To appear in HCI International, 2009
19-24 July 09, San Diego, CA, USA

Toward Cognitive Modeling for Predicting Usability

Bonnie E. John[1] and Shunsuke Suzuki[2]

[1] Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA, 15213, USA
bej@cs.cmu.edu
[2]NEC Corporation, 8916-47, Takayama-cho, Ikoma, Nara 630-0101, Japan
s-suzuki@cb.jp.nec.com

Abstract. Historically, predictive human performance modeling has been successful at predicting the task execution time of skilled users on a desktop computer. More recent work has predicted novice behavior in web searches. This paper reports on a collaborative effort between industry and academia to expand the scope of predictive modeling to the mobile phone domain, both skilled and novice behavior, and how human performance relates to the perception of usability. Since, at this writing, only preliminary results to validate models of mobile phone use are in, we describe the process we will use to progress towards our modeling goals.

Keywords: Cognitive modeling, GOMS, KLM, CogTool, Information Foraging.

1 Introduction

Predictive human performance modeling has been an HCI "holy grail" for decades. If the field had a computational model of a human that could perform like a human (including perception, cognition and motor action), make errors like a human, learn like a human, and experience emotions like a human, then we could test our design ideas as they emerge in the design process, quickly and inexpensively. Just as an automotive crash dummy saves consumers from poorly designed vehicles, a *cognitive crash dummy* could expose design ideas that would harm, or at least annoy, users before they were brought to market. This would save companies substantial investment in building and marketing products that people would have a hard time learning or using or would not enjoy.

The first model to make progress toward these goals was Card, Moran and Newell's Model Human Processor in 1983 [1]; researchers have been making slow but steady progress since then. The most reliable models today predict task execution time of skilled users. Models such as Keystroke-Level Model (KLM; [2]) and GOMS (Goals, Operators, Methods, and Selection rules; [3]) have been validated in over 100 research papers by scores of authors, primarily on tasks performed on a desktop computer. Information Foraging Theory [4], a more recent entry into the realm of computational predictive modeling, predicts exploratory behavior of people searching

for information. It has been operationalized in an ACT-R [5] model called SNF-ACT [6] and validated against human data on web search tasks.

Despite research progress creating and validating theory, UI developers have not adopted predictive human performance modeling as a frequently-used tool for design. Recent work has embodied these theories into tools that allow practicing developers to achieve the benefits of modeling without investing considerable time in learning to model and constructing each new model (e.g., [7, 8, 9]).

However, it is difficult to make a trustworthy tool for practical design problems. This paper explains the process of doing so in the context of collaborative research between NEC, PARC and Carnegie Mellon University. Our project is aimed at producing a tool for predicting task execution time of skilled users, novice exploration to accomplish a goal, and subjective perception of the usability of mobile phones.

2 The Process of Making a Trustworthy, Practical Tool for Design

The process of making a trustworthy, practical tool for design is shown in Figure 1. Each time a new domain is entered, or a new metric is added, the theory, tool and models must be validated with data from appropriate users to produce a trustworthy tool for prediction. If the models' predictions do not match the human data sufficiently, either the theory or the tool, or both, must be revised until valid predictions are produced. The next question is whether the tool is learnable and usable by UI designers in their work process. User-centered design techniques should be used to design, evaluate, and redesign, until the tool is practical for design.

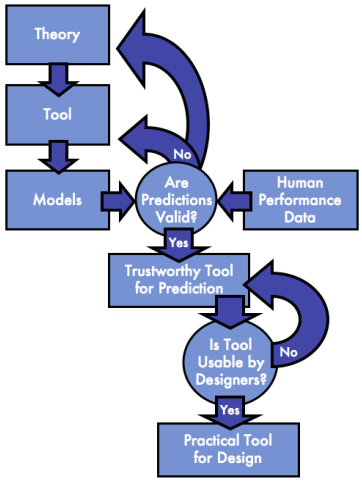


Fig. 1. General process of human performance modeling research that leads to a practical tool for design

