Beyond Performance: Feature Awareness in Personalized Interfaces

Leah Findlater
The Information School
University of Washington
Seattle, WA, USA
leahkf@uw.edu

Joanna McGrenere
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
joanna@cs.ubc.ca

Abstract
Personalized graphical user interfaces have the potential to reduce visual complexity and improve interaction efficiency by tailoring elements such as menus and toolbars to better suit an individual user's needs. When an interface is personalized to make useful features more accessible for a user's current task, however, there may be a negative impact on the user's awareness of the full set of available features, making future tasks more difficult. To assess this tradeoff we introduce *awareness* as an evaluation metric to be used in conjunction with performance. We then discuss three studies we have conducted, which show that personalized interfaces trade off awareness of unused features for performance gains on core tasks. The first two studies, previously published and presented only in summary, demonstrate this tradeoff by measuring awareness using a recognition test of unused features in the interface. The studies also evaluated two different types of personalized interfaces: a layered interfaces approach and an adaptive split menu approach. The third study, presented in full, focuses on adaptive split menus and extends results from the first two studies to show that different levels of awareness also correspond to an impact on performance when users are asked to complete new tasks. Based on all three studies and a survey of related work, we outline a design space of personalized interfaces and present several factors that could affect the tradeoff between core task performance and awareness. Finally, we provide a set of design implications that should be considered for personalized interfaces.

Keywords
Personalization; Adaptive user interfaces; User studies

1. Introduction
Feature-rich graphical user interfaces dominate the software application landscape, from word processors to integrated development environments. These interfaces support a wide range of tasks and provide many necessary features, but are not tailored to an individual user's needs. The complexity can be daunting for novice users, who may be overwhelmed by the sheer number of options available, and it can be problematic for expert users, who tend to use only a small subset of features (Linton et al., 2000; McGrenere and Moore, 2000). To reduce complexity and improve interaction efficiency, personalized graphical user interfaces (GUIs) modify interface elements such as menu and toolbar items to better suit an individual's pattern of use.

Many GUI personalization approaches have appeared in research and commercial applications. With *adaptive split menus*, for example, the menu items most likely to be needed by the user are automatically copied to the top of the menu to make them more easily accessible (Sears and Shneiderman, 1994). As a user-controlled personalization example, *layered interfaces* allow the user to switch between several interfaces to the application, choosing the one that best suits his or her needs at a given point in time (Shneiderman, 2003); novice users may begin working in an interface layer that contains only a small, core set of features before transitioning to a more complex layer (e.g., moving from the minimal layer to the full layer in Figure 1). We use the term personalization to refer to approaches that are *adaptive* (system-controlled), *adaptable* (user-controlled), or *mixed-initiative* (a combination of the previous two).

Findlater, L., McGrenere, J. 2009. Beyond performance: feature awareness in personalized interfaces. *International Journal of Human Computer Studies*, to appear. doi:10.1016/j.ijhcs.2009.10.002

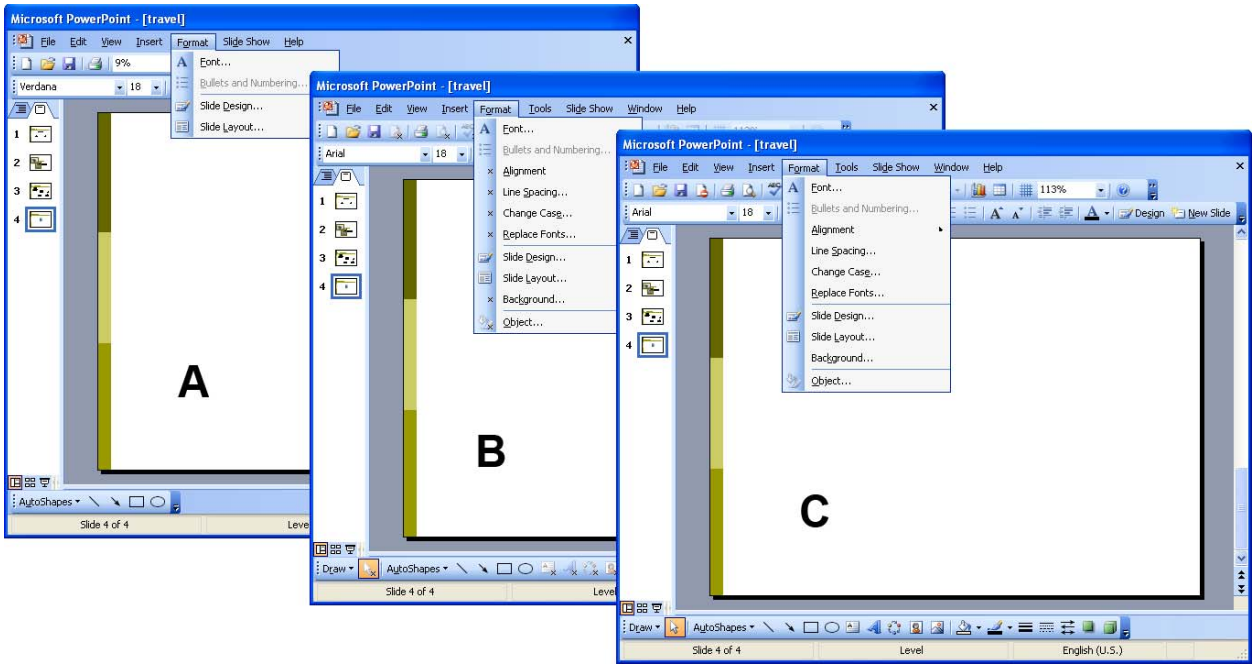

Figure 1. Sample screenshots from the interface layers used in Study 1: minimal interface layer (A), marked interface layer (B), and full interface layer (C).

Working in a personalized GUI has many advantages: such interfaces can make novice users faster, more accurate and more satisfied (Carroll and Carrithers, 1984), and can even be preferred by a large portion of more experienced users (McGrenere et al., 2002). Adaptive personalization of menus and toolbars is especially effective when the adaptation greatly reduces the amount of navigation needed to reach items (Findlater and McGrenere, 2008; Greenberg and Witten, 1985), or tailors the layout of interface elements to the specific abilities of motor-impaired individuals (Gajos et al., 2008b). Researchers have identified drawbacks of adaptable and adaptive mechanisms (e.g., Höök, 2000; Mackay, 1990), yet the underlying premise of previous work is that personalization is inherently beneficial if the mechanisms can be designed right. While we do not dispute that interface personalization can be beneficial, we argue that working in a personalized interface can impact users in ways that are not necessarily captured by the *traditional* measures of user satisfaction and performance.

When the menu and toolbar items in an interface are personalized to make it easier to access features most relevant to a user's current task, this personalization may negatively impact the user's awareness of the full set of available features, features that might be potentially useful in the future. A survey of 53 users of Microsoft Word 97 reflects this tension between core task performance and the ability to learn about new features: while many users requested that their unused features be tucked away, many also indicated that it was important to be able to continually discover new features (McGrenere and Moore, 2000). More recently, we conducted an interview study with 14 users of a complex integrated development environment that provides personalization based on the user's work role (Findlater et al., 2008). We found that more than half the participants were concerned about hiding features, in part because it could impair their ability to learn about and use new features in the interface.

To explore the tension between personalizing to improve performance versus the user's ability to learn about new features, we defined feature *awareness* as a new evaluation measure (Findlater and McGrenere, 2007). Awareness of a feature in an application is a degree of knowledge about that feature; when applied to features that have not yet been used, it is a measure of the secondary, incidental learning that may occur as the user performs a primary task. We have proposed that measuring awareness in conjunction with performance (efficiency) is particularly valuable for personalized interfaces. Design decisions in non-personalized interfaces will also impact awareness and performance (e.g., choosing a

broader vs. deeper hierarchical menu structure), but the impact on these measures is likely greater in a personalized interface.

Together, awareness and performance offer a broader understanding of the impact of working in a personalized interface. Although awareness does not impact performance on routine or known tasks (those supported by the personalization and familiar to the user), it has the potential to impact performance on new tasks. We thus distinguish *core task performance* from *new task performance*. We have operationalized awareness using two methods (Findlater and McGrenere, 2007): (1) as the recognition rate of unused features in the interface, and (2) as the user's performance when completing new tasks that require previously unused features. Although awareness is not a replacement for longer-term field studies, it does provide designers with a lower-cost tool that can indicate the potential impact a design will have on future performance.

