# Impact of Screen Size on Performance, Awareness, and User Satisfaction With Adaptive Graphical User Interfaces

**Leah Findlater and Joanna McGrenere**
Department of Computer Science
University of British Columbia, Vancouver, Canada
{lkf, joanna}@cs.ubc.ca

**ABSTRACT**
Adaptive personalization, where the system adapts the interface to a user's needs, has the potential for significant performance benefits on small screen devices. However, research on adaptive interfaces has almost exclusively focused on desktop displays. To explore how well previous findings generalize to small screen devices, we conducted a study with 36 subjects to compare adaptive interfaces for small and desktop-sized screens. Results show that high accuracy adaptive menus have an even larger positive impact on performance and satisfaction when screen real estate is constrained. The drawback of the high accuracy menus, however, is that they reduce the user's awareness of the full set of items in the interface, potentially making it more difficult for users to learn about new features.

**Author Keywords**
Adaptive interfaces, personalization, small screen devices, menu design, user study, interaction techniques.

**ACM Classification Keywords**
H.5.2 [User Interfaces]: Evaluation/methodology, interaction styles.

**INTRODUCTION**
With the proliferation of mobile phones and PDAs, small screen devices are now pervasive, but smaller screens can make even basic tasks such as reading and web browsing more difficult [9,19]. The reduced screen size means that, even with high resolution screens, designers must choose only the most important features to display. Additionally, users tend to use mobile devices in contexts where their attention is limited in comparison to traditional environments [24], which may make it more difficult to navigate a complex interface. To address the limitations of small screen devices, several researchers have proposed that adaptive interfaces, where the system tailors the interface to an individual user's needs, may be beneficial [2,19].


*CHI 2008,* April 5–10, 2008, Florence, Italy.


Despite the potential theoretical benefits, research on adaptation for small screens has focused largely on adaptive web content (e.g., [19,28]) rather than on adaptive graphical user interface (GUI) control structures. GUI control structures, such as menus, present unique challenges in comparison to adaptation of content, for example, a higher user expectation for consistency [5]. In the context of mobile devices, there has been a small amount of work on adaptive menu structures for phones [3,25], but evaluations have been informal. The bulk of adaptive GUI research, rather, has been conducted on desktop-sized displays, where evaluations have been inconclusive: in some cases, adaptive menus or toolbars have been faster and preferred to their static counterparts [13,15], whereas other research has shown the opposite [10,22,23]. As a result, adaptive GUIs have been conceptually controversial and very few have appeared in commercial applications. If the benefit of adaptivity is more evident for small screens than large screens, adaptivity may be less controversial in this context and should be reconsidered as a viable design alternative.

The main goal of the work reported here was to investigate the impact of an adaptive GUI on small screen displays relative to desktop-sized displays. The results should shed light on the degree to which prior findings directly apply to smaller displays: for instance, an adaptive algorithm that was less efficient than a static counterpart may no longer be so when the two are used on a smaller screen. We also sought to extend prior work [13,29] by assessing the potential interaction between adaptive accuracy and screen size. We conducted an experiment with 36 users, comparing adaptive split menus [26] on a desktop screen to a PDA-sized screen. Since adaptive accuracy can affect performance and use of adaptive predictions [13,29], we included two levels of accuracy (50% and 78%) and a static control condition. Further, we specifically accounted for the predictability and consistency within our two accuracy levels, something that has not been done before.

Our study shows that high accuracy adaptive menus have a larger positive impact on user performance and satisfaction in small screens compared to large screens. This suggests that the potential of adaptive interfaces may be best realized in situations where screen real estate is constrained. We had thought this performance and satisfaction differential would be due to reduced navigation (i.e., scrolling) in small screens, but, interestingly, screen size also impacts user

behavior: people are more likely to take advantage of the adaptive predictions in a small screen. As expected, a low accuracy interface performs poorly regardless of screen size, which reinforces that research findings on adaptivity must be understood in the context of accuracy levels.

A secondary goal of our work was to measure the impact of screen size and adaptive accuracy on awareness. Recently introduced [11], *awareness* quantifies the degree to which the user is aware of the full feature set of an application, and provides insight into the potential performance tradeoff of working in a personalized interface. For example, an adaptive menu may focus the user's attention on a small set of frequently used features, with the drawback that the user may not see and thus learn about additional features. Our study shows that despite the performance benefits of a high accuracy adaptive interface, it can result in reduced awareness. It also suggests that awareness is impacted more negatively in small screens than in large screens, an important tradeoff that designers will need to consider.

The primary contribution of this paper is empirical evidence demonstrating the relative benefits of adaptive GUIs for small displays in comparison to large displays. A secondary contribution is to show that this benefit is not purely due to a reduction in the amount of navigation needed to access features, but that screen size also impacts user behaviour. Finally, the measurement of awareness provides a richer understanding of the impact of working in adaptive interfaces. Combined, our findings motivate the need to revisit previous adaptive research in the context of small screen devices, especially for those studies with negative outcomes for adaptive approaches.

## RELATED WORK

A large body of work exists on usability of non-adaptive small screen interfaces, generally showing that tasks are more difficult on small screens. For example, in comparison to a large screen, reading text requires more navigation [9], and searching the Internet is slower [19].

### Adaptive Interfaces for Small Screens

Approaches to adaptation can be broadly grouped into two categories: content and GUI control structures; our focus is on the latter. Research in this area has largely been done on desktop displays. One exception is SUPPLE, which automatically adapts interfaces based on device constraints and usage, but evaluations have been small and informal [12]. In other work, Bridle and McCreath compared a static mobile phone menu structure to six approaches that adaptively predicted a single shortcut item [3]. Simulation on logged user data suggested that some of the adaptive approaches would be more efficient than the static one, but no formal user evaluation was reported. Bridle and McCreath stress that stability should be considered in adaptive interface evaluations, which we did for the adaptive menus in our study (note: we call this *consistency*).

