Analytical Prototyping of Personal Technologies: Using predictions of time and error to evaluate user interfaces

Chris Baber & Neville Stanton*

School of Elec. & Elec. Eng., The University of Birmingham. UK

*School of Design, Brunel University. UK

c.baber@bham.ac.uk

Abstract: In this paper, we present a technique for analytical prototyping of personal technology. The technique allows conceptual design to be evaluated in terms of two predictive measures of user performance (time and error). The technique is based on the theory of rewritable routines, which is presented in this paper as a novel approach to considering mental models.

Keywords: Analytical Prototyping, Personal Technology, User Modelling

1 Introduction

Wichansky (2000) reviewing developments in usability evaluation concludes that future usability evaluation needs to be 'quick and clean'. "Quick studies, which are highly efficient…will be highly useful to decision-makers…developing products that evolve as rapidly as 3 months. Clean studies are necessary…to provide valid and reliable data for correct decisions to be made." [p.1004]. In this paper, we consider an approach to the development of usable products that is both quick and clean. The approach, 'analytical prototyping', is a means of incorporating a model of user behaviour into early stages of design.

Analytical prototyping comprises three stages:

- functional analysis;
- scenario-based analysis;
- structural analysis.

The first stage is aimed at defining the features and functions of a product. This takes the form of sketching the product and its uses in order to develop design concepts, e.g., through the use of storyboards (Kiljander, 1999). In our work, we use a state-based description that relates user activity to product functions (see section 4.0). In parallel with functional analysis, the use of scenario-based descriptions helps to illustrate and develop the state-based descriptions. In other words, the state-based descriptions are constructed in order to demonstrate a sequence of human-machine actions in pursuance of a particular user goal.

While storyboarding and state-transition diagrams can provide designers with a means of viewing and debating alternative designs, these approaches lack the ability to produce data relating to actual use. On the one hand, one could simply say that this means that users trials are absolutely essential, and need to be conducted in addition to the other approaches. On the other hand, one could argue that the approaches that are currently being used could be modified to include models of users in order to conduct performance evaluation. It is not our intention to dismiss the value of user trials. However, we feel that it might be possible to modify an approach to describing usage to accommodate models of user performance.

The final stage, and the one that will be demonstrated in this paper, is that of structural analysis. When we say structural analysis, we mean the ability to conduct destructive testing (in an engineering sense) of the conceptual design, i.e.,

testing the design to failure. However, our aim has been to develop techniques that can be applied prior to the construction of working prototypes, i.e., to enable conceptual designs to be analysed and refined at the earliest stages of design. We propose that a form of evaluation is required that satisfies the following criteria:

i. the technique is theoretically-grounded, using an easily assimilated theory of user behaviour;
ii. the technique can be used in the earliest stages of design. In this way, proposals for redesign can be cheaply and speedily implemented;
iii. the technique can be used to produce quantitative data.

2 Theory

Norman (1988) has spoken of the 'gulfs of evaluation and execution', which exist between what users of a computer (or other product) wish to do and what they perceive the product will allow them to do. This gulf can be considered from the perspectives of the product's system image and the user's mental model. The product presents a set of functions to the user (via its features, its display, its controls etc.), in the form of a 'system image'. The user conceptualizes a representation of how the product works (a mental model), possibly in terms of pairing features with functions, e.g., the user might assume that pressing a specific button would turn the product on. When there is a mismatch between the mental model and the system image, then the user could make errors or become frustrated.

Taking these notions of system image and mental model has led us to develop the notion of rewritable routines. In order to present this notion, it is necessary to take a slightly longer detour into the realm of mental models. It is well-known that there are many different and competing views on the nature of mental models. In very broad terms, mental models could be considered as either declarative or procedural, and can be permanent concepts held in long-term memory or temporary frameworks developed during a task. For example, Johnson-Laird (1980) presents mental models as temporary frameworks developed during the task of reading in order to allow the reader to maintain the 'gist' of a story and to develop hypotheses of plot development or character action. In this form, the 'model' is developed on an ad-hoc basis and requires the recruitment of knowledge from long-term memory and its combination with information derived from the text. Consequently, the mental model represents a specific state of affairs in the text (or the world). Alternatively, the notion of mental models as representations of the workings of products appears to be popular in the HCI literature. Researchers might accept that mental models allow users to make inferences about the system with which they are interacting, and that a mental model provides a 'problem space' (in Newell and Simon's, 1972 terms). However, it is generally accepted that such mental models are imprecise, incomplete and inconsistent. Consequently, the problem space might itself be messy and problematic. O'Malley and Draper (1992) have suggested that, rather than the mental model containing a representation of the product, what is required is some means of filling the gaps in the system image and interpreting any information provided by the product. The proposals of O'Malley and Draper (1992) have more than a passing resemblance to the suggestions of Johnson-Laird (1980), and we take this as the starting point for our concept of rewritable routines (see below).