We have conducted three controlled lab studies to demonstrate and characterize the tradeoff between core task performance and awareness when working in a personalized interface. The first two studies, previously published, compared different types of personalized interfaces to a control condition to demonstrate a measurable tradeoff between core task performance and awareness, as measured using the recognition rate of unused features. Study 1 (Findlater and McGrenere, 2007) evaluated a layered interfaces approach to personalization, while Study 2 (Findlater and McGrenere, 2008) evaluated an adaptive split menu approach. As both of these studies have already been published, we present them only in summary. The goal of Study 3, presented in full in this paper, was to extend results from the first two studies to provide evidence that personalization impacts our second measure of awareness: performance on new tasks. Based on Study 2, users worked with a static control condition or one of two adaptive split menu conditions that predicted the user's needs with different degrees of accuracy (50% vs. 78%). As expected, participants were fastest at completing new tasks in the control condition, while the high accuracy adaptive condition provided the best core task performance. The low accuracy adaptive condition provided neither a core task nor a new task performance benefit over the control condition.

This paper brings together work to date on awareness and personalized interfaces. We first outline a design space of GUI personalization, identifying four factors that are particularly important when considering the interplay between awareness and core task performance: control, granularity, visibility, and frequency of personalization. We then refine our definition of awareness and how we operationalize it, and summarize findings from Studies 1 and 2 that demonstrate a previously unidentified tradeoff between awareness and core task performance. Those studies evaluated two types of personalized interfaces (adaptive split menus and layered interfaces), which allows for a degree of generalizability and suggests that evaluating awareness will be important for a range of personalized GUIs. We then present Study 3, which provides new empirical evidence showing that working in a personalized interface can negatively impact performance when users are asked to complete *new* tasks. Finally, we discuss the findings from all three studies within the context of our design space, present several design guidelines for personalized GUIs, and identify fruitful areas for future work.

2. Design Space of Personalized GUIs

Based on a survey of related work, we have identified four design factors for personalized interfaces that are particularly important for the interplay between core task performance and awareness: control, visibility, frequency, and granularity of the adaptation. In this section we discuss related work within these design factors, including previous evaluations that have measured performance and user satisfaction. Later (Section 6) we also return to the design space to reflect on our study results and on how each of the factors affects core task performance and awareness. The design space is summarized in Table 1; note that shaded cells represent combinations of factors that have not yet been evaluated in the research literature.

Generally speaking, personalization to reduce complexity or improve core task performance can be grouped into two main categories: (1) personalization of content (e.g., reducing software code complexity

(Kersten and Murphy, 2005), collaborative filtering (Terveen et al., 2002), and recommender systems (Herlocker et al., 2004)), and (2) personalization of GUI control structures (e.g., reducing menu and toolbar options). While it is not always possible to clearly distinguish between the two types of personalization, our design space focuses on GUI control structures. We are particularly interested in lightweight GUI customization mechanisms, in contrast to approaches such as end-user programming and tailorable systems (e.g., Eisenberg and Fischer (1994); Stamoulis et al. (2001)), which typically require deeper technical expertise and effort on the part of the user.

Control		*Adaptive*		*Adaptable*		*Mixed-initiative*	
Granularity		*Coarse*	*Fine*	*Coarse*	*Fine*	*Coarse*	*Fine*
Visibility of change	*Hidden*		MS Office 2003 adaptive menus	Layered interfaces (Clark and Matthews, 2005; Findlater and McGrenere, 2007; Gustavsson Christiernin et al., 2003; Plaisant et al., 2003; Shneiderman, 2003) User role-based (Findlater et al., 2008; Greenberg, 1991)	Multiple interfaces (McGrenere et al., 2002)		Incremental interfaces (Brusilovsky and Schwarz, 1997) Adaptive bar (Debevc et al., 1995) Adaptively supported multiple interfaces (Bunt et al., 2007)
	Moved		Original split menus (Sears and Shneiderman, 1994) Frequency based menus (Mitchell and Shneiderman, 1989) Adaptive hierarchical menus (Greenberg and Witten, 1985) Adaptive split menus (Findlater and McGrenere, 2008) and toolbars (Gajos et al., 2006)				
	Resized		Ability-based interfaces (Gajos et al., 2008b) Morphing menus (Cockburn et al., 2007)				
	Replicated		Replicated split interfaces (Findlater and McGrenere, 2008; Gajos et al., 2006; Gajos et al., 2008a)		Facades (Stuerzlinger, et al., 2006)		
	Marked		Colour highlighting (Tsandilas and schraefel, 2005) Ephemeral adaptation (Findlater et al., 2009)	Marked layered interface (Findlater and McGrenere, 2007)			

Table 1. Design space for personalized GUIs, outlining existing approaches. The adaptive approaches listed here all provide frequent adaptation (except Shneiderman (2003) and Gajos et al. (2008b)); the adaptable and mixed-initiative approaches provide persistent adaptation. Shaded regions represent combinations that have not been evaluated.

2.1 Control of Personalization

Personalization approaches may be adaptive, adaptable, or mixed-initiative. Adaptive personalization mechanisms require no effort on the part of the user, and can improve performance especially when they greatly reduce the amount of navigation required to reach items (Gajos et al., 2006; Greenberg and Witten, 1985; Findlater and McGrenere, 2008). However, the costs of adaptive personalization, such as unpredictability or instability of the interface layout, can sometimes be so great that they negatively impact performance (Mitchell and Shneiderman, 1989; Findlater and McGrenere, 2004). An important aspect of adaptive personalization is the accuracy with which the adaptive algorithm can predict the user's needs: users are faster when adaptive accuracy is higher (Findlater and McGrenere, 2008; Gajos et al., 2006; Tsandilas and Schraefel, 2005).

A major challenge of adaptable personalization approaches, on the other hand, is that the extent to which users adapt their interface is dependent on skill level and interest, with some users not making any adaptations at all (Mackay, 1990; MacLean et al., 1990). Although adaptable personalization approaches can improve user satisfaction (McGrenere et al., 2002), there has been less evaluation of the impact of adaptable personalization on user performance. One exception is a comparison of adaptive and adaptable split menus, which showed that users were able to personalize their menus effectively when they chose to do so, resulting in faster performance than the adaptive counterpart (Findlater and McGrenere, 2004). Subsequent research that has improved upon adaptive split interfaces has not compared them to adaptable ones (Findlater and McGrenere, 2008; Gajos et al., 2006).

In addition to adaptive accuracy (mentioned above), several characteristics related to control can also impact the user's experience. For example, the predictability of adaptive personalization can affect user satisfaction (Gajos et al., 2008a). The user's trust in an adaptive algorithm's ability to predict his or her needs can also impact user behaviour, where lower trust often results in users simply ignoring the predictions (Findlater and McGrenere, 2008; Tsandilas and schraefel, 2005). Trust may also be a factor with an adaptable approach, since the user needs to be confident (trust) in his/her own ability to predict future needs. Mixed-initiative approaches to personalization have also been proposed, where the system provides adaptive suggestions to aid the user in adapting the interface (e.g., Brusilovsky and Schwarz, 1997; Debevc et al., 1994; Bunt, Conati, and McGrenere, 2007). Bunt, Conati and McGrenere (2007) showed that including adaptive suggestions in an adaptable system improved user satisfaction and reduced the amount of time the user spent adapting the interface. As can be seen in Table 1, there has been less evaluation of mixed-initiative GUI personalization than of either adaptive or adaptable approaches.

2.2 Granularity

Fine-grained personalization approaches modify the interface one individual feature at a time. For example, with McGrenere et al.'s (2002) multiple interfaces approach, the user can switch between the full version of the interface and a personal version, where he/she has specified each menu or toolbar item that appears in the personal version. The original split menus (Sears and Shneiderman, 1994) and other split interface variants (e.g., Gajos et al. 2006) offer another example of fine-grained personalization: individual items that the adaptive algorithm (in an adaptive case) or the user (in an adaptable case) deems to be the most useful are copied to a designated section of the interface, such as the top of the menu, for easier access.

In contrast, more *coarse-grained* approaches modify the interface by manipulating large groups of related features at once. The layered interfaces approach (introduced by Shneiderman, 2003) is a coarse-grained approach: in the example shown in Figure 1, moving from the minimal layer to the full layer introduces a large set of advanced features. Another coarse-grained approach is to personalize the interface based on the user's work role (e.g., Findlater et al., 2008; Greenberg, 1991), where different sets of features are associated with different roles and only those features associated with the specific user's role are enabled in the interface. Evaluations of coarse-grained personalization approaches have been mainly qualitative (Clark and Matthews, 2005; Findlater et al., 2008; Gustavsson Christiernin et al., 2003; Plaisant et al., 2003; Shneiderman, 2003).

As seen in Table 1, adaptive techniques have generally been fine-grained while coarse-grained techniques have been limited to adaptable approaches. Adaptive, coarse-grained personalization may not be a feasible combination: since every adaptation introduces a relatively large change to the interface it would make sense to have at least some degree of user control over that change (i.e., mixed-initiative control). When the personalization is user-controlled, a coarse-grained approach will require less effort than a fine-grained approach. However, a fine-grained approach should also be able to more precisely tailor the interface to suit a user's needs at any given point in time.

