Assessing Interactive System Effectiveness with Usability Heuristics and Markov Models of User Behavior

Lashanda Lee & David B. Kaber
Edwards P. Fitts Department of Industrial & Systems Engineering
North Carolina State University, Raleigh, NC 27695, USA

The purpose of this research was to create a new measure of usability to aid in determining whether the intent of system designers is realized and to objectively assess user behavior as a basis for interface design recommendations. There are currently few quantitative, objective measures of usability among many subjective measures. A measure combining both objective and subjective data may be valuable in determining user perceptions of system designs and whether performance goals are achieved. The present research used a human computer interaction framework, representing the user, input, the system and output, as a basis for designer and user interface evaluation forms. These forms were used in combination with a Markov model of average number of user actions in task performance in order to determine an overall system usability effectiveness score. The methodology was applied to the evaluation of the design of a web-based ordering interface. The new measurement technique appeared to be sensitive to interface design changes and may be more effective than subjective methods (alone) for determining if interfaces meet design and performance objectives.

INTRODUCTION

Humans and computers are complex biological and mechanical systems that differ in terms of methods of communication and the ways in which they perform tasks. For human computer interaction (HCI) to be successful, interfaces must be designed to effectively translate user intentions, actions and inputs for the computer and effectively translate machine outputs for human comprehension. However, from a usability engineering perspective, there are few quantitative measures available as bases for interface design and, furthermore, few measures integrating subjective and objective responses. In recent years, several comprehensive interactive system design frameworks have been developed, such as Abowd and Beale (1991; see Figure 1) and Norman's model of HCI (1988), to guide HCI application development. Such frameworks can be useful for classifying usability problems and identifying aspects of system designs that are considered critical to usability and effectiveness. However, for the most part, the dimensions of these frameworks have not been translated into operational measures of usability. Quantitative data is also needed as a basis for critical usability feature identification and cost justification of interface designs. Some quantitative human behavior modeling methods (e.g., GOMS) have been developed to reduce the need for user experimentation but they have not been integrated with subjective measures.

In order to facilitate robust usability analyses, a simple framework like Abowd and Beale, can be used to identify critical design dimensions and to set subjective design criteria for the usability analyst. Operations Research techniques have been previously used to provide quantitative information as a basis for HCI design problems. By combining these approaches, the usability analyst may be able to qualitatively and quantitatively compare various design alternatives along several dimensions of usability (e.g. perceived importance of usability issues for designers and number of interface actions ultimately required of users) and create meaningful outputs for a system's design process.

LITERATURE REVIEW

HCI frameworks

There are several models of HCI (e.g., Norman, 1988) that attempt to describe the methods of communication between a user and computer. The purpose of modeling this interaction is to provide a basis for explaining how and why communication may fail, leading to errors. Abowd and Beale's interaction framework (Figure 1) and the Pipe-line information processing model (Mulhem and Nigay, 1996) are other examples. For this research, the Abowd and Beale framework was chosen as a basis to develop a new measure of usability, as it provides a complete representation of the interaction between the human and the computer. It defines a continual cycle of communication including gulfs of articulation and presentation for both the human and computer.