Adaptation of content has been applied more widely to small screens. For example, Smyth et al. have used adaptive hypermedia to personalize web portals for mobile devices, showing that personalization can reduce navigation to access content [27]; follow-up large-scale deployment showed that the approach increased customer satisfaction [28]. Adaptation of content, however, may present different challenges than adaptation of control structures [5]. Users may not expect the same degree of consistency from content as from control structures, and, compounding this, consistency can impact motor memory, one aspect of performance with control structures.

### General Evaluations of Adaptive Interfaces

Several studies have compared adaptive control structures to static and/or adaptable (user-controlled) counterparts with mixed results. Early research by Greenberg and Witten showed that an adaptive menu structure, which provided a shorter search path to the most frequent items, was faster than a static structure [15]. Conversely, Mitchell and Shneiderman compared static to adaptive menus that reordered during usage based on frequency, finding the static menus were faster and preferred [23].

Since being introduced in the form of split menus by Sears and Shneiderman [26], split interfaces have received a relatively large amount of research attention. An adaptive split interface separates adaptive and static sections of the interface. The original work showed that predetermined split menus, where items were *moved* to the adaptive top section of the menu, were at least as fast, or faster, than traditional static menus [26]. More recently, Gajos et al. showed that users had a strong preference for an adaptive split interface, which *replicated* items in the adaptive section, in comparison to a static counterpart [13]. Findlater and McGrenere have shown that adaptive split menus (where items were moved above the split) were slower than both static and adaptable split menus, except, in the case of the latter, when adaptable appeared first in order of experimental presentation [10]. We chose adaptive split menus for our study because they have been widely studied in the literature, and they appear in commercial applications, such as recency-based font selection menus.

### Exploring Adaptive Characteristics

Commonly cited issues with adaptive interfaces include lack of control, predictability, transparency, privacy, and trust [18]. Recently, researchers have begun to explore how these and other qualities may impact the success of an adaptive GUI. For example, several researchers have evaluated personalization approaches that use different degrees of user control [4,8,22]. More directly related to our study, Tsandilas and schraefel compared two approaches for adaptive highlighting of item lists and varied the level of prediction accuracy (100%, 80%, and 60%), finding that the lower accuracy conditions were slower [28]. Results also showed that lower accuracy increased errors for one of the adaptive approaches (*Shrink*, a fisheye-type distortion), which suggests that the effectiveness of adaptive designs may interact with accuracy.

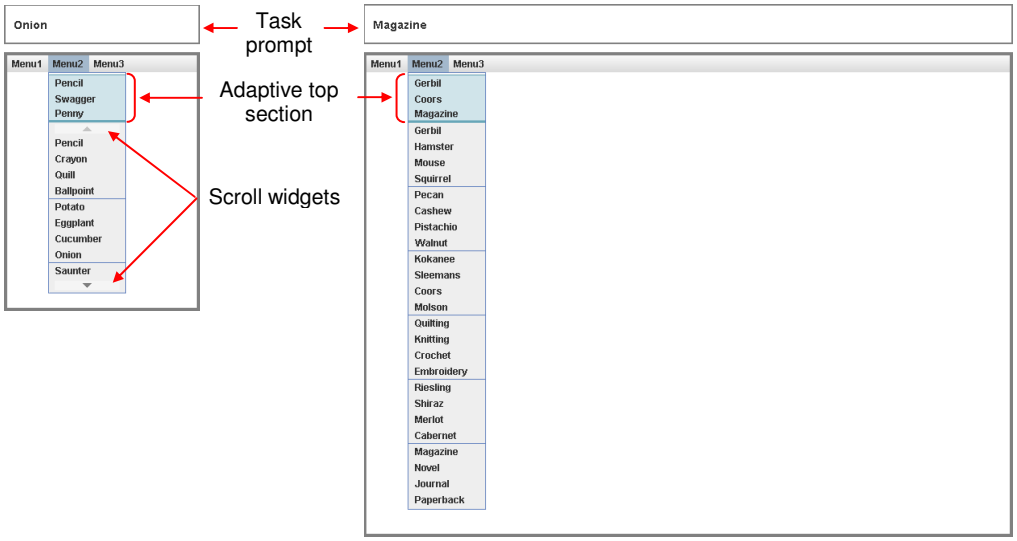


**Figure 1. Screenshots of Small screen (left) and Large screen (right) experimental setups with adaptive menus open, showing task prompt, adaptive top section, and scroll widgets for the Small screen. The High and Low adaptive conditions looked identical; the Static menus did not have an adaptive top section.**

Gajos, Czerwinski, Tan and Weld compared two adaptive toolbars to a static counterpart within two levels of adaptive accuracy [13]. The two adaptive toolbars were implemented as split interfaces, either moving adaptively suggested items to the adaptive section of the toolbar, or replicating the items there. The accuracy levels were achieved by creating two different tasks for which the algorithms were either 30% or 70% accurate. Results of a controlled experiment showed that the split interface that replicated items was significantly faster than the static toolbar. Both of the adaptive interfaces were faster with the higher accuracy condition, and participants took advantage of the adaptive suggestions more often in that condition.

Cockburn, Gutwin and Greenberg developed a model of menu performance that applies to both static and adaptive designs [6]. The model incorporates Hick-Hyman Law and Fitts' Law, and takes stability of an adaptive menu design into account. Results showed the model could accurately predict performance for four types of menus: frequency-based and recency-based split menus, traditional menus, and morphing menus (where items are resized according to frequency). While the model has potential for comparing designs theoretically, it cannot be applied directly to our study because it does not account for scrolling menus, which are used in the small screen condition.

## EXPERIMENT METHODOLOGY

To compare the impact of adaptive menus on a small screen versus a desktop-sized display, we conducted a controlled lab study with 36 subjects. An obvious drawback of designing for a small screen is that not all items can be shown at once, which results in an added navigation cost for accessing the items that are not immediately available. Our hypothesis was that, by reducing this cost, adaptive interfaces should be relatively more beneficial for a small screen than a large screen. Even so, given that previous results for adaptive GUIs on large screens have been mixed, it was not clear how an adaptive interface for a small screen would compare to a static one. We compared two adaptive menus (with 50% and 78% accuracy) and a static menu; the main task was to select a series of menu items. Support for our hypotheses would underscore the need for designers to revisit adaptive approaches in the context of small screens, where they may be more useful.