2.3 Visibility of Change

Personalization approaches offer a variety of options for the visual affordance of the adaptation. Some personalization approaches *hide* unnecessary features from view (McGrenere et al., 2002; Shneiderman, 2003), while others *move* (Findlater and McGrenere, 2004; Greenberg and Witten, 1985; Mitchell and Shneiderman, 1989; Sears and Shneiderman, 1994), *replicate* (Findlater and McGrenere, 2008; Gajos et al., 2005; Gajos et al., 2006; Stuerzlinger et al., 2006), *mark* (Findlater et al., 2009; Gajos et al., 2005; Tsandilas and Schraefel, 2005), or *resize* (Cockburn et al., 2007; Gajos et al., 2008b) the most salient features to reduce navigation time and/or visual or cognitive complexity. When hiding features, the interface can still provide a degree of visibility, such as the chevron at the bottom of the MS Office 2003 adaptive menus. The early training wheels approach (Carroll and Carrithers, 1984; Catrambone and Carroll, 1987) provided *no visual cue* as to which features were blocked; although that work did show some promise, more recent work that used this approach in the context of a graphical user interface yielded negative results (Bannert, 2000). Overall, it does not seem to be a viable approach and so it is not included in Table 1.

Hiding, moving, replicating and resizing features are all *spatial* adaptation techniques. Spatial adaptation techniques introduce spatial instability into the interface layout, but can improve performance when they greatly reduce navigation, for example, through a hierarchical menu (Greenberg and Witten, 1985). In contrast, the goal of *marking* techniques is to reduce visual search time by drawing the user's attention to important features. Gajos et al. (2005) and Tsandilas and schraefel (2005) have proposed adaptive colour highlighting, where a small set of menu or toolbar items that the adaptive algorithm predicts the user will need are highlighted in a light purple colour to draw visual attention; however, a recent follow-up evaluation has not demonstrated a performance benefit to this technique (Findlater et al., 2009). An alternative technique that does offer a performance benefit is ephemeral adaptation: for example in a pull-down menu, a small set of adaptively predicted items appear immediately when the menu is opened, drawing visual attention, while the remaining items gradually fade in after a brief delay (Findlater et al., 2009).

2.4 Frequency of Change

The frequency with which the personalized UI changes may range from as frequent as every user interaction to a much longer term, such as weeks, months or years. As seen Table 1, adaptive approaches generally change the interface after every interaction (e.g., Findlater and McGrenere, 2004; Gajos et al., 2006; Mitchell and Shneiderman, 1989; Tsandilas and schraefel, 2005), although this does not have to be the case. For example, one exception is the original split menu work by Sears and Shneideman (1994), where the menus adapted only once over a 5 week study. Gajos et al. (2006) note that frequency of adaptation appears to impact the cost/benefit of adaptive GUIs, since the relatively slow pace of adaptation in the original split menu study yielded more positive results than a follow-up study by Findlater and McGrenere (2004), where adaptation occurred more frequently. Researchers have also proposed adaptively personalizing the layout of the interface according to the current document (Debevc et al., 1994). Adaptable approaches only change as often as the user chooses to make modifications. Realistically, however, it does not make sense for users to adapt the interface after every interaction, so the frequency of change is likely to be less often than many of the adaptive approaches.

2.5 Summary

We have outlined four personalized interface design factors. Although these factors are not exhaustive, they are particularly important for understanding the impact that personalized interfaces can have on performance and awareness of the full set of features in an application, and we used them to inform the design of our studies. Although many of the shaded cells in Table 1 have simply been unexplored and offer an opportunity for future work, we have also highlighted several combinations of design characteristics that are not practically feasible. =-

3. Awareness and Performance Definitions

We define *awareness* of an unused feature in an application as a degree of knowledge about that feature that has developed consciously or unconsciously as the user accomplishes a primary task; it is thus a measure of secondary learning.

The focus of personalized interfaces is generally to improve *core task performance*, that is, performance of completing known or routine tasks, rather than to build awareness. We predict that lower awareness in the present will, in turn, impact performance when the user is asked to complete new tasks. We thus differentiate *new task performance* from core task performance. For new, complex tasks, performance is impacted both by the time it takes to complete the steps with which the user is already familiar (core task performance), which are those most likely to be supported by the personalized interface, and the time it takes for the user to "discover" how to complete new, unfamiliar steps (new task performance). The time to complete the latter steps will be in part related to prior awareness of unused features.

We operationalize two measures of awareness:

1. *Recognition rate of unused features.* The ability of experienced users to recognize features that are available in the interface, but that they have not yet used (see Section 5.1.4 for more detail on the format of our recognition test).
2. *New task performance*. The speed with which experienced users can locate, when prompted, previously unused features. In contrast to the recognition rate, this is an applied measure of awareness and is more effortful to assess because it requires asking users to perform new tasks. However, it may be a more direct indicator of the impact that awareness can have in the longer term.

Awareness is only one component of performance when selecting graphical user interface elements. Performance and user satisfaction also depend on a number of factors, including user characteristics, such as experience or cognitive and motor abilities, and interface characteristics, such as layout. However, awareness is one aspect of the user's experience that is particularly important for personalized approaches, where the impact on awareness may be greater than in more traditional user interface designs. Since personalization approaches will impact core task performance and awareness to differing degrees, distinguishing between the two measures allows for a more nuanced comparison of designs than measuring performance alone.

Awareness is undoubtedly related to learning more generally. In a recent evaluation of methods for assessing learnability, Grossman, Fitzmaurice and Attar (Grossman et al., 2009) identified awareness of functionality as one category of learnability issues, in addition to understanding task flow, locating functionality, understanding how to use functionality, and transitioning to more efficient interaction. Although there has been some work on mechanisms to suggest new functionality to users (e.g., Linton, 2000), previous research has not studied awareness and personalization. Awareness, and knowledge in general, is a complex phenomenon, and the secondary learning assessed with our measures of awareness likely includes both implicit and explicit knowledge about a feature. It is also likely that these types of knowledge are captured to differing degrees by the two distinct measures of awareness, which is a motivation for providing more than one measure. It would be interesting, however, for future work to explore more specifically what type of learning contributes to each measure of awareness.

We have incorporated awareness into three controlled laboratory studies: Studies 1 and 2 demonstrate a tradeoff between awareness and core task performance using the recognition rate measure of awareness, while Study 3 extends these results to the new task performance measure. Awareness and core task performance may be impacted by the design factors discussed in Section 2, which we have also taken into account in our studies. In Study 1, we evaluated an adaptable approach, called layered interfaces, that either hides or visually marks advanced features in the interface. Studies 2 and 3 focused on adaptive split menus, where those items predicted to be the most useful to the user are replicated at the top of the menu for easier access.

4. Impact of Personalization on Recognition Rate of Unused Features: Studies 1 and 2

We previously published two studies measuring the impact of personalized interface designs on awareness and core task performance. Each study evaluated a different type of personalization approach and showed that working in a personalized interface can positively impact core task performance but negatively impact awareness. We briefly summarize the results here.

4.1 Study 1: Awareness Recognition Rate and Layered Interfaces

We first conducted a proof-of-concept study (Findlater and McGrenere, 2007) with 30 participants to demonstrate that a measurable tradeoff exists between core task performance and awareness for at least one type of personalized interface.

In a controlled lab setting, the study compared two 2-layer interface designs for Microsoft PowerPoint 2003 (Minimal and Marked) to a control condition. The conditions were based on the interface layers shown in Figure 1: the Minimal and Marked conditions provided both a reduced-functionality interface layer (minimal or marked) and a full interface layer, while the control condition provided only the full interface layer.

We hypothesized that the personalized conditions would provide better core task performance than the control but would result in lower awareness of unused features because they did not offer as much opportunity for interacting with the full feature set. The experimental conditions allowed us to evaluate specific design elements that we predicted would impact core task performance and awareness. The first design element is the visibility of the personalized features: (1) the Minimal approach *hid* advanced features in the reduced-functionality layer, and (2) the Marked approach visually *marked* advanced features in the reduced-functionality layer. We anticipated that visually distinguishing (marking with a small 'x' as in Figure 1), but not removing blocked features could offer a compromise between core task performance and awareness. Based on interviews with 10 PowerPoint users, we defined core task features as those used by at least 80% of users.

All participants were novice users of PowerPoint and completed both a core and advanced task in one of the three experimental conditions (a between-subjects design); the advanced task required features beyond those used in the core task. With the personalized conditions, the core task was done in the reduced-functionality interface layer, and the advanced task was done in the full interface layer. This simulated the projected behaviour of users in a layered interface as they move from more basic tasks in simple layers to more advanced tasks in more complex layers (Shneiderman, 2003). The Control condition used the full interface for both tasks. Each task consisted of a series of step-by-step instructions to edit an existing slide presentation (30 steps for the core task and 48 steps for the advanced task). Each step required a specific menu or toolbar item but participants were not told the exact item, for example, "Draw an arrow from the rectangle to the triangle." Awareness was measured by: (1) administering a recognition test after both tasks were completed, and (2) analyzing performance on those steps that required new features in the advanced task.