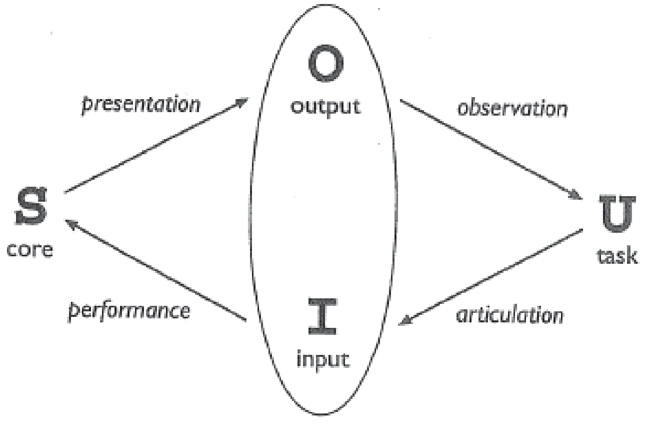


Figure 1: Abowd and Beale HCI framework

Measures of Usability

Measures of usability can be generally grouped into two categories, quantitative and qualitative. Each category tends to be used more in certain interface design or evaluation phases and each has pros and cons for assessing usability. Qualitative measures, such as inspection methods, heuristic evaluations and questionnaires, make comparisons of interface designs based on qualities that cannot be quantified and statistical analyses typically cannot be conducted. The benefits of these methods include low cost and quick discovery of design problems. They can help eliminate usability problems before user testing begins and reduce analysis and reporting time. Inspection methods are also considered to be low skill methods. Research has shown heuristic evaluations with non-experts can correctly identify problems; however, problems that are missed tend to be severe (Cockton et al., 2002; Molich and Nielsen, 1990).

Quantitative measures are used to make comparisons of designs based on physical quantities associated with interface features. Researchers often use quantitative measures during iterative design processes. The most common quantitative measures include number of correct tasks completed and task performance accuracy, task completion time and degree of variation from an optimal solution (Hornbaek, 2005). All of these measures provide a count of concrete occurrences during interface testing and are not based on user opinions. Others include measures of satisfaction, such as rankings or ratings of interfaces based on ease of use. These measures are typically collected through surveys after system use. Hornbaek (2005) found that in 93% of 180 usability studies, surveys were used to collect satisfaction data. Because surveys are collected after the fact and questions can be misunderstood by users, it may be difficult to link specific survey items to parts of an interface. Measures of satisfaction can be quantified and presented as graphs and charts; however, these measures are subjective and rely on the responses of users instead of observed data. Consequently, results may not correspond with objective quantitative data from observational studies.

User modeling is another form of quantitative usability assessment. Empirical data can be collected and used to develop a model to predict, for example, user interface action sequences. A common user modeling technique that has been researched over the last 25 years is GOMS modeling. It is used to predict and explain real-world task performance with interactive systems. A GOMS model shows how a user interacts with a system and how he or she performs tasks based on the hierarchical structure of Goals, Operators, Methods and Selection Rules (Card et al., 1983).

Quantitative measures and models of user performance can provide meaningful information to companies for design. Statistical analyses can be prepared along with charts and graphs to present usability evaluation results. Company management can determine the impact of usability issues in the design of a system from such outcomes. Quantitative measures are however, usually recorded through user testing, which can be expensive.

In general, quantitative measures may be better than qualitative measures for detailed usability problem analysis and formulating design recommendations. However, qualitative measures may be helpful in eliminating some problems before user testing begins. Since quantitative measures are often necessary to gain support from management for addressing usability problems in design, an integrated usability measurement method combining user modeling with subjective usability measures, may provide an integrated analysis approach that is effective in terms of finding problems, reducing the cost of user testing, and providing an adequate basis for interface redesign.

Operations Research Methods for Usability Evaluation

Operations Research (OR) is the use of mathematical models to analyze real-world and complex systems and improve and optimize performance. OR has historically been instrumental in improving the design of entire systems (e.g., Polaris Missile Project; Ravindran, Phillips & Solberg, 1987). Only a few HCI studies have previously used OR methodologies for improving systems usability (e.g., Jenamani et al., 2001; Sholl et al., 1991). However, a considerable amount of research is currently being conducted in this area to determine how usability can be assessed and improved using mathematical methods (e.g., Baber and Mellor, 2001; Gray et al., 1992; Kitajima et al., 2005). Some of the techniques being used include stochastics and optimization, such as Markov models and probabilistic finite state models (Thimbleby et al., 2001; Sholl et al., 1991).

Markov models have a memory-less property; that is, each state predicted by a model is only dependent upon the immediate preceding state and no others. This type of model is relevant to analysis of user interfaces that provide all the information needed to move to the next step at the current state. Markov models can be used to determine quantitative outcomes such as average number of steps to reach a goal state. Markov Chains have been used for website customization by creating algorithms to predict user behavior with pages and to determine which links to display for future actions (Jenamani et. al., 2001). Markov models have also been used to predict the number of mouse clicks to find relevant information on a website or the number of steps to complete a desktop computing task. Kitajima et al. (2005), applied an algorithm to determine the average number of clicks in searching a website. The algorithm was first applied to an original interface design and then used again after usability improvements were applied to the site to determine if the average number of clicks was reduced. In order to create such Markov models, user testing is required with the original system interface. User behavior state observations were coded in transitional probability matrices, and the algorithm was created and validated by comparing it to the initial data. (This research was considered as a key basis for the new usability evaluation presented here.) Thimbleby et al. (2001) also applied Markov chains to several applications, such as a microwave oven interface and a cell-phone dialing interface. They used Markov chains, to model states of these devices and to predict the number of steps it would take a user to perform specific tasks. In order to gather data, the authors created a Mathematica simulation of a microwave control panel. The simulation, however, modeled the device and not the user. Data was used to determine ease of use for human operation.

