Predictability and Accuracy in Adaptive User Interfaces

Krzysztof Z. Gajos[1], Katherine Everitt[1], Desney S. Tan[2], Mary Czerwinski[2], Daniel S. Weld[1]
[1]University of Washington
Seattle, WA 98195
{kgajos,everitt,weld}@cs.washington.edu
[2]Microsoft Research
One Microsoft Way, Redmond, WA 98052
{desney,marycz}@cs.washington.edu

ABSTRACT
While proponents of adaptive user interfaces tout potential performance gains, critics argue that adaptation's unpredictability may disorient users, causing more harm than good. We present a study that examines the relative effects of predictability and accuracy on the usability of adaptive UIs. Our results show that increasing predictability and accuracy led to strongly improved satisfaction. Increasing accuracy also resulted in improved performance and higher utilization of the adaptive interface. Contrary to our expectations, improvement in accuracy had a stronger effect on performance, utilization and some satisfaction ratings than the improvement in predictability.

AUTHOR KEYWORDS
Adaptive interfaces, predictability, accuracy, user study

ACM Classification Keywords
H5.2 Information Interfaces and Presentation: User Interfaces – *Interactions Styles, Evaluation/Methodology*

INTRODUCTION

Despite considerable debate, automatic adaptation of user interfaces (UIs) remains a contentious area. Proponents of machine learning-directed adaptation (e.g., [1,5]) argue that it offers the potential to optimize interactions for a user's tasks and style. Critics (e.g., [2,10]), on the other hand, maintain that the inherent unpredictability of adaptive interfaces may disorient the user, causing more harm than good.

Fortunately, recent studies have presented suggestions for which properties of adaptive UIs increase user confusion and which improve satisfaction and performance [3,4,11]. But the design space for adaptive UIs is large, with a multitude of characteristics that may determine an adaptive interface's success or failure. The tradeoffs between many of these characteristics are still poorly understood.

In this paper we explore the relative effects of predictability and accuracy on the usability of adaptive interfaces. We say that an adaptive algorithm is *predictable* if it follows a strategy users can easily model in their heads. We use the term *accuracy* to refer to the percentage of time that the necessary UI elements are contained in the adaptive area (see Task section). We focus on these properties because they reflect a common design trade-off in adaptive UIs: whether to use a simple, easily-understood strategy to promote functionality, or whether to rely on a potentially more accurate but also more opaque machine learning approach.

We present a study showing that increased accuracy significantly improved both performance and adaptive interface utilization. Furthermore, both predictability and accuracy significantly increased participants' satisfaction. Contrary to our expectations, we found that in our particular design, increasing the adaptive algorithm's accuracy had more beneficial effects on the participants' satisfaction, performance and utilization of the adaptive interface than did improved predictability. Our results suggest that machine-learning algorithms deserve further consideration in the context of adaptive UIs, because the benefits of a large improvement in accuracy may outweigh the disadvantages of decreased predictability.

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

EXPERIMENT

Hypotheses

Building on previous research, we hypothesized: (1) the higher the accuracy of the adaptive algorithm, the better the task performance, utilization and the satisfaction ratings; (2) the more predictable the adaptive algorithm, the better the task performance, utilization and the satisfaction ratings; (3) increased predictability would have a greater effect on satisfaction and utilization than increased accuracy. We formulated this last hypothesis based on the design heuristic asserting that successful user interfaces should be *easy to learn* [6].

Participants

Twenty-three volunteers (10 female) aged 21 to 44 (M=35 years) participated in this study. All participants had normal vision, moderate to high experience using computers and were intermediate to expert users of Microsoft Office-style applications, as indicated through a simple screener. Participants were given a software gratuity for their time.

Task

In order to explore the effects of accuracy and predictability, we used a generalization of the Split Menu concept [8], which [4] termed a *split interface*. In a split interface, func-

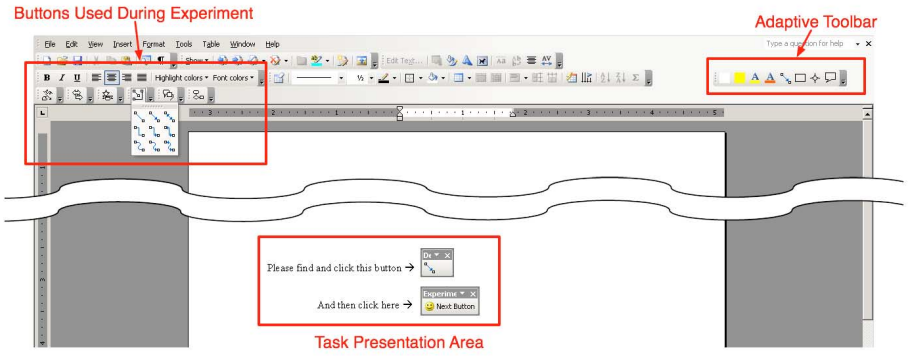

Figure 1. Task Setup.

tionality (e.g., buttons) that is predicted to be immediately useful is *copied* to an adaptive area (clearly designated as hosting changing content). This allows the user to either follow their familiar route or potentially save time by exploiting the adaptation.