To conclude this brief discussion, users' mental models demonstrate highly fragmented knowledge of the product, rely on a variety of metaphors to contrast the product with other products and can be heavily influenced by the system image. This suggests that 'mental models' need not be complete, coherent internal representations of a product (i.e., not a 'model' in any physical sense) nor that 'mental models' are sufficiently coherent to predict the consequences of actions (i.e., not a 'model' in a mathematical sense).

Figure 1: Mitsubishi mt401 mobile telephone

By way of introduction to the notion of rewritable routines, consider the mobile telephone shown in figure 1. A person who has not previously encountered this model of mobile telephone is set the tasks of turning it on. According to our proposal, the first task (turning on the telephone) will require the user to identify a button or key that can be expected to turn on the telephone. The user might examine the side or top of the telephone, the user might see the 'green, lifted-handset' icon and assume this means make a call (and by analogy, turn on), the user might recognise the ⏻ icon as an ISO symbol for 'on', the user might assume that 'OK' meant turn on or could press any of the other keys on the basis of some other assumption. The main point to be made is that, for the first-time user there are multiple possible routines that can be activated in the goal of turning on the telephone. The selection of a routine will be influenced by prior experience and by the system image. One could imagine scenarios in which a user assumes that the 'green, lifted handset' indicates the initial action that one makes using a 'conventional' telephone. In other words, the user brings a 'routine' to the use of the telephone that incorporates the following sequence of tasks: "pick up handset, dial number, hold conversation, replace handset". The 'system image' provides cues to the first step in this sequence of tasks, i.e., 'pick up handset'. Of course, this is the wrong action and the user should press the key labelled with a 'red, lowered handset', i.e., on this model of mobile telephone, one begins by 'hanging up' (by way of explanation, notice the 'on' symbol above this key).

The users choice of action will be based on the appearance of the telephone (its system image) and the user's prior experience, which leads them to assign meanings to the keys, icons and other objects in the system image.

We begin by assuming that people require very little knowledge of the internal workings of a product in order to use it. Thus, one does need to know how a car engine works in order to drive the car. We also assume that all interaction with products will be goal-based and purposeful, i.e., that people have a reason for using the product and that they seek to match their goals with the products' functions. This is the notion of human-computer interaction that Norman (1988) described in his seven-stage action model. We further assume that interaction between user and product proceeds through a series of states, i.e., as hypothesized by Card et al. (1983). At each state, the user interrogates the system image for a correspondence between goal and function. Ideally, the goal and function would match and the action would be obvious. In the ergonomics literature, such matching is known as 'stimulus-response compatibility', e.g., turning a control knob clockwise causes a pointer to move to the right on a linear scale. People appear to have well developed stereotypes for some forms of stimulus-response compatibiliy and to base their actions on these stereotypes. Thus, at one level of behaviour users can simply match the system image with a stereotyped response. Such stereotyped responses can be thought of as *Global Prototypical Routines*. Errors can arise when users mistakenly match an object in the system image with their goal, or when a strong stereotype over-rides the correct action.

One rarely achieves the overall goal with a single action, and users need to keep track of their position in a sequence of goals (as proposed by GOMS of Card et al., 1983). Further, there may be occasions when there does not appear to be a clear match between goal and function. At this stage, users need to engage in problem-solving, and to infer the appropriate action on the basis of the system image. Rather than carrying a representation of the product throughout the interaction, users only require information when they are unsure of the appropriate action. The 'routine' represents a set of actions (or a single action) that is deemed appropriate for a given state. These can be thought of as *State-Specific Routines*. Interpretation of the system image in terms of the current goal state might draw on knowledge related to other products, i.e., through analogy or metaphor, in order to infer an appropriate action. Once the action has been performed, then the knowledge is no longer required. In this way, the routines are 'rewritable' in that they can be overwritten by subsequent information. Thus, users might invoke one or two pieces of information from long-term memory (using metaphor or analogy) in order to determine an appropriate action. However, in order to minimize working memory load, the users will rarely need to maintain this information throughout the interaction, and will concentrate on monitoring their progress towards the goal.