Results were in line with our hypotheses. First, core task performance was better in the Minimal than the Control condition; that is, participants accessed required menu and toolbar items more quickly. However, once both tasks were completed, Control participants were more aware than Minimal participants of features that had not been used for either task. Trend-level results suggested the Marked condition may have had a small positive effect on core task performance and awareness, but almost all participants felt they would prefer to use the full interface alone over the Marked one. Across all conditions, the core task took on average 16 minutes and the advanced task took 26 minutes.

These results are encouraging and suggest that incorporating an awareness measure into evaluating personalized interfaces can add value. Taken in isolation, the core task performance results replicated related work on training wheels interfaces (Carroll and Carrithers, 1984), and could lead us to reach the

straightforward conclusion that the Minimal 2-layer interface is better than the full interface alone. By teasing apart performance and demonstrating that improved performance on core tasks can come at a cost of decreased awareness, we provided a richer understanding of the impact of working in a layered interface.

Unfortunately, the two measures of awareness in our study produced inconsistent results. Unlike the recognition test, the new task performance measure provided no support for our hypotheses that the Minimal condition would result in the least awareness, the Control the most, and the Marked condition would be in between. However, since the recognition test scores provided partial support for the hypotheses, we still believed there should be an indirect impact of awareness on performance. Our inability to detect this difference may have been due to a lack of statistical power for the new task performance measure: the impact of awareness on the complex task could have been small relative to the overall difficulty and time needed to find new features in the full interface. We address this issue in Study 3.

4.2. Study 2: Awareness Recognition Rate and Adaptive Split Menus

Building on the promising results from Study 1, we conducted a second study (Findlater and McGrenere, 2008) to replicate and extend those results to another type of personalized interface. Study 2 included 36 participants and evaluated the impact of adaptive split menus and screen size on core task performance, awareness and user satisfaction. With an adaptive split menu (shown in Figure 2), the items predicted to be most useful to the user are copied to the top part of the menu, above a "split" (Sears and Shneiderman, 1994).

Adaptive split menus are sufficiently different from layered interfaces to increase our confidence in the generalizability of the tradeoff between core task performance and awareness. The two most important differences between adaptive split menus and layered interfaces are: (1) the system controls the adaptation, and (2) the personalization mechanism spatially rearranges items in the interface, but does not otherwise visually mark or hide any from view. The focus of Study 2 was not on awareness, so we summarize only the relevant results here.

Since the accuracy of personalization can affect performance and satisfaction with adaptive interfaces (Tsandilas and schraefel, 2005; Gajos et al., 2006), we included two adaptive conditions whose predictions matched the user's needs with different levels of accuracy (50% and 78%); accuracy indicates how often the user opened a menu to search for an item and that item had been replicated at the top of the menu (see Section 5.1.1 for detail since we used similar menu conditions in Study 3). We compared the two adaptive conditions (low and high accuracy) to a static control condition. We also varied screen size to be either a desktop-sized screen or a PDA-sized screen.

Participants completed a series of menu selections using each of the three types of menus in either the small (PDA-sized) or large (desktop-sized) screen condition. Menu type was a within-subjects factor and screen size was a between-subjects factor. Since not all menu items could be displayed at once on the smaller screen, participants had to scroll to select some items in that condition.

For each selection, participants were provided with the name of the item to be selected, but were not told specifically in which of the three menus it would be found. This provided a more constrained task in contrast to Study 1, where participants were told what steps they needed to complete but not how to do those steps at the detail of individual menu or toolbar items. Another difference from Study 1 is that Study 2 was not designed to detect an impact of awareness on new task performance. Instead, participants simply repeated the same block of menu selections twice, after which they completed an awareness recognition test of unused features.

Results showed a tradeoff between core task performance and awareness recognition test scores for the high accuracy adaptive menus in comparison to the control condition. The high accuracy adaptive menus were faster than the control condition for the small screen, and those two conditions were no different for

the large screen. Overall, however, the high accuracy menus resulted in significantly lower awareness scores than the control condition. Although the low accuracy menus resulted in higher awareness scores than the high accuracy menus, they did not offer a performance benefit over the control condition. Finally, a trend indicated that screen size may impact awareness: the smaller screen resulted in lower awareness than the larger screen, likely because not all menu items could be viewed at once.

4.3. Discussion of Studies 1 and 2

Studies 1 and 2 provide evidence that personalized interfaces can tradeoff core task performance for the user's reduced awareness of unused features, which validates subjective concern (Findlater et al., 2008; McGrenere and Moore, 2000) that personalization impacts the ability to learn about new features. The results also provide a degree of generalizability because they demonstrate this tradeoff for two personalization techniques, layered interfaces and adaptive split interfaces, which vary on several characteristics. The main limitation of these studies, however, is that they do not conclusively show whether or not differences in awareness recognition test scores translate to a performance impact when the user is asked to complete new tasks. For Study 1, this was likely due to a lack of statistical power, while Study 2 did not measure the indirect impact of awareness on performance at all. The main goal of Study 3 is to address this limitation by using the more constrained task and adaptive split menus from Study 2, but with an experimental design that allows us to measure a potential indirect impact of awareness on performance.

5. Impact of Adaptive Split Menus on New Task Performance: Study 3

The first two studies confirmed users' concerns that working in a personalized interface could impact the ability to learn about new features, as measured with the awareness recognition test. However, those studies did not provide us with an understanding of whether the user's current level of awareness impacts *performance* in the future. To revisit this hypothesis, we conducted a controlled lab experiment with 30 participants, using a similar menu selection task and within-subjects design as in Study 2. To keep the study sessions to a reasonable length, using a within-subjects design meant that participants would have less exposure to each interface than in our previous attempt to measure this hypothesis (Study 1). The advantage, however, was that statistical power would be increased. This study has not been reported in the literature, so we include all the details here.

5.1 Experimental Methodology

5.1.1 Conditions

The experimental conditions each displayed a set of 3 menus, and differed as follows:

1. *High*: Adaptive split menus that predicted the user's needs with 78% accuracy, on average; that is, 78% of the time the user needed to select an item, it could be found within the top 3 items in the menu.
2. *Low*: Adaptive split menus that predicted the user's needs with 50% accuracy, on average.
3. *Control*: Traditional static menus.

The adaptive conditions were the same as those used in Study 2's large screen condition: each menu contained 24 items, and the 3 items most likely to be needed by the user were copied to the top of the menu, above the split (see Figure 2). Although performance results from Study 2 showed that the adaptive conditions were more beneficial in the small screen condition, the awareness recognition rates in that condition were uniformly low. Since our primary goal in the current study was to examine the effect of adaptive personalization on the two measures of awareness, we chose only to include the large screen condition.

We modified the Control condition from Study 2 by adding 3 extra menu items at the top of the menu, in addition to the 24 regular items (see Figure 3). This made the Control menus the same length as the adaptive menus, eliminating menu length as a confound. The extra items were never selected in the

experimental tasks and they created a more conservative measure of awareness than using only a 24-item menu for Control: we hypothesized that Control would result in the highest awareness, but increasing the total number of words to which participants were exposed in that condition should negatively impact awareness-related measures.

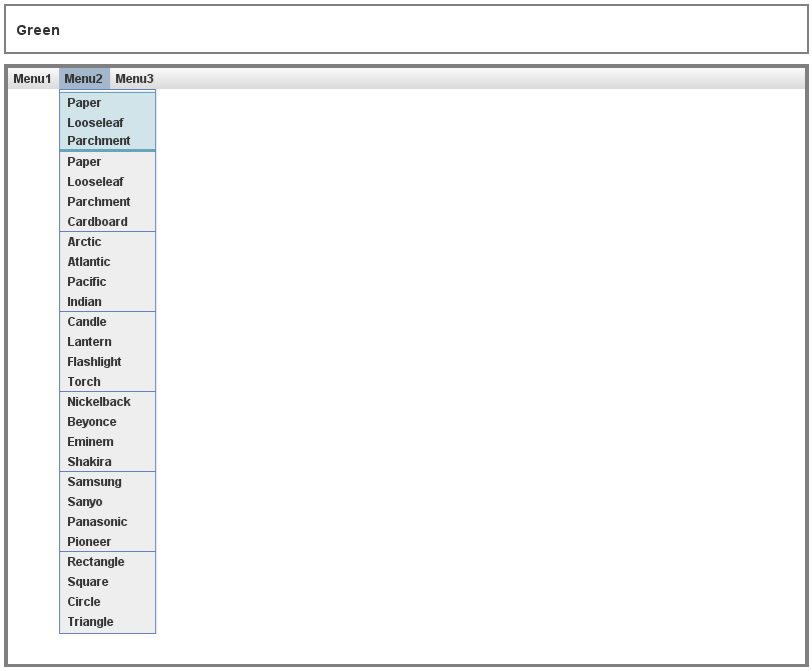

Figure 2. Experimental setup showing an adaptive menu.

Figure 3. Static menu with 3 extra items at top.