Our project started with CogTool, a tool that allows UI designers to create valid Keystroke-Level Models in one tenth the time of doing them by hand as originally demonstrated by Card, Moran and Newell [2]. It has been shown to be easily learnable by users with no background in psychology or cognitive modeling ([8] and tutorials at professional conferences like HFES, BRIMS, and HCII). Recent research with CogTool has extended it beyond KLM and predictions of skilled task execution time to information foraging theory and predictions of novice exploration behavior [10, 11]. From this starting point, we set out to expand CogTool's ability to predict human behavior to a new domain, mobile phones, and to a new metric, subjective impressions of usability as measured by the Mobile Phone Usability Questionnaire (MPUQ) developed by Ryu [12, 13].

Thus, this project will touch all the points in Figure 1. We start by using CogTool as it exists to make predictions of skilled execution time and novice exploration behavior and test those predictions against human data on mobile phones, fully expecting that adjustments to the underlying theory and tool will need to be made. After making changes to the theory and tool to produce valid predictions of these metrics, we intend to correlate various aspects of the predictions with people's perceptions of usability. After verifying that we can make trustworthy predictions, we will determine whether CogTool can be used by mobile phone designers and adjust CogTool's UI until it becomes a practical tool. At this writing, we are at the first part of the process, making predictions with CogTool as it exists and comparing those predictions to human data. The remainder of this paper will describe the current state of the research.

3 The New Domain – Mobile Phones

Fig. 2. N905i mobile phone shown at the screen that was the start for each task.

Mobile phones were chosen as the domain in which to pursue this approach. This product category is important to the corporation and the discrete nature of the tasks users perform on mobile phones makes them relatively easy for collecting human data and to model. In addition, CogTool had previously been shows to make good predictions of skilled use of a similar hand-held device (PDAs, [14, 15]).

Although the project will evaluate several different mobile phones, this paper will use the N905i, shown in Figure 2, as an example of our research process.

The tasks we are examining are varied, as follows.

1. Call a number from a phone book
2. Store a number into a phone book
3. Put an event into a Schedule
4. Change a Security Setting
5. View a previously sent mail message
6. Set a previously stored picture to be the wallpaper.
7. Delete a previously stored picture
8. Add a function into a shortcut
9. Check memory info
10. Shoot a movie, check it, and save it

Data was collected from skilled users who had owned their phones for at east two months and from novice users who had never used this model of phone. The phone screen was captured on video, which was later transcribed to identify which buttons were pressed, when each button was pressed, and how long it took the phone to respond to each button press (system response time).

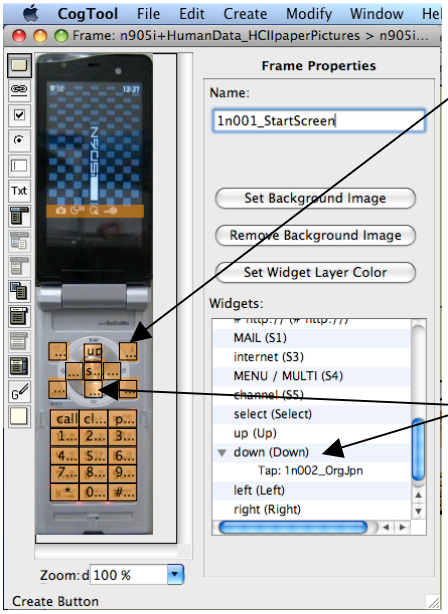
4 CogTool and Initial Models

CogTool is a prototyping and cognitive modeling tool created to allow UI designers to rapidly evaluate their design ideas. A design idea is represented as a storyboard (inspired by the DENIM Project [16]), with each state of the interface represented as a node and each action on the interface (e.g., button presses) represented as a transition between the nodes. Figure 3 shows the start state of the storyboard, where buttons are placed on top of an image of the phone.