Based on this review, OR methods appears to be viable and useful approaches to evaluate interface usability. They provide quantitative and objective measures that can be used to compare design alternatives and create benchmarks for later investigations. Once user testing has been completed, data can be used in, for example, Markov models to repeatedly predict objective outcomes, such as number of clicks in application use, with different interface design alternatives. Markov models represent a form of user modeling that can be combined with subjective usability measures to provide an integrated evaluation method for assessing overall system effectiveness and comparing various alternative designs.

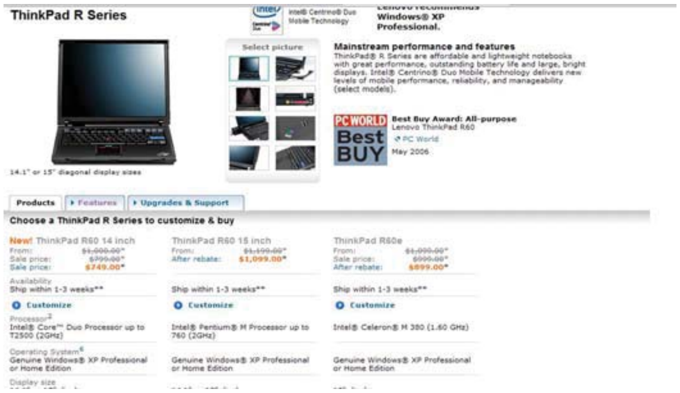
METHOD

Overview of System Effectiveness Score

Designer and user surveys were developed based on the Abowd and Beale interaction framework and used in conjunction with Markov model predictions of the average number of clicks at an online interface to provide an overall system effectiveness score. Nodes in the Abowd and Beale model include the user, interface and system. Links between the nodes represent user articulations, interface translations or systems processing. More specifically, the link between the user and the interface (articulation; see Figure 1) represents how well the user is able to match their goals to the input states of the system. The link between the interface input and the system (performance) determines if each input has a logically mapped function. The link between the system and the interface output (presentation) represents the accuracy and completeness of system state information provided to the user. The link between the interface output and the user (observation) represents how well a user can interpret the information from the system. These links all represent key design dimensions for HCI. In order to determine a system effectiveness score, we decided that all links needed to be evaluated by both users and designers. The designers were to subjectively rank the importance of each dimension. The users were to use the design dimensions as a basis for rating specific web interface alternatives. The designer rankings and user ratings were ultimately to be integrated across the design dimensions.

*Weighting Factor Determination***.** Four designers were surveyed using the Abowd and Beale framework and asked to conduct pair-wise comparisons of the links (HCI design dimensions). All designers were experienced web interface and application developers from a major PC manufacturer. The sample represented the population of a web application development group. The comparisons were used to calculate weighting factors, for each dimension, ranging between 0 and 0.5. The weighting factors were averaged across designers and later used in calculating the overall system subjective score (see end of section). Designers were expected to be most concerned with cognitive load, which was represented by articulation and observation in the framework. In general, commercial web-based ordering system design attempts to minimize user memory load to facilitate easy and quick buying decisions.

Experimental Task. Participants included frequent online shoppers as well as PC shoppers who had never purchased products online. They were asked to use a version of a PC manufacturer's website (see Figures 2 and 3) to find and order a laptop. There were 20 participants in total, including 11 males and 9 females, between the ages of 17 and 25. Half of the participants used an old version of the web interface, which required a minimum of 11 clicks to buy. The other half used an improved version, which included a multi-level navigation interface, more salient buttons and all information about a particular type of computer contained on one page. It required a minimum of 9 clicks to buy.