The adaptive algorithm's predictions were based on recently and frequently selected items. To achieve two different levels of adaptive accuracy, we followed the adaptive algorithm and two-step process used for Study 2 (more detail can be found in Findlater and McGrenere (2008)). First, for each participant we randomly generated a selection sequence (see Section 5.1.2) and applied the adaptive algorithm to predict

a set of the 3 items most likely to be needed next by the user; this algorithm resulted in prediction accuracy of 64% on average for all participants. Second, for Low accuracy we randomly adjusted 18 trials so that they were no longer correct, and for High accuracy we randomly adjusted the same number of incorrect predictions to be correct. This resulted in the accuracy conditions listed above.

As with Studies 1 and 2, we needed users to have no previous experience with the experimental interface in order to accurately measure how the different conditions impacted awareness. Since we also wanted to use a within-subjects design for increased statistical power, we chose to use a custom experimental interface rather than using a real application (unlike in Study 1, where we used Microsoft PowerPoint). This allowed us to create three interface layouts that were similar in every respect other than the personalization mechanism.

5.1.2 Task

The experimental task was a sequence of menu selections. A prompt at the top of the screen displayed the item to be selected by the user, but did not specify in which menu that item would be found (see Figure 2). The three menus were located just below the prompt. Once the participant had correctly selected the item, the prompt for the next trial would be shown.

The task was split into two blocks for each condition: a *training block* and a *testing block*. The purpose of the training block was to give participants experience with the menus, to develop a base level of awareness that we hypothesized would, in turn, impact performance when selecting new items in the testing block. The training block included selections of only 8 of the 24 items in each menu, whereas the testing block included an additional 4 items in each menu, to simulate an experienced user completing new tasks. The selection sequence for the training block was generated using a Zipf distribution (Zipfian $R^2 = .99$) over 8 randomly chosen items from each menu (i.e., within each menu, the selection frequencies of the 8 items were: 15, 8, 5, 4, 3, 3, 2, 2); this resulted in 126 selections and is the approach taken in Study 2 and used by Cockburn et al. (2007). The testing block was a randomly generated permutation of the exact same set of selections as the training block plus 2 additional selections for each new item, resulting in 150 selections.[1]

Selection sequences were randomly generated for each participant to mitigate the effect of a single sequence. The same underlying sequence for the training block and the testing block were used in each condition for a given participant, but to minimize learning effects the menu items were masked with different labels for each condition. These labels were randomly chosen in groups of 4 semantically related items (e.g., Chardonnay, Shiraz, Merlot, Cabernet) from a larger set of labels such that each label appeared only once for each participant. The labels for the extra three items in Control were generated similarly.

In Study 2, participants completed 252 menu selections before the recognition test. Based on feedback about fatigue from pilot participants and Study 2 participants, it was impractical to keep such a long training block when we were additionally asking users to complete a testing block. Instead, in Study 3 participants completed only half as many selections in the training block before the recognition test. The implication is that we would not expect as much of an impact on recognition test scores in Study 3 as in Study 2.

5.1.3 Design, Participants and Apparatus

The design was within-subjects, with a single factor: menu type (High, Low or Control). Order of presentation was fully counterbalanced and participants were randomly assigned to an order. Thirty participants (19 female) aged 19-56 (average 25 years) were recruited through on-campus advertising.

[1] Because of the additional items in the testing block task, the accuracy of the adaptive algorithm necessarily drops slightly.

Participants were students and community members who were regular computer users. Each participant was reimbursed $10 per hour to defray the costs of participation.

The experiment used a 2.0 GHz Pentium M laptop with 1.5 GB of RAM, with an 18" LCD monitor at 1280x1024 resolution and Microsoft Windows XP. The application was coded in Java 1.5 and it recorded all timing and error data.

5.1.4 Quantitative and Qualitative Measures

Our main measure was the indirect impact of awareness on new task performance, defined as the time to select items in the testing block that were not selected in the training block.

We also used a recognition test of unused items to more directly assess awareness (similar to Studies 1 and 2). It listed 12 randomly chosen items that were found in the menus for each condition, but were *not* selected in either the training or testing blocks. It also included 6 items randomly chosen from a set of distracter items; the full distracter 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., distracter for the group "soccer, basketball, baseball, football" was "rugby"). Valid and distracter items were chosen evenly across menus. For each item, subjects were asked to note if they definitely remembered it. From this, we calculate the *corrected recognition rate*: the percentage of valid targets correctly remembered minus the percentage of distracters incorrectly chosen. 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 (Baddeley, 1976).

We measured core task performance as the time to select those items in the testing block that had appeared in the training block. Time to select items in the training block was also used as a secondary measure of novice core task performance. Subjective feedback on each of the menu types was collected using six Likert scales. We were most interested in the first three of these scales, which measured awareness-related subjective responses: ease of learning the full set of menu items, ease of selecting infrequent items, and ease of remembering items that were not selected. The remaining Likert scales were on efficiency, difficulty, and satisfaction.

5.1.5 Procedure

The study procedure was designed to fit in a single 1.5 to 2 hour session. Participants first completed a background questionnaire. Then, for each menu condition, participants completed the training block, followed by the paper-based awareness recognition test, then the testing block. Short breaks were given between blocks and between conditions. We collected subjective feedback by questionnaire at the end of each condition and, for comparative comments, at the end of the session.

5.1.6 Hypotheses

Our main hypotheses were:

H1. *Impact of awareness on new task performance.* Control and Low will be faster than High (extension of the recognition test results from Study 2).

H2. *Core task performance*. High and Control will be faster than Low, but will be no different from each other (based on Study 2).

H3. *Perception of awareness.* Control and Low will be perceived to be easier than High for the three awareness-related subjective questions (following H1).

5.2 Results

A 3x6 (menu type x presentation order) repeated measures (RM) ANOVA showed no significant main or interaction effects of presentation order on the main dependent variable (new task performance), so we simplify our results by only examining effects of menu type. We ran a one-way RM ANOVA for each of the main dependent measures, using the same approach taken in Study 1. All pairwise comparisons were protected against Type I error using a Bonferroni adjustment. Along with statistical significance, we

report partial eta-squared (η^2), a measure of effect size. To interpret this value, .01 is a small effect size, .06 is medium, and .14 is large (Cohen, 1973).

One outlier was removed from the analysis for being more than 3 standard deviations away from the mean in one condition for new task performance. We report on results from 29 participants.

5.2.1 New Task Performance

Participants took on average 7.6 minutes to complete the testing block across conditions. As predicted, participants performed poorly with both of the personalized interfaces in comparison to the Control condition when asked to select new items in the testing block (Figure 4): menu type significantly impacted the speed of selecting new items ($F_{2,56} = 21.4$, $p < .001$, $\eta^2 = .433$). In High, participants took on average 3.7 seconds to select a new item, which was significantly longer than the average of 3.2 seconds for Low ($p = .002$) and the average of 2.9 seconds for Control ($p < .001$). Control was also faster than Low ($p = .011$). These results reflected our expectations that Control would allow participants to develop a better awareness of the full set of menu items and the location of those items in the interface.

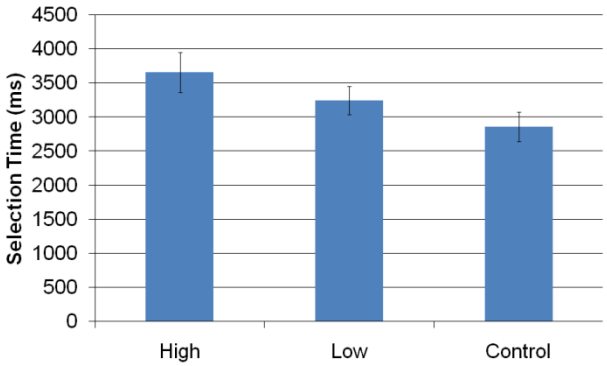

Figure 4. Performance impact of awareness, measured as speed of selecting new items in testing block; 95% confidence intervals shown. (*N* = 29)

5.2.2 Awareness Recognition Test

We also analyzed the paper-based awareness test scores. On average, test scores for each condition followed the same pattern as the new task performance results; that is, faster performance when selecting new items corresponded to higher scores here. Scores were 20.7%, 24.4% and 27.0% for High, Low, and Control, respectively. However, this did not translate to a significant main effect of menu type ($F_{2,56} = .988$, $p = .379$, $\eta^2 = .034$) as we found in Study 2. In retrospect, this is not entirely surprising given that we administered the test after participants had completed only half as many selections as in the previous study.

5.2.3 Core Task Performance

Shown in Figure 5, there was a main effect of menu type on core task performance of selecting items in the testing block that also appeared in the training block ($F_{2,56} = 58.9$, $p < .001$, $\eta^2 = .678$). High was faster for selecting old items than both Control and Low ($p < .001$ for both comparisons). Participants were faster in Control than Low at selecting old items ($p = .002$).