Figure 4 shows a storyboard for six instances of the first task, calling a person who is already listed in the phone's contact list. The first action at the start state is to press the down button called out in Figure 3. Because different contacts are located at different points of the phone book, the task takes different paths from the start screen to completion of the task. We will use Calling Person4 as the example in the remainder of this paper.

After creating the storyboard, the next step is to demonstrate the correct actions to do the task. CogTool automatically builds a valid Keystroke Level Model from this demonstration. It creates ACT-R code that implements the Keystroke-Level Model and runs that code, producing a quantitative estimate of skilled execution time and a visualization of what ACT-R was doing at each moment to produce that estimate (Figure 5).

Since mobile phones are a new domain for CogTool, we do not expect that the predictions it makes "out of the box" will be very accurate when compared to human data. We expect to have several iterations of comparing the predictions to human data and fixing the underlying theory and CogTool's implementation of that theory, before we can make trustworthy predictions to help design. The next section presents preliminary analysis of one such iteration.



Each Button in the picture of the phone has a button "widget" drawn on top. Actions on widgets follow transitions as defined in the storyboard (Fig 4).

If you tap the down button, CogTool transitions to the next frame in the storyboard (Fig. 4)

Fig. 3. Start screen of the CogTool prototype.

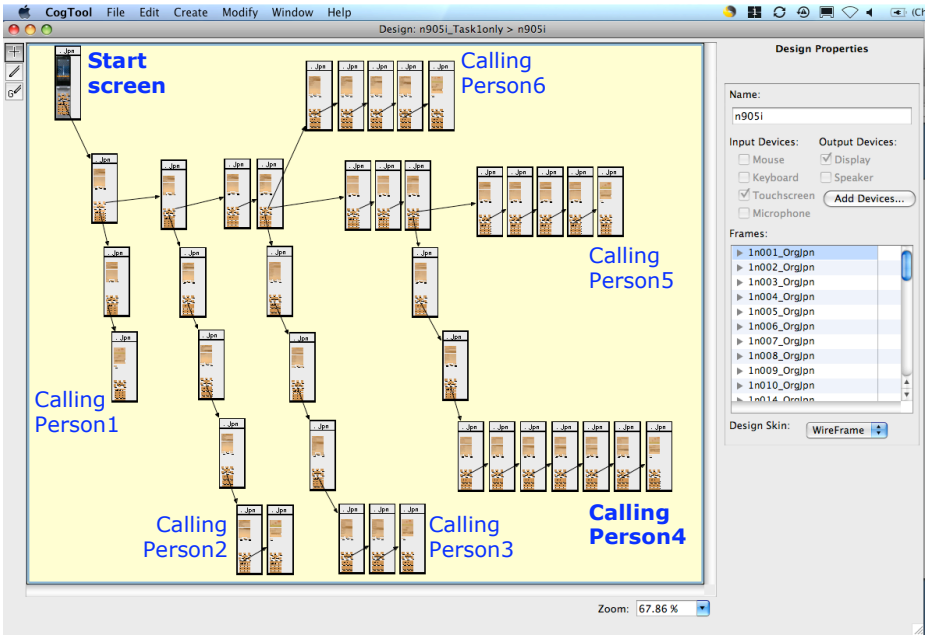


Fig. 4. Storyboard of the screens a person would pass through to accomplish Task 1 (making a call from the phone book) for six different instances of the task, i.e., calling six difference people in the phone book. We will use the instance of Calling Person4 as an example throughout this paper.

3.2 Comparing initial models to human performance data

The first step in comparing human performance data to the predictions of models is to make sure the same metrics are used in both the data and the models. For example, CogTool models predict not only when a button will be pressed, but also the thinking time and visual perception that precedes pressing the button. Only button presses were recorded in the empirical study, so we cannot directly compare the "total" time predicted by the model against the "total" time observed in the experiment. Adjusting for this difference, and comparing the time from first key press to the appearance of "Calling Person4" on the screen, the CogTool model predicted 11.049 seconds. The mean of five skilled participants was 9.770 seconds, an over-prediction of the average by 13% and an average absolute percent error of 15% between the predicted time and each observed time. This level of prediction is within the 20% error typically claimed by KLM and is an excellent prediction for an initial foray into a new domain and device.