Mainstream performance and features
ThinkPad® R Series are affordable and lightweight notebooks with great performance, outstanding battery life and large, bright displays. Intel® Centrino® Duo Mobile Technology delivers new levels of mobile performance, reliability, and manageability (select models).

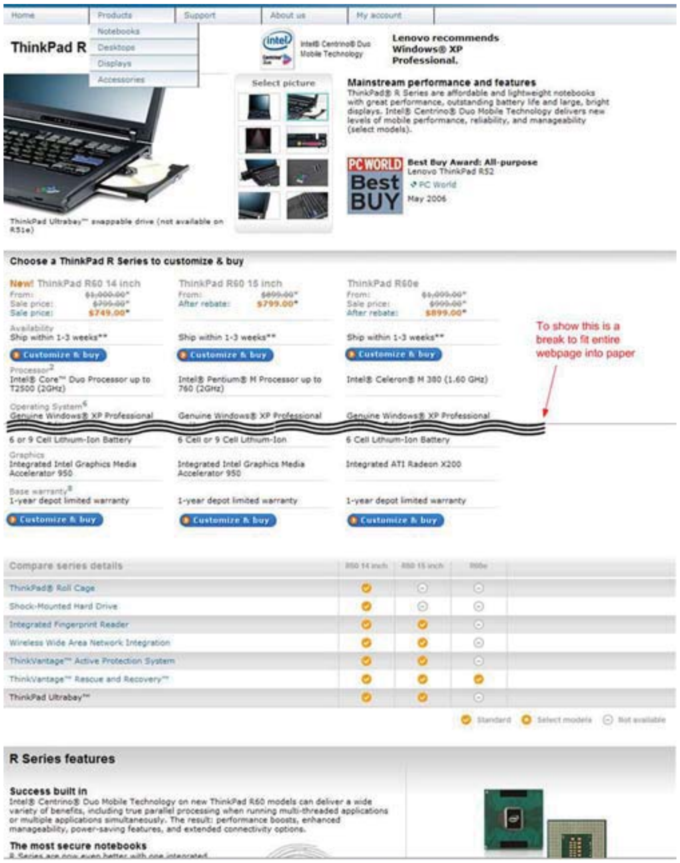
Best Buy Award: All-purpose
Lenovo ThinkPad R60
PC World
May 2006

14.1" or 15" diagonal display sizes

Choose a ThinkPad R Series to customize & buy

New! ThinkPad R60 14 inch	ThinkPad R60 15 inch	ThinkPad R60e
From: ~~$1,000.00*~~ Sale price: ~~$799.00*~~ Sale price: $749.00*	From: ~~$1,199.00*~~ After rebate: $1,099.00*	From: ~~$1,099.00*~~ Sale price: ~~$999.00*~~ After rebate: $899.00*
Availability Ship within 1-3 weeks**	Ship within 1-3 weeks**	Ship within 1-3 weeks**
Customize	Customize	Customize
Processor[2] Intel® Core™ Duo Processor up to T2500 (2GHz)	Intel® Pentium® M Processor up to 760 (2GHz)	Intel® Celeron® M 380 (1.60 GHz)
Operating System[6] Genuine Windows® XP Professional or Home Edition	Genuine Windows® XP Professional or Home Edition	Genuine Windows® XP Professional or Home Edition
Display size		

Figure 2: Old version of web interface



Mainstream performance and features
ThinkPad® R Series are affordable and lightweight notebooks with great performance, outstanding battery life and large, bright displays. Intel® Centrino® Duo Mobile Technology delivers new levels of mobile performance, reliability, and manageability (select models).