We used a carefully controlled performance task that allowed us to eliminate complications associated with more complex tasks. Our task and procedures are modeled on those used in [4]'s second experiment. In the study, we showed participants a picture of a target UI button, which they had to find and click within the interface. They could use either the traditional (static) toolbars and menus or the adaptive toolbar if the button was displayed upon it.

We used a modified Microsoft Word interface, where several toolbars were placed toward the left and the center of the toolbar area (Figure 1). Some of the toolbar buttons revealed popup menus, in which additional functions were located. In the center of the screen was a panel that displayed the target button as well as a "Next" button, which participants used to proceed through the study. We only asked participants to find buttons originating in popup menus, which were one level deep.

The adaptive toolbar, whose contents changed during the experiment, was located in the upper right so that it was far enough from the closest, relevant, non-adaptive button (>20° visual angle). This ensured that it required an explicit change of gaze to discover if a helpful adaptation had taken place. Eight buttons were always shown in the adaptive toolbar, and no more than one button was replaced per interaction. Participants clicked on 60 target buttons in each task set. We considered the first 10 clicks to be a "ramp-up" time and did not include performance metrics for these clicks in our analysis.

Design and Procedure

Our study considered two accuracy levels: 50% and 70%. Because it is difficult to implicitly measure predictability, and measuring it explicitly might influence user performance [7], we considered two extreme cases. In the unpredictable condition, updates to the adaptive toolbar were entirely random – a worst-case simulation of a complex algorithm's inscrutability. In the predictable condition, we chose a most recently used (MRU) strategy, placing the eight MRU buttons in the toolbar. Post-experimental interviews confirmed that our participants easily formed a mental model of this MRU policy. Hence, the study was a 2 (accuracy: 50% or 70%) × 2 (predictability: *High* (most recently used) or *Low* (random)) factorial design.

We predetermined the sequence of buttons that had to be pressed in all conditions as well as the contents of the adaptive toolbar in the random condition to ensure the desired level of accuracy.

After familiarizing themselves with the task and completing a practice set using a non-adaptive interface, participants performed four counterbalanced task sets, one for each of the four conditions. Participants filled out a brief satisfaction survey after each task set, and an additional survey and an exit interview following the last session. Participants took 2.5 to 5 minutes per task set, and the whole study took less than an hour.

Equipment

We ran participants in pairs on two 2.8 GHz Pentium 4 Compaq PCs with 2G of RAM, each with a Compaq keyboard and an optical mouse. Each computer had two NEC 18″ color LCD displays but only one display per computer was actively used by the participants. Participants did not interact with each other.

Measures

We collected overall task times as well as the median times to acquire individual buttons (i.e., the time from selecting the "Next Button" to clicking on the indicated target), distinguishing times for buttons in their original locations from those located on the adaptive toolbar. We also measured the adaptive toolbar utilization levels, or the number of times that the participant selected the requested UI element from the adaptive toolbar divided by the number of times that the requested element was present on the adaptive toolbar. Additionally, we collected the subjective accuracy of the adaptive algorithm, and participant satisfaction ratings (on a 7-point Likert scale). Finally, we asked a random subset of participants to perform an extra set of tasks following the main experiment; here we used an eye-tracker to determine which strategies our participants employed. Performance considerations prevented us from using the eye tracker during the main part of the experiment.

Results

We analyzed all continuous measures using a 2 (50% or 70% accuracy) × 2 (High *vs*. Low predictability) repeated measures analysis of variance (RM-ANOVA). For the analysis, we took the logarithm of all timing data - standard practice to control for non-normal distributions found in such data. Because we cannot justify any assumptions about

		Individual conditions				Averaged over accuracy settings			Averaged over predictability settings		
		Random, 50%	Random, 70%	Predictable, 50%	Predictable, 70%	50%	70%	significant?	Random	Predictable	significant?
Duration		196	176	199	177	197	177	*	186	188	
Utilization		69%	89%	73%	84%	71%	86%	*	79%	78%	
Satisfaction ratings	Useful	3.36	4.77	3.73	4.77	3.55	4.77	*	4.07	4.25	
	Predictable	2.41	3.00	3.82	4.43	3.11	3.71		2.70	4.12	*
	Knew†	1.95	2.64	3.24	4.24	2.60	3.44		2.30	3.74	*
	Frustrating	4.23	2.55	3.50	2.73	3.86	2.64	*	3.39	3.11	
	Confusing	4.14	3.36	3.24	2.95	3.69	3.16		3.75	3.09	
	Satisfied	3.86	4.64	4.41	5.05	4.14	4.84		4.25	4.73	
	In Control	3.19	4.27	4.41	5.05	3.80	4.66	*	3.73	4.73	*
	Efficient	2.59	3.95	3.18	4.59	2.89	4.27	*	3.27	3.89	