The next step is to go to a deeper level of comparison and look at the predictions for each individual action. We expect the quantitative comparisons to get worse, as explained by Card, Moran and Newell [1], but we are looking for patterns in behavior at this point, not absolute quantitative match.

The first types of patterns we hope to see are those predicted by the model. Consider Figure 5, a timeline of the model's predictions for the Calling Person4 task provided by a CogTool as a visualization of its behavior. The rows in the timeline represent different types of actions in the model. The changes in the phone's screens are on the top gray line, with the estimates of system wait time (between button press and when the screen can be read) in the second row (light gray). The three purple rows show activity associated with vision: Vision-Encoding, Eye Move-Execute and Eye Move-Preparation. The central gray row represents the cognition that controls behavior, both the long "Mental operators" empirically established by Card, Moran and Newell, and the short ACT-R cognitive acts that control vision and hand motions. The bottom red row shows the button presses, in this case, with the right hand (the thumb). The model predicts a pattern:

1 press (at time=0), pause, 6 presses, pause, 8 presses, pause, 1 press.

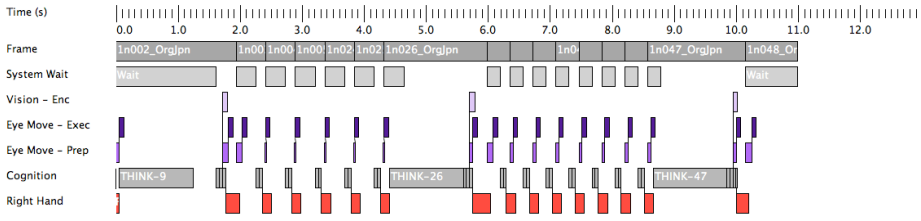


Fig. 5. Timeline of a CogTool model prediction.

Consider Figure 6, where the data from five participants are placed below the model's timeline, aligned so that their first key presses all start at 0.0 sec. The top four participants display a pattern in keeping with the model's prediction (1 press, pause, 6 presses, pause, 8 presses, pause, 1 press), except for P9 who does not pause for long before the last key. However, the bottom participant, P1 does not show this pattern at all. When we went back to the video of this participant, we found that although P1 used the same number of keys to complete the task, he did not use the same keys as the other participants or the model. Further investigation is needed to understand whether this was due to an error or whether it represents an alternative correct method for this task. Either way, in the majority of cases of this small sample, CogTool automatically predicted a pattern of behavior that was observed in human performance, even without modifying CogTool for the mobile phone domain.