Best Buy Award: All-purpose
Lenovo ThinkPad R52
PC World
May 2006

ThinkPad Ultrabay™ swappable drive (not available on R51e)

Choose a ThinkPad R Series to customize & buy

New! ThinkPad R60 14 inch	ThinkPad R60 15 inch	ThinkPad R60e
From: ~~$1,000.00*~~ Sale price: ~~$799.00*~~ Sale price: $749.00*	From: ~~$899.00*~~ After rebate: $799.00*	From: ~~$1,099.00*~~ Sale price: ~~$999.00*~~ After rebate: $899.00*
Availability Ship within 1-3 weeks**	Ship within 1-3 weeks**	Ship within 1-3 weeks**
Customize & buy	Customize & buy	Customize & buy
Processor[2] Intel® Core™ Duo Processor up to T2500 (2GHz)	Intel® Pentium® M Processor up to 760 (2GHz)	Intel® Celeron® M 380 (1.60 GHz)
Operating System[6] Genuine Windows® XP Professional	Genuine Windows® XP Professional	Genuine Windows® XP Professional
6 or 9 Cell Lithium-Ion Battery	6 Cell or 9 Cell Lithium-Ion	6 Cell Lithium-Ion Battery
Graphics Integrated Intel Graphics Media Accelerator 950	Integrated Intel Graphics Media Accelerator 950	Integrated ATI Radeon X200
Base warranty[8] 1-year depot limited warranty	1-year depot limited warranty	1-year depot limited warranty
Customize & buy	Customize & buy	Customize & buy

To show this is a break to fit entire webpage into paper

Compare series details	R60 14 inch	R60 15 inch	R60e
ThinkPad® Roll Cage	✓	⊖	⊖
Shock-Mounted Hard Drive	✓	⊖	⊖
Integrated Fingerprint Reader	✓	✓	⊖
Wireless Wide Area Network Integration	✓	✓	⊖
ThinkVantage™ Active Protection System	✓	✓	⊖
ThinkVantage™ Rescue and Recovery™	✓	✓	✓
ThinkPad Ultrabay™	✓	✓	⊖

✓ Standard ○ Select models ⊖ Not available

R Series features

Success built in
Intel® Centrino® Duo Mobile Technology on new ThinkPad R60 models can deliver a wide variety of benefits, including true parallel processing when running multi-threaded applications or multiple applications simultaneously. The result: performance boosts, enhanced manageability, power-saving features, and extended connectivity options.

The most secure notebooks
R Series are now even better with one integrated

Figure 3: New version of web interface

Developing Markov Chain Models. JavaScript was used to record user actions at the interfaces and the action sequences were used to create transitional-state probability matrices, which were based on the actual number of users going from interface state *i* to state *k* (e.g., initial ThinkPad series selection page to the final laptop purchase page). The assumptions of the Markov model include: (1) the sum of each row (in the probability matrix) must equal 1; and (2) the probability of the next interface state can only depend on the current state. To determine the average number of clicks to ordering task completion, the equation from Kitajima et al., $u^{(i)} = 1 + \Sigma\ P_{ik}\ u^{(k)}$, was used. The equation was applied to the probabilities (P_{ik}) from the transitional probability matrices (collected from user actions) and estimates of the average number of steps from state *i* to state *k* ($u^{(k)}$), collected from designers.

*Rating System Effectiveness***.** At the end of the experiment, subjects were asked to rate the interfaces, using the Abowd and Beale framework by assigning each link a rating from 1 to 10. This subjective data was averaged for each link and used in the overall system effectiveness score.

*Overall System Effectiveness Score***.** This score was used to compare the alternative interface designs. The average designer weights were multiplied by the average end-user ratings for each design dimension and summed across dimensions. This was considered a partial system score. The partial score was then divided by the predicted average number of clicks from the Markov model to determine the "overall system effectiveness" score. The interface alternative with the highest ratio of perceived usability per click was considered to provide greater ease of use, etc. The novelty of this score is that it reflects the designers' intent for an application and users' perceptions of a system.

To validate the Markov model, and consequently the new score, t-tests were used to determine if the actual number of observed clicks was significantly different from the number predicted by the model. Statistical comparisons were also made among the scores for the old and improved interfaces.

RESULTS

A Durbin Watson test was initially used to determine if the assumptions of the Markov model were met. The new interface showed no first order autocorrelation, while the old interface did. A normalization procedure was applied to both the new and old transitional probability matrices to eliminate any first order autocorrelation. The Durbin Watson test was again applied showing no first order autocorrelation for the new interface and a slight improvement for the old interface. The test revealed mixed results across interfaces; however, since websites are generally created to support customers in making decisions through information on the current page, the normalized transitional probability matrices were used to determine the average number of clicks for both designs.