### Conditions

Figure 1 shows the layout of the experimental conditions.

#### *Screen Size*

To simulate two distinct screen sizes, the window containing the experimental application was either 800x600 pixels (Large screen) or 240x320 pixels (Small screen). For the Large screen, this was big enough to display a full-length menu for our experimental task. The Small screen condition, which was the size of many Pocket PC PDAs, was only big enough to display a subset of menu items at once. To access all items in the Small screen, the user had to hover or click on scroll widgets that appeared at the top and bottom of the menu (similar to menus in Windows Mobile 6). Based on pilot testing with 4 subjects, scroll speed was set at 1 item per 75 ms. This was reported to be the best compromise between efficiency and ease of reading item labels; with faster scrolling speeds pilot subjects often overshot their target and would have to recover by scrolling back. We controlled for input device and display characteristics by using a mouse for both conditions, and simulating the screen sizes on an 18" LCD flat panel monitor with 1280x1024 resolution.

#### *Menu Type*

We included a Static control condition, and High and Low accuracy adaptive conditions. The menus in the Static condition were traditional pull-down menus, while the High and Low adaptive conditions were adaptive split menus. With the split menus, adaptively chosen items were replicated, rather than moved, above the split (as preferred

by users in [12]); this necessarily made the split menus slightly longer than the Static menus. The bottom section was identical to the Static menus, while the top section contained three items (as suggested by [6,26]).

For each condition, the menu bar contained three individual menus, each with 24 items. The 24 items were further separated into semantically related groups of 4 items. (The length and group size were based on averages from four desktop applications, Firefox 2.0, Microsoft Excel 2003, Adobe Reader 7.0, and Eclipse 3.2, including both top-level and cascading submenus.)

*Adaptive algorithm detail.* To achieve two levels of accuracy, Tsandilas and schraefel changed the set of adaptive predictions for each trial, either including the item to be selected or not [29]. As acknowledged by the authors, this approach would result in a high level of unpredictability. Gajos et al. took another approach, using two different experimental tasks that resulted in different levels of accuracy for the same interface [13], which makes it difficult to directly compare performance. To address these limitations, we used an identical underlying set of selections for each condition, and determined the adaptive predictions in advance using a two-step process:

1. *Apply base algorithm.* Using a simple base algorithm (shown in Figure 2), we pre-calculated the items to appear in the adaptive top. This algorithm incorporated both recently and frequently used items, as suggested by the literature [10,14] and is commonly used in commercial adaptive user interfaces such as Microsoft Office 2003's adaptive menus. For the randomly generated selection streams in our study (described later), this resulted in 64.2% accuracy on average (*SD* = 1.7).
2. *Adjust accuracy.* To adjust accuracy, we then randomly selected 14% of trials (18 per block, as discussed later) that could be manipulated to increase accuracy (i.e., by swapping the item to be selected into the adaptive top) and 14% that could be manipulated to decrease accuracy (i.e., by swapping the item to be selected out of the adaptive top). This resulted in 50% and 78% accurate adaptive conditions, two somewhat arbitrarily chosen levels of accuracy, as we cover in the Discussion. We also enforced several constraints on this manipulation in an effort to maintain consistency and predictability (e.g., the most recently selected item always had to appear in the adaptive top).

1. set top section to the *most recently* selected item and the *two most frequently* selected items (as pre-calculated from the selection stream)
2. if there is overlap among these three slots or if this is the first selection in the stream (i.e., no recently selected item exists)

   then the third most frequently selected item is included so that 3 unique items appear in the top
3. order top items in the same relative order as they appear in the bottom section of the menu

**Figure 2. Base adaptive algorithm**

*Consistency and predictability of the menus.* We chose the above approach because we wanted the adaptive interfaces to behave as similarly as possible in aspects other than accuracy. We considered: (1) *consistency*, which we defined as the percentage of total trials where no items changed in the adaptive top (similar to [3]), and (2) *predictability*, which we defined as the percentage of trials where the adaptive top contained the item to be selected, *and* this could be predicted because that item had been in the adaptive top for the previous trial as well. The accuracy, predictability, and consistency of the Low and High conditions is summarized in Table 1. Note that the Low condition had both lower accuracy and lower consistency than the High condition. While it would have been ideal to achieve the same level of consistency for both High and Low, this compromise at least paired high consistency with high accuracy, and vice versa. The relative importance of these factors is covered in the Discussion section.

### Task

The main experimental task was a sequence of menu selections. As shown in the task prompt in Figure 1, the system displayed the name of a menu item for each trial but did not specify which menu should be used. Only once a subject had correctly selected the item, the next one would be displayed. To mitigate the impact of any particular set of selections (i.e., item locations), a new set was randomly generated for each subject. However, this underlying set of selections was used for all of an individual subject's conditions, and different menu *masks* (or item labels) were applied in each condition to reduce learning effects, similar to previous work [10,29]. For example, if *item 3* on *menu 1* was selected first, this was the case for each condition. The menu masks for each subject were created by randomly assigning 54 semantically related groups of 4 item labels, such that each group appeared once and only once per subject (as in [6]). For example, "diamond, topaz, emerald, sapphire," represented the precious stones group. All menu item labels were single words, 5-10 letters long.

Previous work has shown both that users only use a small subset of items (for Microsoft Word: 8.7% [20] to 21.5% [21] of items), and that usage can often be modeled by a Zipf distribution [16,17]. Following the approach of Cockburn, Gutwin and Greenberg [6], we simulated this type of selection pattern: we generated a Zipf distribution (Zipfian $R^2$=.99) across only 8 randomly chosen items out of the 24 items in a menu (with respective frequencies of: 15, 8, 5, 4, 3, 3, 2, 2). The final selection stream was also randomized, for a total of 126 trials per task block (42 trials in each of 3 menus). Each subject completed the same task block twice per condition.

| | Accuracy | | Predictability | | Consistency | |
|---|---|---|---|---|---|---|
| | *M* (%) | *SD* (%) | *M* (%) | *SD* (%) | *M* (%) | *SD* (%) |
| **Low** | 50.0 | 1.7 | 94.1 | 2.3 | 19.7 | 2.6 |
| **High** | 78.5 | 1.7 | 94.4 | 2.0 | 36.5 | 3.9 |

**Table 1. Accuracy, predictability, and consistency of adaptive conditions. Since task selection streams were randomly generated, values were not identical for each subject (*N* = 36).**

### Quantitative and Qualitative Measures

*Performance.* Speed was measured as time to complete both task blocks per condition. Error rate was also recorded, although there was an implicit penalty for errors since subjects had to correctly complete a trial before advancing.