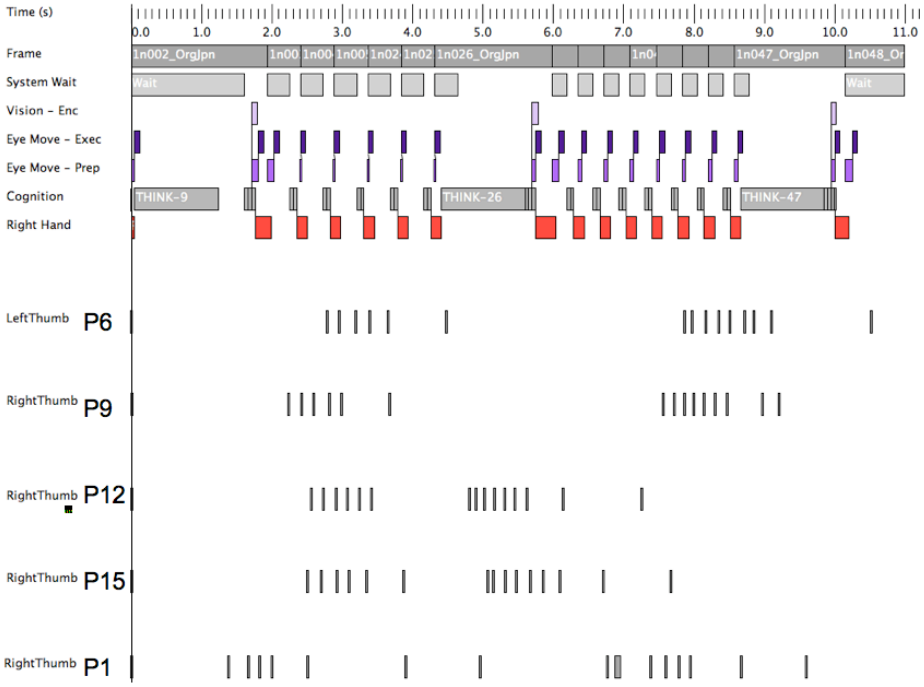


Fig. 6. Timeline of a CogTool model prediction with keypress data from five participants aligned below it.

Looking more closely at the data of the people who used the same keystrokes as the model (the top four), another pattern can be seen, one not predicted by the un-modified CogTool. Each participant shows two grouping of keys pressed close together in time, one of six keys in the beginning of the task and one of eight keys at the end of the task. Of these eight groupings, six show a distinct pause before the last keystroke in the group (P6, 1st group; P9, both groups; P12, 2nd group; P15, both

groups). The groupings of six are repeated pressing of the S3 key to move across a set of icons at the top of the screen, some of which drop down a list of items that can be selected. When the desired icon is reached and its list drops down, the user then hits the Down key eight time to move down to the desired contact and hits the Call button to complete the task. The pauses come before the last S3 key press and the last Down key press. In both cases, the user is watching a highlight move across (or down) the screen and can anticipate when the next key press will bring the highlight to the desired item. The pause before the last key press might represent a strategy to avoid over-shooting. This monitoring activity was not included in the original systems tested by Card, Moran and Newell and therefore is not represented in the original Keystroke-Level Model. Thus, we have identified a case where we may need to develop new theory about monitoring and anticipatory keystrokes (i.e., iterate on the theory) and build it into the tool (i.e., iterate on the tool) before we can produce trustworthy predictions in this domain.

Another case where the predictions do not match the data is in the inter-keystroke times. All four users who did the task in the same way as the model pressed the same key far faster than CogTool did, as seen by the denser grouping of keystrokes in the participants' timelines than in the model timeline. In this case, we will have to iterate on the underlying theory of motor movement to allow it to produce faster keystrokes. The timeline shows us that CogTool inserts visual perception of a key between each keystroke, which likely to be wrong for repeated keystrokes, especially given the monitoring activity described above where the user's eyes are presumably on the screen not the buttons.

With a model of just one instance of one task and data from five participants, the timeline visualization has suggested that the model is making reasonable predictions of the grouping of actions but is missing some important patterns of human behavior. More tasks and more data will have to be analyzed to be sure it is necessary to change the underlying theory and build it into CogTool to get trustworthy predictions. However, this small example illustrates the process of model validation this project has undertaken.

5 Future work

In addition to following the process in Figure 1 for skilled task execution time predictions on mobile phones, this project will also examine the prediction of novice exploration behavior, with CogTool-Explorer ([10, 11], a version of CogTool that predicts novice behavior). As with skilled behavior, we do not expect CogTool-Explorer to be able to predict a new domain (mobile phones instead of web searches) in a new language (Japanese instead of English) without iteration on the theory and tool. We have already identified improvements to the tool required for mobile phones, for example, mobile phones have "soft keys" where the label of the key is displayed on the screen instead of being printed on the key and CogTool-Explorer was not originally designed to represent that relationship.