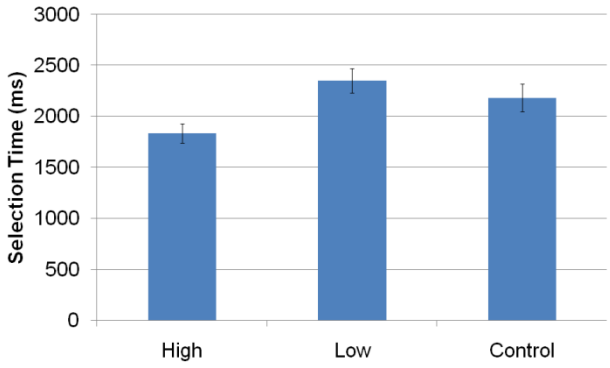

Figure 5. Experienced core-task performance; 95% confidence intervals shown. (*N* = 29)

Although it was not one of our main measures, we performed a secondary analysis on core task performance in the training block to assess inexperienced usage. As with the testing block, there was a significant main effect of menu type on speed of selections ($F_{2,56} = 30.7$, $p < .001$, $\eta^2 = .523$). High was faster than both Control ($p < .001$) and Low ($p < .001$). In comparison to the testing block, however, no difference was found between Control and Low in the training block.

These results differ from Study 2, where no difference was found on core task performance between the higher accuracy adaptive menus and the control condition, but both were faster than the lower accuracy adaptive menus. The difference between High and Control found in Study 3 is likely due to the additional extra items included in the Control menus.

5.2.4 Errors

We analyzed testing block error rates separately for newly introduced items and for old items (ones that had appeared in the training block). On average, the error rate was 1.9% across conditions for new items, and there was no significant effect of menu type on error rate ($F_{2,56} = .260$, $p = .772$, $\eta^2 = .009$). For the old items, however, there was a significant effect of menu type ($F_{2,56} = 4.23$, $p = .019$, $\eta^2 = .131$), but with a Bonferroni adjustment none of the pairwise comparisons were significant. For old items, error rates were 1.4%, 2.1% and 3.0% in High, Low, and Control, respectively.

5.2.5 Subjective Measures

We ran a Friedman test on each of the Likert scale questions and used Wilcoxon signed ranks tests with a Bonferroni adjustment for pairwise comparisons. One participant's questionnaire data was incomplete and is excluded from the analysis.

Subjective responses regarding learning mirrored the new task performance results. Significant differences were found for ease of learning the full set of menu items ($\chi^2_{(2,N=28)} = 9.08$, $p = .011$) and ease of finding infrequently selected items ($\chi^2_{(2,N=28)} = 12.7$, $p = .002$). Participants found that Control made it easier to learn the full set of menu items than Low ($p = .039$), and possibly High (trend: $p = .093$). It was also easier to select infrequent items with Control than with either Low ($p = .012$) or High ($p = .021$).

The ease of remembering items that were in the menus but that were not selected was also impacted by menu condition ($\chi^2_{(2,N=28)} = 6.50$, $p = .039$), but no pairwise comparisons were significant. A trend suggested that menu type may impact the perceived efficiency of finding items ($\chi^2_{(2,N=28)} = 5.31$, $p = .070$). No other significant differences were found.

5.2.6 Summary

We summarize our results according to our hypotheses:

H1. *Impact of awareness on new task performance: Supported.* Control and Low were faster than High when selecting new items in the testing block, showing an indirect impact of awareness on new task performance. Control was also faster than Low.

H2. *Core task performance: Partially supported.* High and Control were both faster than Low when selecting old items in the testing block, but, contrary to our hypothesis, High was also faster than Control.

H3. *Perception of awareness: Partially supported.* Participants found Control easiest for selecting infrequent items and a trend suggested this was also the case for learning the full set of menu items. However, the results for ease of remembering unused items were inconclusive, and Low was not found to be easier than High for any of the measures.

5.3 Discussion of Study 3

The findings from Study 3 show that the level of awareness gained from working in a personalized interface impacts the user's performance when completing new tasks; that is, different levels of awareness have the potential to impact future performance. The high accuracy adaptive split menus offered the best core task performance, but also resulted in the worst new task performance; subjective feedback also supported these findings. In comparison to the control condition, the low accuracy adaptive menus resulted in poor performance on both core and new tasks. This supports Study 2 results that show the low accuracy menus do not offer a viable alternative to traditional single-length pull-down menus for desktop-sized screens.

We had expected to find a significant impact of menu type on the recognition test scores in addition to the impact on new task performance. In Study 2, we found that high accuracy split menus resulted in lower awareness recognition test scores than both low accuracy adaptive menus and a static control. However, we did not find a significant difference here for the awareness recognition test. Overall, awareness test scores were lower than the previous study (on average, 24% here versus 31% in the large screen condition of the previous study). This is likely due to changes in the study design: we administered the recognition test after the training block, which is half the total time that participants spent in each condition before completing the recognition test in the previous study. The reduced length of exposure to each interface likely explains the lower recognition test scores and the lack of sensitivity of the measure. Another factor may be that the control condition had three extra items in each menu, which would have made it more difficult for participants to remember the full set of menu items in that condition. Descriptively, however, the pattern of mean scores is similar to that found previously, and mirrors the new task performance results. We expect that with a longer training block there would be both higher awareness test scores and statistically significant differences between the three conditions.

Achieving consistent awareness results on both the recognition test and new task performance measures was difficult in Studies 1 and 3. In Study 1, we found statistically significant differences on the recognition test but not on new task performance; Study 3 yielded the opposite result. This is likely due to the differing primary goals of the two studies and accompanying methodological choices. Study 1 was designed to provide a more realistic and cognitively demanding experience for participants, so was not optimized to isolate differences in new task performance. It did, however, allow for much longer exposure to the application (42 minutes on average), which may have resulted in more reliable awareness recognition test scores. Because of its more constrained task, Study 3 was better able to isolate differences in new task performance, but because participants only spent on average 8.5 minutes in each interface condition before completing the awareness recognition test, individual variability obscured possible statistical differences due to menu type. It would be useful to consider in the future whether a standardized memory test could be used as an appropriate covariate to account for some of this variation.

6. Revisiting the Design Space of Personalized GUIs in Light of Awareness

In our studies, we explored layered interfaces and adaptive split menus, two personalization approaches that offer contrasting points in the design space of personalized interfaces. However, many other

personalization approaches exist, and these approaches could impact the tradeoff between core task performance and awareness differently. In Section 2 we introduced several design factors that we consider to be particularly important for the interplay between core task performance and awareness. Previous evaluations of personalized interfaces have often measured core task performance (although it may not explicitly have been distinguished as such). However, since awareness has not been measured before, we revisit the design factors in the context of awareness, to incorporate our results into the discussion and to identify areas for future work.

6.1 Control of Personalization

We studied an adaptive mechanism (Studies 2 and 3) and an adaptable mechanism (Study 1), and found a measurable tradeoff between performance and awareness for both. We also saw that accuracy of an adaptive interface can impact core task performance (Study 2) and both the measures of awareness (Studies 2 and 3). Previous work has examined the impact of adaptive and adaptable personalization mechanisms on core task performance, as discussed in Section 2, but has not measured awareness.

The choice of adaptable versus adaptive mechanisms should impact awareness in at least two major respects. First, the cognitive overhead required for the user to adapt an interface, choosing which items to promote or demote, should result in a higher level of awareness of the full set of features than a comparable adaptive approach where this cognition is offloaded to the system. How long this effect lasts beyond the initial adaptation effort, however, would need to be explored. Second, although adaptive approaches to date have been designed with personalization accuracy as the main goal, they also have the potential to draw the user's attention to unused or infrequently used features. Recent recommender system research has begun exploring how recommendations that are not necessarily the most accurate may positively impact the user's satisfaction (Ziegler et al., 2006), a technique that could be explored for adaptive GUIs.

6.2 Granularity

In Section 2, we identified coarse-grained personalization, where large groups of features are manipulated at once, and fine-grained personalization, where features are manipulated individually. Although we studied both fine-grained and coarse-grained personalization (adaptive split menus and layered interfaces, respectively), we did not explicitly compare the two. Finer-grained approaches should allow for improved core task performance since they can be more accurately personalized to the user's needs at any given point in time. For example, a user may include in their personalized interface only the exact set of features they use in their daily work, rather than having to select from a more generalized grouping of features that a designer has deemed relevant to that user's type of work. However, coarse-grained approaches, if designed correctly, should be able to contribute to awareness by personalizing the interface to emphasize not only features known to the user, but related features as well.

6.3 Visibility of Change

The visual change in a personalized interface can take on many forms. Spatial techniques, for example, include hiding, moving, replicating, and resizing features. Of these options, hiding (completely removing) features deemed to be unnecessary strongly emphasizes the speed of selecting the remaining features. However, as seen in Study 1, this approach also negatively impacts awareness, even after transitioning to a more complex interface. Studies 2 and 3 demonstrated a similar tradeoff when features are replicated rather than hidden, at least when personalization accuracy is high.