It is assumed that the majority of routines that one uses will be local. This is because movement through states in human-machine interaction is punctuated by brief periods in which the machine responds to user actions. A further example of local routines relates to work previously reported on ticket-vending machines (Baber and Stanton, 1996). One common error that we observed in our investigations of ticket-vending machines on the

London Underground was that users would attempt to insert their money as the first stage (this 'error' is further supported by the prevalence of labels on these machines informing users 'Do not put your money in first'). It is proposed that this error is the result of a conflict in two local routines. In one local routine, users follow a 'Vending machine routine' (i.e., insert money; make selection; retrieve item), and in another local routine users follow a 'Ticket-kiosk routine' (i.e., request ticket; pay money; collect ticket). In would appear that the designers had assumed that latter routine, when users often follow the former.

From this perspective, users will employ multiple routines to determine relevant action as and when required. This suggests that users neither need nor use mental models of the product being used. Rather, users seek to draw on previous experience to infer actions only when necessary. This further suggests that basing a design on a single metaphor (or even on a collection of metaphors) need not be useful; users will have different experiences and so need not appreciate the relevance of the supplied metaphor. Consequently, a general design proposal is that the 'exits' from each state in a transaction need to be clearly marked, need to be clearly related to potential user goals and need to be kept to a minimum (so as not to overload or challenge problem-solving abilities). In this way, the description would seek to consider knowledge-in-the-world, knowledge-in-the-user's-head, and knowledge-in-the-context.

While the ideas presented in this section are by no means radical, they have led us to propose that there is a need to represent the interaction between user and product in a manner that makes it easy to consider these ideas during initial design activity. Furthermore, we want our approach to produce quantitative data, i.e., time and error that will support early evaluation of products in terms of 'user performance'.

3 Design

Task Analysis for Error Identification (TAFEI) was developed to allow designers to investigate products in terms of their potential for human error (Baber and Stanton, 1994, 1996, 1999). The approach describes goal-directed user performance in relatively simple systems. The approach is based on state-transition diagrams, and considers potential user errors in terms of transitions between states. Thus, TAFEI could be used as an adjunct to the storyboarding and state-transition diagramming that designers currently use.

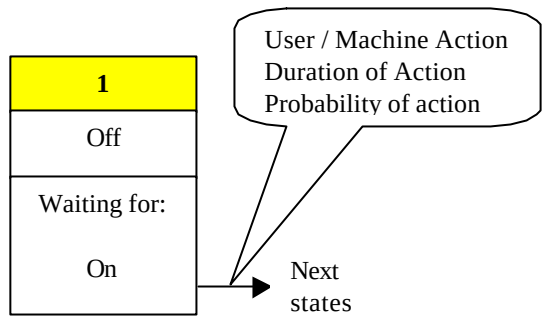

Figure 2: Object in TAFEI diagram

The original method was based on a combined description of user activity and the product (through state-space diagrams). Transitions between states were caused by user or machine action. In the modified version, we have included a duration for each action and a probability that the action will be performed (see figure 2). As we shall discuss below, the probability of an action occurring can be related to the notion of rewritable routines in that the evaluation can ask what would make a user press a given key in error and to assign a probability of occurrence to this action.

4 Evaluation

We have had some success with error prediction, and have been able to demonstrate that TAFEI can predict some 70-80% of errors that are observed in subsequent user trials (Baber and Stanton, 1996).

The description of the user's interaction with a product is described through state-transition diagrams. Each state is considered in terms of the user's current goal, and the match between system image and user goal is explored. .

4.1 Predicting Transaction Time

The idea that one can predict user performance has been popular in human-computer interaction since to publication of Card, Moran and Newell's (1983) keystroke-level model. Recent years have seen an interest in describing parallel activity through critical path analysis (Gray et al., 1993) and a growing interest in the use of Fitts law to predict movement time. Thus, one could argue that there is a well-established tradition in HCI for the development of

models that can predict transaction time, i.e., the total time that it will take a user to perform a task or set of tasks. Furthermore, it is generally accepted that such methods can predict transaction time within around 20% of average observed times (Card et al., 1983; Olson and Olson, 1990).

Keystroke level models were originally developed to address human interaction with computers (Card et al., 1983). Consequently, it is possible that the predicted times might need to be modified when applied to personal technology. However, if one considers the mobile telephone as an example of personal technology, then it is clear that many of the tasks described in KLM can also be applied to such artefacts. For example, the dominant mode of entering data is through buttons and often the main form of dialogue is selection of items from menus. Consequently, it is proposed that times from KLM can be applied directly to personal technology; at least until subsequent research has established a reliable data-set for such products.