†Knew = "I knew when the extra toolbar would have what I needed"

Table 1. Summary of the results. Times are in seconds and satisfaction ratings are on a 7-point Likert scale.

the distribution underlying Likert scale subjective responses, and because three participants omitted answers to some of the questions, we used ordinal logistic regression [12] (a non-parametric test, which can accommodate missing data) to analyze those data. Subjective data from one participant were lost due to a software error. Table 1 summarizes the results.

Perception of Predictability

In the free response part of the post-task questionnaire, 11 out of 23 participants spontaneously commented, after at least one of the two random conditions, that the toolbar behavior was "random," "confusing" or otherwise unpredictable. In contrast, after the predictable conditions only two participants commented that they did not understand the adaptive toolbar's behavior, while three specifically observed that it behaved more predictably than in earlier conditions.

Similarly, when debriefed after the study, the majority of participants correctly described the algorithm in the predictable condition as selecting the most recently used items, while a few felt that the system selected the most frequently used items. Participants often described the algorithm in the random condition as behaving in an apparently random manner though they often assumed that the behavior was purposeful (even if inscrutable) and that the algorithm was trying to "guess" or "help".

Satisfaction

We observed main effects of predictability ($\chi^2_{1,N=1049}$=17.22, p<.0001) and accuracy ($\chi^2_{1,N=1049}$=34.59, p<.0001) on the combined satisfaction ratings. We further analyzed each of the 8 responses separately. To preclude any effects that might have arisen purely by chance, we applied the Bonferroni correction [9] and considered as significant only those effects where p≤.05/8 = .00625.

This further analysis showed that participants' feeling of being in control increased both with improved predictability ($\chi^2_{1,N=87}$=11.69, p<.001) and with improved accuracy ($\chi^2_{1,N=87}$=9.70, p<.002).

In predictable (*vs*. random) conditions participants felt that the adaptive interface behaved more predictably ($\chi^2_{1,N=87}$=17.03, p<.0001), and that they knew better when the adaptive toolbar would contain the needed functionality ($\chi^2_{1,N=86}$=15.03, p=.0001).

In conditions with higher accuracy, participants felt that the adaptive interface was more useful ($\chi^2_{1,N=88}$=14.26, p<.001), less frustrating ($\chi^2_{1,N=88}$=26.47, p<.0001), and that it improved their efficiency ($\chi^2_{1,N=88}$=12.56, p<.001).

Unsurprisingly, a separate 2×2 RM-ANOVA revealed that participants perceived a difference in accuracy between the two accuracy levels ($F_{1,19}$=34.906, p<.001), estimating the lower accuracy condition to be 50.0% and the higher to be 69.0% accurate, on average.

Utilization

A 2×2 RM-ANOVA showed a main effect of accuracy on adaptive toolbar utilization ($F_{1,22}$=11.420, p<.01). At the 50% accuracy level the adaptive toolbar was used 70.6% of the time when it contained the correct button, compared with 86.4% at the 70% accuracy level.

Performance

We found that increased accuracy improved task completion times ($F_{1,18}$=62.038, p<.001) and, in particular, the median time to access buttons located on the adaptive toolbar (from 1.86s to 1.70s, $F_{1,18}$=18.081, p<.001) but not those located in the static part of the interface. No significant effects were observed for the algorithm's predictability.

Relative Impacts of Predictability and Accuracy

Treating the condition with low predictability and 50% accuracy as a baseline, we investigated which change would more greatly impact the participants: raising predictability or improving the accuracy to 70%. Thus, we compared two conditions: predictable but only 50% accurate versus random but 70% accurate.

An ordinal logistic regression analysis (with Bonferroni correction) of the satisfaction responses showed that the participants felt that the toolbar in the predictable but 50% accurate condition was more predictable ($\chi^2_{1,N=44}$=9.14, p<.003), while the adaptation in the random but more accurate condition was more useful ($\chi^2_{1,N=44}$=11.93, p<.001), less frustrating ($\chi^2_{1,N=44}$=19.07, p<.0001), and better improved their efficiency ($\chi^2_{1,N=44}$=14.80, p<.0001).

Increased accuracy also resulted in significantly shorter task completion times (RM-ANOVA, $F_{1,22}$=26.771, p<.001) and higher adaptive toolbar utilization ($F_{1,22}$=5.323, p<.05) than improved predictability.