In contrast to spatial adaptation techniques, marking techniques use visual cues (e.g., colour highlighting) to draw the user's attention to important features. Marking techniques should not have as much of a negative impact on awareness as spatial adaptation, since all features are as easily visible as in a traditional full interface. In Study 1, a trend suggested that the graphical marking technique we used (an 'x') may result in higher awareness than initially hiding features (the minimal layered condition), but we did not find a performance benefit for marking. Other studies have looked at different marking

techniques, such as colour highlighting (Tsandilas and schraefel, 2005) and temporal marking (Findlater et al., 2009), where adaptively predicted items appear briefly before the rest. These may provide more of a performance benefit than the graphical marking technique we used.

With techniques that hide features, the *direction of change* should also be considered. An approach like layered interfaces (Shneiderman, 2003) initially provides only a small, core set of features, adding more features as needed. As seen in Study 1, this improves core task performance, but also negatively impacts awareness. In contrast, an approach that initially provides the full set of features, then removes unnecessary or unused ones after a period of time, may allow the user to develop greater awareness.

6.4 Frequency of Change

Adapting the interface more frequently should theoretically allow it to more closely match the user's needs at a given point in time, improving core task performance. Yet, the lack of persistence with frequent adaptation may ultimately result in a negative impact on core task performance, as shown in several studies of adaptive personalization techniques that spatially reorganize the interface after every user interaction (e.g., Mitchell and Shneiderman, 1989; Findlater and McGrenere, 2004). Future work should explore how this factor can both positively and negatively impact awareness.

7. Design Implications

Based on the results from all three studies and the discussion in the previous section, we present several guidelines for personalized interfaces.

Look beyond accuracy as the ultimate goal of personalization. Our studies demonstrate the value of including both performance and awareness measures in evaluations of personalized interfaces. The personalized interfaces we studied offered better core task performance than a static control condition, but the tradeoff of this improved efficiency for selecting commonly used items is that users are less aware of the full set of features available in the application. Especially for adaptive approaches to personalization, where much of the focus has been on accuracy, designers need to broaden their focus to consider other aspects of the interaction, including awareness.

Identify the desirable balance between core task performance gains and awareness based on the application context. What is considered to be a desirable balance between core task performance and awareness may depend on different design contexts. High awareness of advanced features will be more important for software applications where users are expected to mature into experts, for example, as with a complex integrated development environment. On the other hand, for applications that are used on a less frequent basis (e.g., many websites) or for those applications that cater to a range of users with varying levels of expertise (e.g., ATMs), the need for efficient performance on core tasks may outweigh the need for awareness.

Match design characteristics to core task performance and awareness goals. We have identified four personalization design factors that are particularly important for performance and awareness (control, visibility, frequency, and granularity). Although more work is needed to map out the impact of all of these factors (and possibly identify further factors), we have provided a first step towards understanding their impact. Designers of personalized interfaces should incorporate design elements that support the particular goals of their system.

Use an appropriate awareness measure in evaluations. We presented two methods for measuring awareness and our experience demonstrates the advantages and disadvantages of each. For a more open-ended task or a field evaluation, the recognition test will be easier to administer because the only requirement is that users need to have had some experience with the interface before completing the recognition test. The performance impact on new task completion is more effortful to apply since it requires the design of an experimental task; however, if the evaluation is in a controlled setting and an appropriate constrained task can be devised, this measure will provide an indication of future performance.

Support exploratory behaviour and make de-emphasized features discoverable. Users often exhibit exploratory behaviour when learning an interface (Rieman, 1996), which can be inhibited by personalization. In Study 1 we saw that users explored more in the control condition than the marked layered condition, even though all features were visible in that condition. A trend also suggested that the control condition facilitated more exploration than the minimal condition. To support exploratory behaviour, especially in cases where features are hidden, users should have an easy means of viewing the full set of features. This should somewhat alleviate the user's concern over hiding features (as seen in Findlater et al., 2008; McGrenere and Moore, 2000).

Consider introducing new features to the user. There is the potential for adaptive or mixed-initiative systems to increase the user's awareness of features, by suggesting instances when the user may benefit from unused or underused features (e.g., Brusilovsky and Schwarz (1997) and Linton et al. (2000)). Adaptive suggestions have also been used to improve the overall efficiency of user-controlled personalization (Bunt et al., 2007). Very little work has been done on this type of mixed-initiative interaction, so it is a potentially fruitful area for further research. Our results do not generalize to personalization approaches that adaptively introduce features to the user, but they do offer motivation for the potential utility of such a mechanism.

Ultimately, the outcome of an individual design will depend on a number of the above factors and the interaction among them.

8. Limitations

Study 1 showed that a minimal layered interface impacts core task performance and awareness recognition test scores in comparison to a static control interface, but no support was found for our hypothesis that the layered interface would also impact performance on new tasks. With the goal of exploring multiple points in the design space, we purposely evaluated different personalization techniques in our studies instead of revisiting this hypothesis for layered interfaces (Study 1 looked at layered interfaces; Studies 2 and 3 looked at adaptive split menus). However, it will be important to revisit layered interfaces. We predict that a more controlled task that reduces individual variability will yield a statistically significant impact of the minimal layered interface on new task performance.

All three of our studies were conducted in a controlled laboratory setting, where users may value efficiency over longer-term learning. In contrast, in a more realistic setting when cognitive resources are divided among several, complex tasks and GUI feature selection is only part of any given task, users may value a personalization approach that facilitates awareness over one that emphasizes core task performance. A field study will be important for exploring the relationship between performance, awareness, and user satisfaction.

We focused on measuring core task performance and awareness, which we believe are particularly important for interface personalization, but a number of broader challenges need to also be considered when designing a personalization mechanism. For example, adaptive, adaptable, and mixed-initiative mechanisms offer different advantages. Adaptive mechanisms require little or no effort on the part of the user and do not require the user to have specialized knowledge to adapt the interface (Fischer, 2001), but have several issues related to lack of user control, unpredictability, transparency, privacy and trust (Höök, 2000).

Adaptable approaches, on the other hand, require effort and motivation on the part of the user to adapt the interface. Studies have also found that the extent to which people customize depends on their skill levels and interest (Mackay, 1990; MacLean et al., 1990). Our protocol in Study 1 did not have users interacting with the mechanism to reduce features since the experimenter set the interface layer for each task. Although the goal of adaptable personalization is to reduce complexity, the very inclusion of a mechanism to do so has the potential to make the system less usable, especially if these mechanisms are poorly designed (Kay, 2001). This impact needs to be outweighed by the beneficial effects of working within the personalized interface.

9. Conclusions

There is a strong tendency to add rather than eliminate features in new versions of software applications. The need for managing interface complexity is thus increasing, which underscores a major motivation behind personalization approaches. To provide a more nuanced evaluation of personalized interfaces, we previously introduced awareness as a measure to be included alongside traditional performance, and provided two operationalizations: (1) a recognition test of unused features and (2) user performance of completing new tasks.

GUI personalization research has largely focused on the benefit of personalization, including improved core task performance and reduced visual complexity. Through three controlled laboratory studies, our work reveals a more comprehensive understanding of the impact of personalization: we showed that personalization can negatively impact the user's overall awareness of features using both a layered interface (Study 1) and adaptive split menus (Study 2). In turn, personalization also impacts performance on completing new tasks (Study 3). Although personalization often offers a performance benefit for the user's core tasks, this negative impact on awareness indicates there may be a negative impact on future performance. Based on the study findings and a survey of related work, we also outlined a design space for personalized interfaces, identifying four factors that are likely to impact performance and awareness, and developed a set of design guidelines.

Our design space is not exhaustive, and other possible design factors and extensions should be explored in future work (e.g., our studies did not focus on granularity or frequency of change at all). In particular, since results with the marking interface in Study 1 were inconclusive, another approach to marking should be explored. In very recent work (Findlater et al., 2009), ephemeral adaptation has been shown to improve initial menu selection speed, so may be an improvement over our marked approach.

There is also undoubtedly a connection between awareness and learnability; personalization approaches may be more or less useful depending on whether the user is a learner or long-term user. There is a broad range of literature that investigates interfaces for learnability (e.g., Cox and Young (2001); Jordan et al. (1991)), and personalized or reduced-functionality versions of interfaces have been used in a learning context to provide support for novice users (e.g., Leutner (2000)). The issues of working in a personalized interface may be affected differently in these cases, however, since added direction is provided by a teacher or course material.

It will be important to generalize this research to other GUI interaction techniques, such as the Ribbon in Microsoft Office 2007. Study 1 included both menus and toolbars, but did not differentiate between the two. It is possible that textual control structures (menus) may result in higher awareness of available actions in the interface in comparison to visual control structures (toolbars), which may predominantly result in awareness about the *number* of features rather than the specific actions that can be achieved by those features. The Ribbon, which combines both text and icons, likely provides a different tradeoff between core task performance and awareness than is found with either menus or toolbars. It will be important to characterize the differences among these control structures, and to evaluate how they correspond to user satisfaction.