The Markov model predicted that it would take the users an average of 11.5 clicks to order a ThinkPad R60 using the old interface (actual 12.9) and 9 clicks for the new interface (actual 9.2). A t-test revealed the difference in actual clicks across the interfaces to be significant ($p = 0.0004$). The new interface required significantly fewer clicks. T-tests were also used to compare the actual click counts to the predicted click counts for all subjects with p-values for the new and old interfaces of 0.439 and 0.605, respectively. There was no significant difference between the actual and predicted number of clicks with either interface. A t-test also revealed the number of predicted clicks across interfaces to be significant ($p = 0.033$). The new interface design significantly reduced the number of clicks to order a ThinkPad R60.

As previously stated, designers were expected to rate the articulation and observation dimensions of HCI design as being more important. T-tests comparing designer ratings of articulation and observation with performance dimensions supported our hypothesis. The average designer rankings were multiplied by average user ratings. A t-test was then used to compare the partial score for the new interface with the old interface for all subjects, indicating a significant difference ($p < .0001$). The partial score for the new interface was significantly higher.

The partial scores were then divided by the predicted average number of clicks from the Markov models to yield perceived usability per click for each interface. The new and old interfaces had scores of 0.939 and 0.475, respectively. A t-test revealed the overall scores for the new and old interfaces for all subjects to be significantly different ($p < .0001$).

The purpose of the Markov model was to predict the number of user clicks and to potentially reduce the need for additional user testing of interface design revisions. Designers can speculate an average number of steps to transition among states in a new interface and these estimates can be multiplied by probabilities determined for an original interface (through user testing). By using this method to predict the average number of clicks for the new www.Lenovo.com interface, the predicted clicks to purchase a ThinkPad R60 was 9.35 (recall the actual was 9.2). A t-test revealed that the actual number of clicks was not different than the predicted number of clicks ($p = 0.270$). The Markov model was accurate in predicting the average number of clicks at the new interface.

In order to obtain user ratings to complete the system effectiveness score in this approach, a focus group would be necessary to subjectively evaluate any new design alternatives. However, the approach would still significantly reduce time and money necessary for user testing.

DISCUSSION

Based on the designer rankings of design dimensions, it was found that they were, in fact, more concerned with user cognitive load, as represented by articulation and observation than system processing requirements. This is likely due to the perception that if customers cannot find what they are looking for, it may lead to frustration, lost customers and lost revenue. Thus, designers emphasized articulation and observation in the overall system effectiveness score.

It was also hypothesized that the new interface design would improve perceived usability. The multi-level navigation interface was integrated in the new web page to reduce cognitive load by making it easier to find and view all options. Users could reach many states with one click and it was identified as one of the most usable features. More prominent buttons aided in easily identifying next steps; whereas, in the original interface, users had a difficult time finding the customize button as they often scrolled up and down a page or backtracked to determine what to do next. The partial system effectiveness score was higher for the new interface (8.6) than the old interface (5.2).

It was expected that the new interface would produce a higher overall effectiveness score because it was perceived to have higher usability. Beyond this, the old interface degraded performance. For example, from a notebook's features tab, some users found it difficult to identify what to do next and once they found the products tab, some scrolled up and down trying to determine what option to select. The new interface alleviated both these problems with consolidated pages of information and more prominent buttons. With higher perceived usability and fewer clicks, the new interface led to a higher effectiveness ratio. The fewer number of clicks with the new interface was achieved mainly through a reduction in user errors (vs. a reduction in the number of web pages).

It was also expected that the Markov model would accurately predict the average number of clicks used in the equation developed by Kitajima et al. Because Markov models are used to represent probabilistic behavior they proved valid in the present work, as human interaction with a website is not deterministic in nature. The model revealed variability among participants but did not reveal the magnitude of the error.