*Awareness.* Awareness is a measure of the secondary, incidental learning that may occur as the user performs a primary task [11]. Subjects were given an awareness-recognition test, similar to that used by Findlater and McGrenere [11], for each menu condition. This test listed 12 randomly chosen items that were found in the menus for each condition, but were *not* selected in the tasks. It also included 6 items randomly chosen from a set of distractor items; the full distractor set contained 1 item for each group of 4 items used in the menus, such that the item was related to that group (e.g., distractor for the group "soccer, basketball, baseball, football" was "rugby"). Valid and distractor items were chosen evenly across menus.

For each item, subjects were asked to note if they definitely remembered it. From this, we calculate awareness as the corrected recognition rate of the recognition test score. This is a commonly applied method in psychology to account for individual variation in the amount of caution a subject applies when responding to a memory test; it is simply the percentage of valid targets correctly remembered minus the percentage of distractors incorrectly chosen [1].

*Subjective measures.* Finally, after each menu condition subjects were asked to rank the condition along several 7-point Likert scales: difficulty, efficiency and satisfaction. Additionally, consistency and predictability were also asked for the two adaptive conditions. Lastly, we asked subjects for their overall preference of the three menu conditions.

### Design

A 2-factor mixed design was used: screen size (Small or Large) was a between-subjects factor, while menu type (High, Low or Static) was within-subjects. Presentation order of menu type was fully counterbalanced.

### Subjects

Thirty-six subjects (19 females) between the ages of 19-49 were randomly assigned to either the Small or Large screen condition and to a presentation order for menu type. Subjects were recruited through campus advertising and were screened so that they were not novice computer users (i.e., used a computer for at least 3-5 hours per week). Each subject was paid $15 to participate.

### Apparatus

The experiment used a 2.0 GHz Pentium M laptop with 1.5 GB of RAM, running Microsoft Windows XP, and connected to an 18" LCD monitor at 1280x1024 resolution. The application was coded in Java 1.5. Figure 1 shows a screenshot of the application: instructions were given one at a time at the top of the screen. The system recorded all timing and error data.

### Procedure

The experiment was designed to fit in a 1.5 hour session. Subjects were first given a background questionnaire. Then, to introduce the format of an awareness-recognition test, subjects completed a 5-minute paper-based search task on a list of words, followed by an awareness test of words that appeared on the list but were not included in the task. This was so subjects would be prepared for a similar test after each menu condition.

Following this, the three menu conditions were presented, with 5-minute breaks and paper-based distractor tasks between each. For each condition, the subject completed a short practice block of 15 selections, followed by the same task block repeated twice. To reduce fatigue, 30-second breaks in the middle of each task block and a 1-minute break between blocks were enforced. After the second task block the awareness recognition test was administered. At the end of all three conditions, a preference questionnaire asked for comparative ratings of the three menu types.

Subjects were not told about the different accuracy levels for the conditions. For the first adaptive condition they were simply told that the items in the top section of the menu would change as they performed the task, and for the second adaptive condition that the behaviour of the top section was slightly different from the previous condition.

### Hypotheses

We summarize our main hypotheses:

*H1. Higher adaptive accuracy is faster than lower.* The difference between High and Low would replicate previous findings [13,29]. Previous results of comparing adaptive menus to static ones have been conflicting [10,15,23,26], so it was unclear how the static menu would fare.

*H2. Small screen is slower than Large screen.* Previous research has shown that tasks such as text reading and content retrieval are slower on small screens (e.g., [9,19]), so this should be the case for accessing menu items, especially considering the additional scrolling needed.

*H3. Effect of adaptive accuracy on speed is greater in Small screen than Large screen.* The relative benefit of the adaptive interfaces should be higher for the small screen, largely because they will reduce the amount of scrolling.

*H4. Higher adaptive accuracy results in lower awareness, and Static has the highest awareness.* The higher the adaptive accuracy, the fewer menu items that users will need to navigate through to complete their task blocks. Thus, higher accuracy should result in reduced awareness.

*H5. Small screen results in lower awareness than Large screen.* Since at least half of the menu items are hidden from view at any given time with the Small screen condition, it should result in lower awareness than the Large screen condition.

*H6. Effect of adaptive accuracy on awareness is greater in Small screen than in Large screen.* Combining the

arguments from H4 and H5, we would expect the differences in awareness due to accuracy to be even more pronounced in the Small condition.

## RESULTS

A 2x3x2x6 (screen size x menu type x task block x presentation order) repeated measures (RM) ANOVA showed no significant main or interaction effects of presentation order on the main dependent variable of speed, and showed a main, learning effect of block. Since both of these were expected, we simplify our results by examining only effects of screen size and menu type, collapsing across block. All pairwise comparisons were protected against Type I error using a Bonferroni adjustment. Where *df* is not an integer, this is because we have applied a Greenhouse-Geisser adjustment for non-spherical data. We report measures which were significant ($p < .05$) or represent a possible trend ($p < .10$). Along with statistical significance, we report partial eta-squared ($\eta^2$), a measure of effect size. To interpret this value, .01 is a small effect size, .06 is medium, and .14 is large [7].

Two subjects (1 Large screen and 1 Small screen) were removed from the analyses for each having at least one performance measure more than 3 standard deviations away from the mean. Thus, we report on the data of 34 subjects.

### Speed

We present the speed results first, followed by secondary analysis to understand some of the specific behaviours that may have contributed to the differences in speed (i.e., scrolling and use of adaptive predictions).

