7*. So actually the control sample should be labelled "users with no WIMP-experience", but to find people to join them was not in the realm of possibility for this study – more than 30 years after the market introduction of GUI systems this hardly comes as a surprise. Hence *Windows 7* has an advantage of familiarity during the test, because of its widely spread paradigm.

Before the testing phase, we had suspected that *Windows 8*'s split interface would disturb novel users, but this could not be confirmed. Changing between desktop and tile environment was confusing at first, because users didn't expect to find the desktop after the surprising new start screen. But after a short familiarisation phase it was accepted and not longer recognized as a problem. For example, during task 5, installation of software (*Adobe Reader XI*, not as app version) *Windows 8* changes automatically from the download of the installation file in the *Internet Explorer* app to the desktop where the installation dialog appears. Most of the users did not notice the automatically change of the environment. Working with two interaction paradigms in one system does not appear as a problem according to our test.

The biggest criticism of *Windows 8* was the missing start button, a fact well reflected in popular online discussions of the problem.[5] Many users claimed at the beginning of a task that they would now call the start menu over the start button to start the handling of the task. Because they could not use this function any longer, they had to find different ways to start. Many of the possibilities offered were also available in *Windows 7*, but test subjects were used to the start button, hence they didn't have alternatives on their mind. With the disappearance of the start button in *Windows 8*, Microsoft has removed an important initial point for operating the system. There was a lot of public criticism on Microsoft for this decision, so Microsoft decided to put a new version of the button back in the desktop in *Windows 8.1*.

Because of the suitability of *Windows 8* for touch devices, apps in full screen and the *charms* bar are an important feature. Both caused many problems during the test: As described above, changing or closing apps running in full screen mode was difficult for many users. They did not see any obvious possibilities and spent much time on searching a way out. The hint of changing the cursor appearance when the user hovers over the upper border was not enough to provoke a drag gesture from the top to a middle position. Similar problems with hidden functions occurred with the *charms*. Novel users did not know how to let the hidden bar appear and stay fixed. It fist seemed kind of random how the users tried to activate the *charms* bar with their mouse movements. But after a while, this feature was apprehended and could be used in a controlled way by the test subjects.

---

[5] The quite specific query "*Windows 8*" "missing start button" yields 200,000 hits in *Google*.

After the analysis of the control samples' data, the hypothesis that *Windows 8* is more usable than *Windows 7* could not be confirmed. The differences between the results of each part of the usability analysis are quite remarkable. Because *Windows 7* is a representative of the WIMP paradigm and *Windows 8* is built on its own new interaction paradigm, WIMP has to be seen as more usable than the paradigm introduced in *Windows 8*. It is open, though, whether this finding might change with longer experience phases for the new interaction style.

References

1. Hassenzahl, M., Burmester, M., Koller, F.: Der User Experience (UX) auf der Spur: Zum Einsatz von www.attrakdiff.de. In: Brau, H., Diefenbach, S., Hassenzahl, M., Koller, F., Peissner, M., Röse, K. (eds.) Usability Professionals, pp. 78–82. German Chapter der Usability Professionals Association, Stuttgart (2008)
2. ISO. Information technology – Icon symbols and functions for controlling multimedia software applications (ISO 18035:2003). International Standards Organization, Geneva (2003)
3. ISO. Ergonomic requirements for office work with visual display terminals (VDTs), Part 11, Guidance on usability (ISO 9241-11:1998E). International Standards Organization, Geneva (1998)
4. Microsoft. Search, share, print, and more (2014a),
   `http://windows.microsoft.com/en-gb/windows-8/charms-tutorial`
   (retrieved January 19, 2014)
5. Microsoft. Start screen (2014b), `http://windows.microsoft.com/en-gb/windows-8/start-screen-tutorial` (retrieved January 19, 2014)
6. Nielsen Norman Group. Thinking Aloud: The #1 Usability Tool (2014),
   `http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/` (retrieved January 22, 2014)
7. Sauro, J., Lewis, J.R.: Quantifying the User Experience. Elsevier/Morgan Kaufmann, Amsterdam (2012)
8. TechSmith. Morae (2013), `http://www.techsmith.com/morae.html` (retrieved October 09, 2013)
9. User Interface Design GmbH. AttrakDiff (2013),
   `http://attrakdiff.de/sience-en.html` (retrieved January 20, 2014)