In the Intelligent User Interface community, some researchers have criticized intelligent systems of "dumbing down" the user when a portion of the user's cognitive load is offloaded to the system (Lanier, 1995). Intelligent systems can reduce the user's breadth of experience by reducing opportunities for learning in that domain (Jameson, 2008). Recent research in recommender systems has introduced the notion of topic diversification to improve the user's experience, in contrast to more narrow definitions of accuracy for recommendation lists (Ziegler et al., 2005). Although differences exist between GUI personalization and content personalization (Bunt, 2007), such as for web pages, it will also be interesting to explore parallels in terms of the accuracy of personalization versus the user's breadth of experience.

Finally, although the three studies reported in this paper offered different degrees of ecological validity, a longitudinal field study will be important for assessing how the set of features known to the user is affected over much longer-term by working in a personalized interface.

Acknowledgment

This research was funded by IBM Centers for Advanced Studies and NSERC. We also thank Jessica Dawson for her invaluable help in running this study, and Karyn Moffatt for comments on a draft of this paper.

References

1. Baddeley, A., 1976. The Psychology of Memory. Basic Books, New York.
2. Bannert, M., 2000. The effects of training wheels and self-learning materials in software training. Journal of Computer Assisted Learning 16(4), 336-346.
3. Brusilovsky, P., Schwarz, E., 1997. User as student: Towards an adaptive interface for advanced web-based applications. In: Proceedings of the Sixth International Conference on User Modeling, 177-188.
4. Bunt, A., Conati, C., McGrenere, J., 2007. Supporting interface customization using a mixed-initiative approach. In: Proceedings of Intelligent User Interfaces, 92-101.
5. Bunt, A., 2007. Mixed-Initiative Support for Customizing Graphical User Interfaces [dissertation]. University of British Columbia.
6. Carroll, J.M., Carrithers, C., 1984. Training wheels in a user interface. Communications of the ACM 27(8), 800-806.
7. Catrambone, R., Carroll, J.M., 1987. Learning a word processing system with training wheels and guided exploration. SIGCHI Bulletin 18(4), 169–174.
8. Clark, B., Matthews, J., 2005. Deciding Layers: Adaptive Composition of Layers in a Multi-Layer User Interface. In: Proceedings of HCI International.
9. Cockburn, A., Gutwin, C., Greenberg, S., 2007. A predictive model of menu performance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 627-636.
10. Cohen, J., 1973. Eta-squared and partial eta-squared in communication science. Human Communication Research 28, 473-490.
11. Cox, A.L., Young, R.M., 2001. Device-oriented and task-oriented exploratory learning of interactive devices. In: Proceedings of the 3rd International Conference on Cognitive Modeling, 70-77.
12. Debevc, M., Meyer, B., Donlagic, D., Svecko, R., 1994. Design and evaluation of an adaptive icon toolbar. User modeling and user adapted interaction 6(1), 1-21.
13. Eisenberg, M., Fischer, G., 1994. Programmable design environments: Integrating end-user programming with domain-oriented assistance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 431–437.
14. Findlater, L. McGrenere, J., 2004. A comparison of static, adaptive and adaptable menus. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 89-96.
15. Findlater, L., McGrenere, J., 2007. Evaluating reduced-functionality interfaces according to feature findability and awareness. In: Proceedings of IFIP Interact, 592-605.
16. Findlater, L., McGrenere, J., 2008. Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1247-1256.
17. Findlater, L., McGrenere, J., Modjeska, D., 2008. Evaluation of a role-based approach for customizing a complex development environment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems , 1267-1270.
18. Findlater, L., Moffatt, K., McGrenere, J., Dawson, J., 2009. Ephemeral adaptation: The use of gradual onset to improve menu selection performance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. In press.
19. Fischer, G., 2001. User modeling in human-computer interaction. User Modeling and User-Adapted Interaction 11(1-2), 65–86.

20. Gajos, K.Z., Christianson, D., Hoffmann, R., Shaked, R., Henning, K., Long, J.J., Weld, D.S., 2005. Fast and robust interface generation for ubiquitous applications. In: Proceedings of UBICOMP'05, 37-55.
21. Gajos, K.Z., Czerwinski, M., Tan, D.S., Weld, D.S., 2006. Exploring the design space for adaptive graphical user interfaces. In: Proceedings of the Working Conference on Advanced Visual Interfaces (AVI'06), 201-208.
22. Gajos, K.Z., Everitt, K., Tan, D. S., Czerwinski, M., Weld, D. S., 2008. Predictability and accuracy in adaptive user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08), 1271-1274.
23. Gajos, K.Z., Wobbrock, J.O., Weld, D.S., 2008. Improving the performance of motor-impaired users with automatically-generated, ability-based interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1257-1266.
24. Greenberg, S., Witten, I., 1985. Adaptive personalized interfaces: A question of viability. Behaviour and Information Technology 4(1), 31–45.
25. Greenberg, S., 1991. Personalizable groupware: Accommodating individual roles and group differences. In: Proceedings of the European Conference on Computer-Supported Cooperative Work. (ECSCW'91), 24–27.
26. Grossman, T., Fitzmaurice, G., Attar, R., 2009. A survey of software learnability: metrics, methodologies and guidelines. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09), 649-658.
27. Gustavsson Christiernin, L., Lindahl, F., Torgersson, O., 2004. Designing a multi-layered image viewer. In: Proceedings of the Third Nordic Conference on Human-Computer Interaction (NordiCHI'04), 181-184.
28. Herlocker, J. L., Konstan, J. A., Terveen, L. G., Riedl, J. T., 2004. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1), 5-53.
29. Höök, K., 2000. Steps to take before intelligent user interfaces become real. Journal of Interacting with Computers 12(4), 409-426.
30. Jameson, A., 2008. Adaptive interfaces and agents. In: Sears, A., Jacko, J., editors. Human-Computer Interaction Handbook. 2nd edition. New York: Lawrence Erlbaum.
31. Jordan, P.W., Draper, S.W., MacFarlane, K.K., McNulty, S., 1991. Guessability, learnability, and experienced user performance. In: Proceedings of HCI '91 People and Computers VI: Usability Now!, 237-245.
32. Kay, J., 2001. Learner control. User Modeling and User-Adapted Interaction 11, 111-127.
33. Kersten, M., Murphy, G. C., 2005. Mylar: A degree-of-interest model for IDEs. In: Proceedings of the 4th International Conference on Aspect-Oriented Software Development (AOSD'05), 159-168.
34. Lanier, J., 1995. Agents of alienation. interactions 2(3), 66-72.
35. Leutner, D., 2000. Double-fading support - a training approach to complex software systems. Journal of Computer Assisted Learning 16(1), 347-357.
36. Linton, F., Joy, D., Schaefer, H.-P., Charron, A., 2000. Owl: A recommender system for organization-wide learning. Educational Technology & Society 3(1), 62-76.
37. Mackay, W. E., 1990. Patterns of sharing customizable software. In: Proceedings of the 1990 ACM Conference on Computer-Supported Cooperative Work, 209-221.
38. MacLean, A., Carter, K., Lövstrand, L., Moran, T., 1990. User-tailorable systems: Pressing the issues with buttons. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 175-182.
39. McGrenere, J., Moore, G., 2000. Are we all in the same "bloat"? In: Proceedings of Graphics Interface (GI 2000), 187-196.
40. McGrenere, J., Baecker, R., Booth, K., 2002. An evaluation of a multiple interface design solution for bloated software. CHI Letters 4(1), 163-170.
41. Mitchell, J., Shneiderman, B., 1989. Dynamic versus static menus: An exploratory comparison. SIGCHI Bulletin 20(4), 33-37.

42. Plaisant, C., Kang, H., Shneiderman, B., 2003. Helping users get started with visual interfaces: Multi-layered interfaces, integrated initial guidance and video demonstrations. In: Proceedings of HCI International.
43. Rieman, J., 1996. A field study of exploratory learning strategies. ACM Transactions on Human-Computer Interaction 3(3), 189-218.
44. Shneiderman, B., 2003. Promoting universal usability with multi-layer interface design. In: Proceedings of the 2003 Conference on Universal Usability, 1-8.
45. Stamoulis, D., Kanellis, P., Martakos, D., 2001. Tailorable information systems: Resolving the deadlock of changing user requirements. Journal of Applied System Studies 2(2).
46. Stuerzlinger, W., Chapuis, O., Phillips, D., Roussel, D., 2006. User interface facades : Towards fully adaptable user interfaces. In: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST'06), 309-318.
47. Terveen, L., McMackin, J., Amento, B., Hill, W., 2002. Specifying preferences based on user history. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 315-322.
48. Tsandilas, T., schraefel, m.c., 2005. An empirical assessment of adaptation techniques. In: Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems, 2009-2012.
49. Ziegler, C-N., McNee, S.M., Konstan, J.A., Lausen, G., 2005. Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web, 22-32.