#### *Primary Speed Results*

On average, subjects took 877 seconds to complete both selection blocks in each condition (*SD* = 189). The results are summarized in Figure 3. A 2x3 RM ANOVA for speed (screen size x menu type) showed that the combination of menu type and screen had a significant impact on speed (i.e., an interaction effect $F(2,64) = 9.201$, $p < .001$, $\eta^2 = .456$). To understand the reason for this we conducted pairwise comparisons, as shown in Table 2.

*High accuracy menus are faster than Low ones, but outperform Static menus only in Small screens.* As predicted by H1, High was faster than Low in both screen conditions, showing that a higher accuracy interface is more efficient independent of screen size. Support for H3 is also shown in the pairwise comparisons by looking at the relative performance of Static to the adaptive menus for each screen size. High was no different than Static in the Large condition, whereas it was significantly faster than Static in the Small condition. The Low accuracy menu did not perform better than Static in either condition; in fact, it performed worse than Static in the Large screen. Thus, from a performance standpoint, our results show that there was a benefit to adaptive menus, relative to status quo Static menus, only when they have High accuracy and only in Small screens. Low accuracy is at best no worse than Static (for Small screens) and at worst, it degrades performance relative to Static (for Large screens).

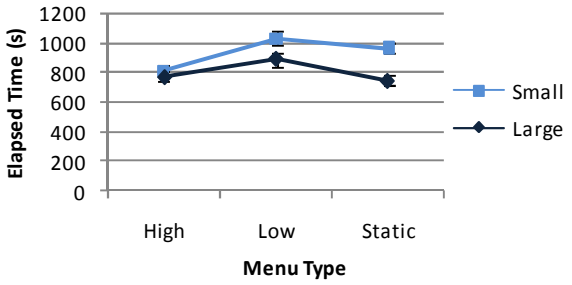


**Figure 3. Speed (*N* = 34).**

*Small screen slower than Large screen.* As predicted by H2, subjects were significantly slower using the Small screen, taking 938 seconds on average to complete both task blocks in that condition, compared with 821 seconds in the Large condition (a main effect of screen, $F(1,32) = 20.923$, $p < .001$, $\eta^2 = .395$).

#### *Secondary Analyses: Scrolling and Adaptive Predictions*

*High accuracy reduces scrolling.* One of the expected benefits of the adaptive menus in the Small screen was that they would reduce the amount of scrolling (there was no scrolling in the Large condition). We counted scrolling as the number of items scrolled upward or downward. The mean items scrolled in High, Low, and Static were 1019, 1750, and 1867 respectively. High indeed resulted in significantly less scrolling than the other two menus, which mirrors the speed results. (A single factor (menu type) RM ANOVA on the Small screen data showed a main effect of menu type on scrolling, $F(2,32) = 31.715$, $p < .001$, $\eta^2 = .665$, and $p < .001$ for both the High-Low and High-Static comparisons.)

*Small screen increases use of adaptive predictions.* Previous work has suggested that lower accuracy adaptive interfaces will result in lower user trust in the adaptive predictions [29], and that users will be less likely to make use of those predictions [13]. To explore this behaviour for the two adaptive menu conditions, we ran a 2x2 (menu type x screen size) RM ANOVA on the percentage of trials where subjects did not use the top section of the menu to make a selection that had been correctly predicted by the adaptive menu. We call these *non-strategic* selections.

| Menu (*i*) | Menu (*j*) | Mean Difference (*i*-*j*) | Std. Error | Sig.[a] |
|---|---|---|---|---|
| *Small screen* | | | | |
| High | Low | -213.394* | 31.178 | <.001 |
| High | Static | -152.359* | 30.561 | <.001 |
| Static | Low | -61.035 | 35.339 | .281 |
| *Large screen* | | | | |
| High | Low | -119.856* | 31.178 | .002 |
| High | Static | 24.387 | 30.561 | 1.000 |
| Static | Low | -144.243* | 35.339 | .001 |

a. Adjusted for multiple comparisons using Bonferroni.
* The mean difference is significant at the .05 level.

**Table 2. Pairwise comparisons for speed, in seconds (*N*=34).**

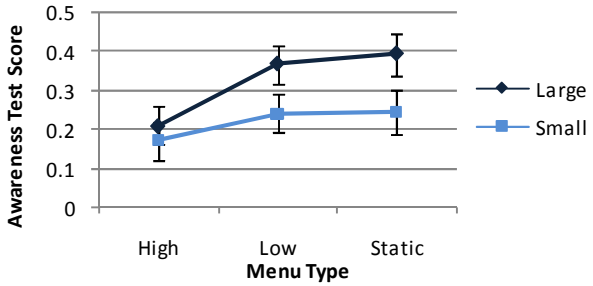


**Figure 4. Awareness recognition test scores ($N$ = 34).**

Subjects in the Large screen condition made significantly more non-strategic selections than subjects in the Small screen condition, 22.7% vs. 9.7% (main effect of screen size, $F(1,32) = 5.706$, $p = .023$, $\eta^2 = .151$). This result suggests that subjects perceived the adaptive predictions to be more useful in the Small screen condition, which may at least partially explain why the High accuracy menus were faster than Static menus for Small screens but no different for Large screens. Also, as expected, subjects made significantly more non-strategic selections in Low (18.9%) than in High (11.4%) (a main effect of menu, $F(1,32) = 7.657$, $p = .009$, $\eta^2 = .193$).

### Awareness

After efficiency, we were most interested in how the menu conditions and screen sizes would impact the user's overall awareness of menu items. Figure 4 shows the overall corrected awareness test scores. A 2x3 (screen size x menu type) RM ANOVA showed that the menu type did significantly impact users' awareness (main effect of menu type on awareness, $F(2,64) = 6.547$, $p = .003$, $\eta^2 = .170$).

*High accuracy results in the lowest awareness*. We found partial support for H4. As expected, High had the lowest awareness, with an average score of 19% on the awareness test, in comparison to both Low (30%) and Static (31%) (pairwise comparisons were $p = .006$ and $p = .009$, respectively). However, there was no significant difference found between Low and Static.