Perhaps more interestingly, when we have succeeded in producing trustworthy predictions of behavior, we intend to correlate this behavior with subjective

impressions of usability as measured by the Mobile Phone Usability Questionnaire (MPUQ) developed by Ryu [12, 13]. Unlike empirical methods that can only correlate observed behavior, like time on task or number of errors, with questionnaire results, we can extract much more varied metrics from the models against which to correlate subjective impressions. For example, total time on task may not correlate with subjective impressions, but time spent in cognition may. Or more complex measures may be needed, like time spent in cognition that is not in parallel with motor movements for skilled users. Or number of keys looked at by CogTool-Explorer before making a choice, for the subjective impressions of novice users. Or amount of system response time not in parallel with cognition (i.e., making the user wait). Because CogTool produces a process model of perception, cognition and motor actions necessary to do a task, many combinations of actions can be explored to see if any can explain a significant part of the variance in subjective impressions. If a significant correlation can be found, then the predictive human performance models will be extended to a subjective metric, moving the field closer to the holy grail.

References

1. Card, S. K., Moran, T. P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, NJ (1983)
2. Card, S. K., Moran, T. P., Newell, A.: The Keystroke-Level Model for User Performance Time with Interactive Systems. Commun. ACM 23, 7, 396--410 (1980)
3. Card, S. K., Moran, T. P., Newell, A.: Computer Text-Editing: An Information-Processing Analysis of a Routine Cognitive Skill. Cognitive Psychology 12, 32--74 (1980)
4. Pirolli, P. Card, S.: Information Foraging in Information Access Environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1995), pp. 51—58. ACM Press/Addison-Wesley Publishing Co., New York (1995)
5. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y. An Integrated Theory of the Mind. Psychological Review, 111(4) 1036--1060 (2004)
6. Fu, W.-T., Pirolli, P.: SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. Human-Computer Interaction, 22, 355--412 (2007)
7. Blackmon, M. H., Kitajima, M., Polson, P. G. (2005). Tool for Accurately Predicting Website Navigation Problems, Non-Problems, Problem Severity, and Effectiveness of Repairs. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '05, pp. 31--40 ACM, New York, NY (2005)
8. John, B. E., Prevas, K., Salvucci, D. D., Koedinger, K.: Predictive Human Performance Modeling Made Easy. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, pp. 455--462 ACM, New York, NY (2004)
9. Wu, C., Liu, Y. Usability Makeover of a Cognitive Modeling Tool. Ergonomics in Design 15 (2) 8--14 (2007)
10. Teo, L., John, B. E.: Towards Predicting User Interaction with CogTool-Explorer. In Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting, pp. 950--954. HFES, Santa Monica, CA (2008)
11. Teo, L., John, B. E., and Pirolli, P.: Towards a Tool for Predicting User Exploration. In CHI '07 Extended Abstracts on Human Factors in Computing Systems, CHI '07, pp. 2687--2692. ACM, New York, NY (2007)

12. Ryu, Y. S.: Development of Usability Questionnaires for Electronic Mobile Products and Decision Making Methods, Doctoral dissertation, State University, Blacksburg, VA, USA, (2005)
13. Ryu, Y. S. Smith-Jackson, T. L. Reliability and Validity of Mobile Phone Usability Questionnaire (MPUQ). Journal of Usability Studies. 2(1), 39--53 (2006)
14. Luo, L., John, B. E.: Predicting Task Execution Time on Handheld Devices Using the Keystroke-Level Model. In: Proceedings of the International Conference on Human Factors in Computing System (CHI 2005), pp. 1605--1608. ACM Press/Addison-Wesley Publishing Co., New York (2005)
15. Luo, L., Siewiorek, D. P.: KLEM: A Method for Predicting User Interaction Time and System Energy Consumption during Application Design. In: Proceedings of the 11th International Symposium on Wearable Computers (ISWC 2007), pp. 69--76 IEEE Press, New York (2007)
16. Lin, J., Newman, M. W., Hong, J., Landay, J. A.: Denim: An Informal Tool for Early Stage Web Site Design. In CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI 2001), pp. 205--206 ACM, New York, NY (2001)