Eye Tracking

In analyzing the eye tracking data (22 task sets performed by 16 participants; each condition was repeated 5 or 6 times), we identified three regions of interest (ROIs): the static buttons on the top left, the adaptive toolbar at the top right, and the task presentation area in the center of the screen (see Figure 1). The small sample collected led to low statistical power, but the data shed some light on the approaches used by the participants.

We looked at transitions between the ROIs to see if the participants were more likely to look at the adaptive toolbar or the static toolbar after being presented with the next button to click. We found that users moved their gaze from the task presentation area to the adaptive toolbar (rather than to the static part of the interface on the left) much less in the low accuracy condition than the high one (66% *vs*. 79%, respectively). The difference between predictable and random conditions did not elicit similar difference in behavior.

We also looked at the percentage of times participants first looked at the adaptive toolbar, failed to find the desired functionality there, and then shifted their gaze to the static toolbar. We found that participants looked but could not find the appropriate button on the adaptive toolbar much more often in the random than the predictable conditions (41% vs. 34%, respectively). Participants seemed to be performing better than the expected 40% failure rate (averaging over the two accuracy levels) in the predictable condition, suggesting that the more predictable algorithm did help the participants to best direct their effort.

Other User Comments

Besides commenting on the predictability of the adaptive toolbar behavior, 10 of the 23 participants commented that they wished the adaptive toolbar were closer to the original locations of the buttons used in order to aid opportunistic discovery of adaptation. Similar comments were also reported by [4] – their *moving interface*, although not statistically better than the non-adaptive baseline, was frequently praised by participants for placing adapted functionality right next to the original location. We chose [4]'s split interface for this study, because it *was* statistically better than the baseline, but a hybrid approach might be even better.

CONCLUSIONS

We have examined the influence of accuracy and predictability on adaptive toolbar user interfaces. Results show that both predictability and accuracy affect participants' satisfaction but only accuracy had a significant effect on user performance or utilization of the adaptive interface. Contrary to our expectations, improvement in accuracy had a stronger effect on performance, utilization and some satisfaction ratings than the improvement in predictability. Our results suggest that even though machine learning algorithms may produce inscrutable behavior, in certain cases they may have the potential to improve user satisfaction. Specifically, if a machine-learning algorithm can more accurately predict a user's next action or parameter value, then it may outperform a more predictable method of selecting adaptive buttons or default values. However, because predictability and accuracy affect different aspects of users' satisfaction, improvements to one of these factors cannot fully offset the losses to the other.

Our contribution is initial evidence showing the relative impact of these two dimensions on adaptive UIs. We believe that much future work remains in moving beyond laboratory studies and into the field with users' applications and projects, as well as understanding the crossover points in the tradeoff between improved accuracy and reduced predictability.

Acknowledgments This work was funded in part by the Microsoft Graduate Research Fellowship, NSF grant IIS-0307906, ONR grant N00014-06-1-0147, DARPA project CALO through SRI grant number 03-000225, and the WRF /TJ Cable Professorship.

REFERENCES

[1] Benyon, D. (1993) Adaptive systems: A solution to usability problems. *User Modeling and User-Adapted Interaction 3* (1), 65-87.

[2] Findlater, L. and McGrenere, J. (2004) A comparison of static, adaptive, and adaptable menus. *Proc. CHI'04*, 89-96. ACM Press.

[3] Findlater, L. and McGrenere, J. (2008) Impact of Screen Size on Performance, Awareness, and User Satisfaction With Adaptive Graphical User Interfaces. *Proc. CHI'08*, ACM Press.

[4] Gajos, K., Czerwinski, M., Tan, D. and Weld, D. (2006) Exploring the design space for adaptive graphical user interfaces. *Proc. AVI'06*, 201-208. ACM Press.

[5] Maes, P. (1994) Agents that reduce work and information overload. *Commun. ACM 37* (7), 30-40.

[6] Nielsen, J. (1994) Enhancing the explanatory power of usability heuristics. *Proc. CHI'94*, 152-158. ACM Press.

[7] Paymans, T., Lindenberg, J. and Neerincx, M. (2004) Usability trade-offs for adaptive user interfaces: ease of use and learnability. *Proc. IUI'04*, 301-303.

[8] Sears, A. and Shneiderman, B. (1994) Split menus: effectively using selection frequency to organize menus. *ACM Trans. Comp.-Hum. Interact. 1* (1), 27-51

[9] Shaffer, J. (1995) Multiple Hypothesis-Testing. *Annual Review of Psychology 46*, 561-584.

[10] Shneiderman, B. (1997) Direct manipulation for comprehensible, predictable and controllable user interfaces. *Proc. IUI'97*, 33-39. ACM Press.

[11] Tsandilas, T. and Schraefel. (2005) An empirical assessment of adaptation techniques. *CHI '05 extended abstracts*, 2009-2012.

[12] Winship, C. and Mare, R. (1984) Regression Models with Ordinal Variables. *Am. Soc. Rev. 49* (4), 512-525.