*Small screens seem to impact awareness more negatively than Large screens*. We found trend level support for H5. The Large screen subjects scored on average 31% on the awareness test, while the Small subjects scored only 22% on average, a difference that was marginally significant (main effect of screen on awareness, $F(2,32) = 3.392$, $p = .075$, $\eta^2 = .096$). However, we did not find any support for H6; the different accuracy levels did not have a greater impact on awareness in the Small screen condition relative to the Large condition (there was no significant interaction effect between screen size and menu type, $F(2,64) = 1.134$, $p = .328$, $\eta^2 = .034$).

*High accuracy fastest for selecting frequent items, but slower than Static for infrequent items.* As a possible indirect effect of awareness on performance, we wanted to know if subjects had more difficulty selecting infrequently accessed items in conditions with lower awareness. To do this, we blocked on *frequency of item*, grouping the 12 items that had been selected only 2 or 3 times per task block separately from the remaining 12 items (i.e., the frequent items) and calculated each subject's average speed for these two groups. This is shown in Figure 5. A 2x3x2 (screen size x menu type x frequency block) RM ANOVA on the speed averages did show that the type of menu differentially impacted both the time it took to select infrequent items as well as frequent items (a significant interaction effect between menu type and frequency block, $F(2,64) = 30.365$, $p < .001$, $\eta^2 = .487$). For frequently selected items, High was faster than Static and Low ($p < .001$ for both). However, for the infrequently selected items, Static was faster than both Low and High ($p < .001$ for both). This shows that High made it very efficient to access a small number of features, but the drawback was that it took longer to access the less frequently used features. While this effect may be partly due to the additional visual search time required to process the additional three items in the adaptive conditions, the higher awareness afforded by the Static menus likely made it easier to learn all the item locations more evenly.

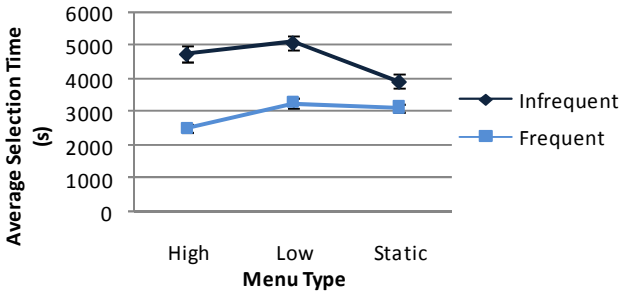


**Figure 5. Individual selection times of frequently versus infrequently selected items ($N$ =34).**

### Errors

A 2x3 (screen size x menu type) RM ANOVA showed no significant differences for error rate. Errors were uniformly low in all conditions ($M = 2.2$, $SD = 1.8$ per condition).

### Subjective measures

*High accuracy most satisfying menu in Small screen condition:* A reliability test showed that our subjective measures of difficulty, efficiency, and satisfaction measured the same internal construct (Cronbach's alpha = .858), so we collapsed these into a single *overall satisfaction* measure. A 2x3 (screen size x menu type) RM ANOVA showed that overall satisfaction was significantly impacted by a combination of the menu used and the screen size (an interaction effect, $F(1.711,58.189) = 3.489$, $p = .044$, $\eta^2 = .093$). Pairwise comparisons showed that there were no differences in satisfaction for the Large screen. For the Small screen, however, subjects were significantly more satisfied with High than they were with Low ($p = .008$) and Static ($p < .001$). This pattern reflects the speed results and is evident from the data in Figure 6.

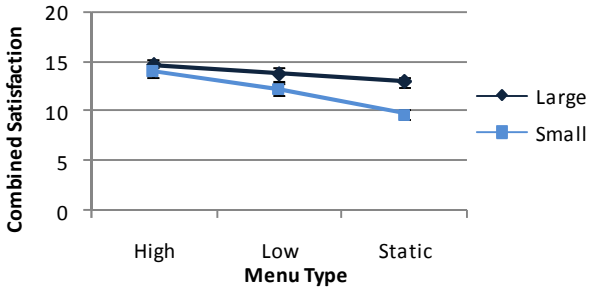


**Figure 6. Subjective satisfaction (*N* = 34).**

*Subjects perceived High to be more consistent and predictable than Low:* Our theoretical calculations for consistency and predictability of the menus aligned with subjects' perceptions. A 2x2 (screen size x menu type) RM ANOVA for the High and Low conditions showed that subjects felt that High was more consistent than Low ($F(2,32) = 7.493$, $p = .010$, $\eta^2 = .190$) and a trend suggested that subjects felt that High was also more predictable than Low ($F(2,32) = 3.868$, $p = .058$, $\eta^2 = .108$).

*High accuracy preferred in Small screens, whereas more even split between High and Static in Large screens:* As summarized in Figure 7, the majority of subjects (12/17) in the Small screen condition chose High as their preferred menu type. In contrast, preference of Large screen subjects was more evenly split between High and Static (8 and 6, respectively). Three subjects in the Small screen condition chose Low even though their speed results showed they were faster with High; when asked afterward to explain their reasoning, they had chosen Low because they found it more predictable. For the Large screen, 3 subjects could not distinguish between Low and High; their speed, order of presentation, and non-strategic selections did not provide an obvious explanation for this.

### Summary

We summarize our results with respect to our hypotheses.

*H1. Higher adaptive accuracy is faster than lower.* Supported. However, performance of Static relative to High and Low depended on screen size.

*H2. Small screen is slower than Large screen.* Supported.

*H3. Effect of adaptive accuracy on speed is greater in Small screen than Large screen.* Supported.

*H4. Higher adaptive accuracy results in lower awareness, and Static has the highest awareness.* Partially supported. High had reduced awareness in comparison to Low and Static, but there were no differences between the latter two.

*H5. Small screen has lower awareness than Large screen.* A trend shows this may be supported with more data.

*H6. Effect of adaptive accuracy on awareness will be greater in Small screen than Large screen.* Not supported. We found no interaction between screen size and menu condition (accuracy level).