CONCLUSION

The objective of this work was to create a new measure of system usability. We observed there to be few quantitative, objective measures and many subjective measures, which may be insufficient for justifying system design changes. We developed novel subjective measures using the Abowd and Beale HCI framework and an objective measure, based on Markov models. The integration of these measures was found to be effective in objectively selecting among alternative interface designs. The method is also easy to implement and can be used with several interface alternatives without the need for additional testing. It cannot, however, be applied to interfaces where selection of the next system state depends on several previous states (and not only the current state). For example, aviation system interfaces, which are often not designed to reduce user cognitive load, may require pilots to remember novel information across system states and stages (e.g., Kaber, Tan, Riley & Endsley, 2001). This type of setup violates the basic assumptions of a Markov model and the number of actions to reach a goal state will not be accurately predicted for such interfaces. This type of analysis is also inappropriate when there is a radical redesign of an interface with few overlapping states with the original interface design. The accuracy of the analysis may also be improved by increasing the number of designers and users surveyed for ranking design dimensions and rating design alternatives, respectively.

Future research includes using Markov models to predict a user's future actions in interface use in real-time in order to make relevant options more salient and to improve usability. We are also interested in determining an approach to incorporate time-on-task in the overall system effectiveness score. Time-on-task is important because it may ultimately impact customer retention.

ACKNOWELEDGEMENT

Lashanda Lee's work on this project was supported by the National Institute for Occupational Safety & Health (NIOSH; Grant No. 2 T42 OH008673-03). All opinions are those of the authors and do not necessarily reflect the views of NIOSH.

REFERENCES

Abowd, G.D. and Beale, R. (1991). Users, systems and interfaces, a unifying framework for interaction. In D. Diaper and N. Hammond (Eds.), *HCI '91 People and Computers VI* (pp. 73-87). Cambridge, NY: Cambridge University Press.

Baber, C., and Mellor, B. (2001). Using critical path analysis to model multimodal human computer interaction. *International Journal of Human Computer Studies, 54,* 613-636.

Card, S.K., Moran, T.P., and Newell, A. (1983). *The Psychology of Human Computer Interaction*. Hillsdale, NJ: Erlbaum.

Cockton, G., Lavery, D., and Woolrych, A. (2002). Inspection-Based Evaluations. In J. Jacko, and A. Sears (Eds.), *The Human Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications (*pp. 1118-1138). Mahwah, NJ: Lawrence Erlbaum Associates.

Gray, W., John, B. E., and Atwood, M. E. (1992). The Précis of Project Ernestine or an Overview of a Validation of GOMS. In *Proceedings of the 1992 Conference on Human Factors in Computing Systems, CHI,* (pp. 307-312). New York, NY: ACM.

Hornbaek, K. (2005). Current Practice in Measuring Usability: Challenges to Usability Studies and Research. *International Journal of Human-Computer Studies, 64,* 79-102.

Jenamani, M., Mohapatra, P. K. J., & Ghose, S. (2002). Online customized index synthesis in commercial web sites. *IEEE Intelligent Systems, 17*(6), 20-26.

Kaber, D. B., Riley, J., Tan, K-W and Endsley, M. R. (2001). On the design of adaptive automation for complex systems. *International Journal of Cognitive Ergonomics*, *5(1),* 37-57.

Kitajima, M., Kariya, N., Takagi, H., & Zhang, Y. (2005). Evaluation of website usability using Markov Chains and latent semantic analysis. *IEICE Transactions on Communications, E88-B*(4), 1467-1475.

Molich, R and Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, *33*, 338-342.

Mulhem, P., and Nigay, L. (1996). Interactive information retrieval systems: From user centered interface design to software design. *Proceedings of the 1996 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 96,* (pp. 326-334). Zurich, Switzerland: ACM.

Norman, D. (1988). The Psychology of Everyday Things. New York, New York: Doubleday.

Ravindran, A., Phillips, D.T. and Solberg, J.J. (1987). Operations Research: Principles and practice. Wiley: New York

Sholl, H. A., Ammar, R. A., and Weiss, W. S. (1991). Using probabilistic finite state models to evaluate human-computer interfaces. In *Conference proceedings of the 1991 IEEE international conference on systems, man, and cybernetics,* (pp 1145-1150). Piscataway, NJ: IEEE

Thimbleby, H., Cairns, P., and Jones, M. (2001). Usability Analysis with Markov Models. *ACM Transactions on Computer-Human Interaction*, *8*(2), 99-132