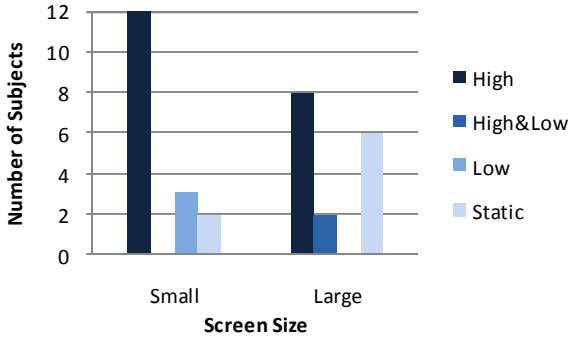


**Figure 7. Overall preference (*N* =34).**

## DISCUSSION

*Adaptive interface is more beneficial when screen real estate is constrained.* Strong evidence shows that the adaptive accuracy conditions fared better in the small screen. The high accuracy adaptive menus were significantly faster and more satisfying than the static menus for the small screen, but these differences disappeared for the large screen. Secondary analyses showed that this was likely due to a combination of the high accuracy condition reducing navigation (i.e., scrolling), and the increased use of adaptive predictions for the small screen. The latter behaviour suggests that users implicitly recognize the added benefit of the adaptive interfaces when screen real estate is constrained. These findings indicate that previous work on adaptive GUIs conducted with desktop-sized applications does not adequately generalize to small screens. Because of the increased potential benefit, researchers and designers should revisit adaptive approaches in the context of reduced screen size.

*Adaptive interfaces are low risk for small screens.* From a design standpoint, given that it is likely difficult to predict the accuracy of an adaptive interface at design time, our results suggest that there is little performance risk of using adaptive menus in a small screen (for our menu design, as long as the accuracy is at least 50%). For the small screen, the low accuracy adaptive menus were no worse than the static ones, and, if there is *potential* for them to exhibit higher than 50% accuracy, then from a performance perspective they should be beneficial. For large screens the risk is much higher: accuracy of 80% provided no performance gain relative to static menus, and 50% degraded performance. As a result, for the adaptive menus to be beneficial on the large screen, the accuracy level would theoretically need to be very high (above 80%). This analysis only considers performance, but subjective measures would need to be considered as well.

*Higher accuracy results in reduced awareness.* Extending previous work on awareness [11] to an adaptive interface, our results show that the higher accuracy condition resulted in reduced awareness. Perhaps most interesting is that the high accuracy condition had reduced awareness in comparison to the static condition, but it was not faster for the large screen, indicating that it provided no real benefit.

This suggests that a static interface may be optimal for the large screen, at least for these two measures. However, for the small screen, the high accuracy condition was significantly faster the static one, so it may be a better overall choice for the reduced screen size. The differences in awareness also suggest that the designer needs to consider what the goals are for an interface: for example, if the goal is to have the user ultimately become an expert with knowledge of a wide range of features, an interface that affords higher awareness may be preferred. Alternately, if expertise in a small number of features is sought, then awareness may be less of an issue.

*Sensitivity of awareness measure needs improvement.* We had hypothesized that the smaller screen would result in even stronger differences in awareness between the menu conditions. That this did not significantly affect the outcome could be due to a floor effect: the measure may not have been sensitive enough to detect differences in the small screen condition where awareness scores were low.

*Accuracy, consistency, and predictability require more research*. In initial pilot testing, 2 out of 4 users commented that our original low accuracy menus were more predictable than the high accuracy menus. Previous work has not studied the relative impacts of consistency, predictability, and accuracy on performance and user satisfaction, so we had planned to eliminate a possible confound by creating two accuracy conditions that had similar consistency and predictability in the full study. However, given that we required the same task for each condition and had other constraints, such as using a Zipf distribution over items, this was not a straightforward problem. The compromise was to pair higher accuracy with higher consistency and lower accuracy with lower consistency. As a result, it is unclear whether the poor performance of the low accuracy condition is attributable to accuracy, low consistency, or, most likely, to a combination of the two.

While recent work has highlighted the need to report accuracy [13,29] and consistency (alternatively called *stability*) [3,6] in addition to efficiency, our findings stress the importance of all three in combination, as well as predictability. Since the distinction between consistency and accuracy has not been addressed in previous accuracy research [13,29], further work will be needed to understand how much these two factors separately contribute to performance and satisfaction. For example, a study with fewer task constraints than the one reported here could be designed to include both consistency and accuracy as independent variables. The exact accuracy levels in our study were based on a need to have two reasonable levels that were distinct enough to impact results, but beyond that, they were based on artificial manipulations (similar to [13,29]). Further work is needed to understand how similar the findings would be for other levels of accuracy.

*Adaptive menu models should account for differential usage of adaptive predictions.* Cockburn et al. have provided compelling results for modeling adaptive menus [6]. However, their model for adaptive split menus assumes that users will select from the top, adaptive section if the item is there; both our results and those of Gajos et al. [13] show this is not always the case. In addition, Cockburn et al. acknowledge that their model does not incorporate incidental learning (which we measured as awareness). Since an adaptive interface can impact awareness, an obvious extension of the model would be to incorporate it.

*Generalizability of the results to other GUI control structures.* Although further work is needed, the speed and awareness differences between the small and large screens should be equally applicable to other types of GUI control structures, such as toolbars and the MS Office 2007 Ribbon. It is also possible that the particular visual display of features provided in toolbars and the Ribbon will result in similar awareness of the *number* of features available, but lower awareness of the specific *actions* that may be carried out by those features, since the images may not as directly convey this information to users as menu labels do.

### Limitations of the Experiment

*Replication in realistic task context.* For a task consisting of only menu item selections, such as the one included in our study, users may be more likely to utilize the adaptive component of the menu because they will value efficiency over other aspects of the interaction. It will be interesting to replicate this work in a more realistic setting where the user's cognitive resources for any given task are divided, and menu selection is but one part of the task. For example, 6/17 subjects preferred the static menus in the large screen condition, but in a more realistic setting this may increase. It will also be interesting to study the long-term impact of differences in awareness, such as on an experienced user's ability to complete a new task.

*Task appropriate for small screen devices.* Further work is needed to understand how our results will apply to tasks specific to mobile computing with small screen devices, and to replicate the work on a mobile device, using pen or stylus input, instead of the simulation we used. Even if mobile application interfaces are simpler than desktop ones, the relative benefit of an adaptive interface may be greater since the user's attention is more fragmented in a mobile context than in a more standard computing context [24].

## CONCLUSION

Through a controlled lab study, we have provided empirical evidence to show that high accuracy adaptive menus may have a larger positive benefit on small screen displays than regular desktop-sized displays. Not only was this shown through direct performance and user satisfaction measures, but we also found that screen size impacts user behaviour: subjects were more likely to make use of the adaptive predictions in the small screen condition than the large screen one. We also found that high adaptive accuracy menus negatively impacted the user's overall awareness of features in the interface, which may be important for longer-term performance and satisfaction. Finally, our

results highlight the importance of considering adaptive performance in the context of accuracy, since the lower and higher accuracy adaptive menus performed differently in relation to their static counterpart when screen size varied.

Overall, these findings stress the need to revisit previous adaptive research in the context of small screen devices. Approaches which may not have been shown to be beneficial on larger screens may be more advantageous in a small screen context. Further work is needed to understand how well our results will generalize in the field, where user tasks are more complex and there are many more demands on the user's attention. Nonetheless, the study presented here provides encouraging evidence that GUI adaptation is a viable design direction for small screen devices.

## ACKNOWLEDGMENTS

We thank Andrea Bunt, Heidi Lam, and Tony Tang for their valuable comments. We also thank IBM Centers for Advanced Studies and NSERC for funding.

## REFERENCES

1. Baddeley, A. *The Psychology of Memory*. Basic Books, New York, 1976.
2. Billsus, D., Brunk, C. A., Evans, C., Gladish, B., and Pazzani, M. 2002. Adaptive interfaces for ubiquitous web access. *CACM 45*, 5 (2002), 34-38.
3. Bridle, R., and McCreath, E. Inducing shortcuts on a mobile phone interface. *Proc. IUI*, (2006), 327-329.
4. Bunt, A., Conati, C., and McGrenere, J. Supporting interface customization using a mixed-initiative approach. *Proc. Intelligent User Interfaces*, (2007), 92-101.
5. Bunt, A. *Mixed-Initiative Support for Customizing Graphical User Interfaces*. Ph.D. Thesis, University of British Columbia, 2007.
6. Cockburn, A., Gutwin, C., and Greenberg, S. A predictive model of menu performance. *Proc. CHI*, (2007), 627-636.
7. Cohen, J. Eta-squared and partial eta-squared in communication science. *Human Communication Research*, 28, Oxford Journals, (1973), 473-490.
8. Debevc, M., Meyer, B., Donlagic, D., and Svecko, R. Design and evaluation of an adaptive icon toolbar. *User modeling and user adapted interaction* 6, 1(1994), 1-21.
9. Dillon, A., Richardson, J., and McKnight, C. The effects of display size and text splitting on reading length text from screen. *Behaviour and Information Technology* 9, 3(1990), 215-227.
10. Findlater, L. and McGrenere, J. A comparison of static, adaptive and adaptable menus. *Proc. ACM CHI 2004*, (2004), 89-96.
11. Findlater, L. and McGrenere, J. Evaluating reduced-functionality interfaces according to feature findability and awareness. *Proc. IFIP Interact 2007*, (2007), 592-605.
12. Gajos, K., Christianson, D., Hoffmann, R., Shaked, R., Henning, K., Long, J.J., and Weld, D.S.. Fast And Robust Interface Generation for Ubiquitous Applications. *Proc. UBICOMP '05*, 2005), 37-55.
13. Gajos, K. Z., Czerwinski, M., Tan, D. S., and Weld, D. S. Exploring the design space for adaptive graphical user interfaces. *Proc. AVI '06*, (2006), 201-208.
14. Greenberg, S. *The computer user as toolsmith: The use, reuse, and organization of computer-based tools*. Cambridge: Cambridge University Press, 1993.
15. Greenberg, S. and Witten, I. Adaptive personalized interfaces: A question of viability. *Behaviour and Information Technology 4*, 1(1985), 31-45.
16. Greenberg S. and Witten, I. Directing the user interface: how people use command-based computer systems. *Proc. 3rd IFAC Conference on Man-Machine Systems*, (1988), 349-356.
17. Hansen, S., Kraut, R., and Farber, J. Interface design and multivariate analysis of Unix commands. *ACM TOIS 2*, 1(1984), 42-57.
18. Höök, K. Steps to take before intelligent user interfaces become real. *Journal of Interacting with Computers 12*, 4(2000), 409–426.
19. Jones, M., Buchanan, G., and Thimbleby, H. Improving web search on small screen devices. *Interacting with Computers* 15, (2003)4, 479-495.
20. Linton, F., Joy, D., Schaefer, H.-P., and Charron, A. Owl: A recommender system for organization-wide learning. *Educational Technology & Society* 3, 1(2000).
21. McGrenere, J. and Moore, G. Are we all in the same "bloat"? *Proc. of GI 2000*, (2000), 187-196.
22. McGrenere, J., Baecker, R., and Booth, K. An evaluation of a multiple interface design solution for bloated software. *CHI Letters 4*, 1(2002), 163–170.
23. Mitchell, J. and Shneiderman, B. (1989). Dynamic versus static menus: An exploratory comparison. *SIGCHI Bulletin 20*, 4(1989), 33-37.
24. Oulasvirta, A., Tamminen, S., Roto, V., and Kuorelahti, J. Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI. *Proc CHI '05*, (2005), 919-928.
25. St. Amant, R., Horton, T. E., and Ritter, F. E. Model-based evaluation of expert cell phone menu interaction. *ACM TOCHI 14*, 1(2007).
26. Sears, A., and Shneiderman, B. Split menus: Effectively using selection frequency to organize menus. *ACM TOCHI 1*, 1(1994), 27 - 51.
27. Smyth, B., and Cotter, P. Personalized adaptive navigation for mobile portals. *Proc. ECAI'02*, (2002), 608-612.
28. Smyth, B., Cotter, P., and Oman, S. Enabling intelligent content discovery on the mobile internet. *Proc. AAAI*, (2007), 1744-1751.
29. Tsandilas, T. and schraefel, m. c. An empirical assessment of adaptation techniques. *CHI '05 Extended Abstracts*, (2005), 2009-2012.