Silfverberg et al. (2000) consider text entry speeds on a 12-key keypad that is commonly used on mobile telephones. They used Fitts law to describe the movement time required for entering digits using either thumb or forefinger. The equation used in their paper was: $MT = a+b.\log_2(A/W+1)$. For the handset used in their studies, the key were 10mm x 6mm, and interkey distances ranged from 9mm to 38mm (with a mean of 20.6mm). From user trials, they proposed the following values for a and b: Thumb (a = 176; b = 64); forefinger (a= 165; b = 52). Using these values produced correlations in excess of 0.95. Overall, the average movement time between keypresses was 273ms for the forefinger and 309ms for the thumb. This is reasonably consistent with the 'standard times' of keystroke level models.Using the data in table one, it is possible to develop predictions of transaction time for activity. For example, to select the third item from an unfamiliar menu, one would take 3(314+320) to read the screen and 265ms to select the item, i.e. 3(314+320) + 3(265) = 2697ms, but to select the same item when it was known and familiar would take 3(265) = 795ms. Thus, one might anticipate the 'expert' to be some 70% faster than the 'novice', in this instance.

Unit-task	*time*	*Source*
Press key (forefinger)	273ms	Silfverberg et al., 2000
Press key (thumb)	309ms	Silfverberg et al., 2000
Select from menu	265ms	Fitts law[i]
Read item on screen	314ms	Olson and Olson, 1990
Scan to next item	320ms	Olson and Olson, 1990
Recall command	990ms	Olson and Olson, 1990

Table 1: Standard times for Unit-tasks

4.2 Predicting User Error

In addition to measuring time, it is often useful to measure error in human performance. Indeed, Newell and Card (1985) have argued that "…error categories must be integrated into the theories that describe performance and learning." (p.231). It is somewhat surprising that there are very few approaches to modelling user performance that consider human error. This is partly because KLM (and its derivatives) assumed error-free performance on the part of the user. It is also based on the problem of defining and predicting user error; one could simply follow Murphy's law and say that, if anything can go wrong, it will. Early work on error in HCI tended to focus on defining taxonomies of error and applying these taxonomies to production rule description of performance (Green et al., 1985). More recently, Young and Whittington (1990) demonstrated that one could develop a Programmable User Model that can make errors based on misinterpretation of 'local' knowledge. This suggested that the information presented to the user could lead to difficulty in deciding the appropriate action to take.

[i] For selecting an item in a menu, the display screen is relatively small and the selection of items is typically by means of a joystick. This could be described by Fitts law, using values of a and b for a joystick and A = 5mm and W = 20mm to give 265ms.

From the ergonomics literature, there have been various attempts to develop probabilities of human errors (Kirwan, 1992). One approach that has proved popular is HEART, which uses the probabilities of types of error given generic contextual factors. Table three presents some of the values that might be used in this approach. In the revised version of TAFEI, the 'illegal' action is assigned a probability of occurrence These probabilities are derived from HEART but reflect the level of match between user goal and illegal action. Each action is assigned a standard time (see section 4.1).

4.3 Combining Time and Error Data

From the TAFEI diagram, the transition probabilities and the standard times, it is possible to construct a network diagram. Using MicroSaint[ii], the network can be run as a Monte Carlo simulation. We have successfully employed MicroSAINT to describe speech-based interaction with computers (Hone and Baber, 1999). The output from MicroSaint provides an average time, and a frequency distribution of actions, i.e., in this case, deviations from the 'legal' path (errors). These time and error data form the basis for comparative evaluation.

5 Example

In this example, the simple act of setting a personal CD player on Repeat Play (i.e., continuously playing a single track) is considered for novice and expert users. One might anticipate that a novice user would be more likely to make errors than an expert (although we would assume that the expert could still make some errors). Furthermore, we assume that experts do not necessarily perform unit-tasks faster than novices. This is an assumption in GOMS but the skills literature does not tend to support this; experts are not necessarily able to perform individual tasks faster but are able to combine tasks more efficiently, so that overall the performance is faster.

One way of considering efficiency is in terms of potential for error. Thus, we propose that experts are less likely to make errors, which leads to less need for revision of action and hence, faster performance. Assuming that the CD is inserted in the player, the sequence of tasks required is: Recall plan, Select key, Press key, Read display, Press key. A TAFEI analysis (see figure four for an extract of the diagram) suggested that the CD player under consideration suffered from the following problems: No indication of which key to press; Poor labelling of Play Mode key; Play-Mode key cycles through five states: Normal – Intro – Single music – Program – Random (back to normal). Thus, in terms of user error we can make the following assumptions: (i.) 'Press Play' might be omitted from the user's plan and the user could press the 'Playmode' first; (ii.) Playmode cycles through all five modes to return to Normal; the user could not realise this and select the wrong mode; (iii.) There is no 'undo' and the only solution is to switch the CD player off and on and start again.

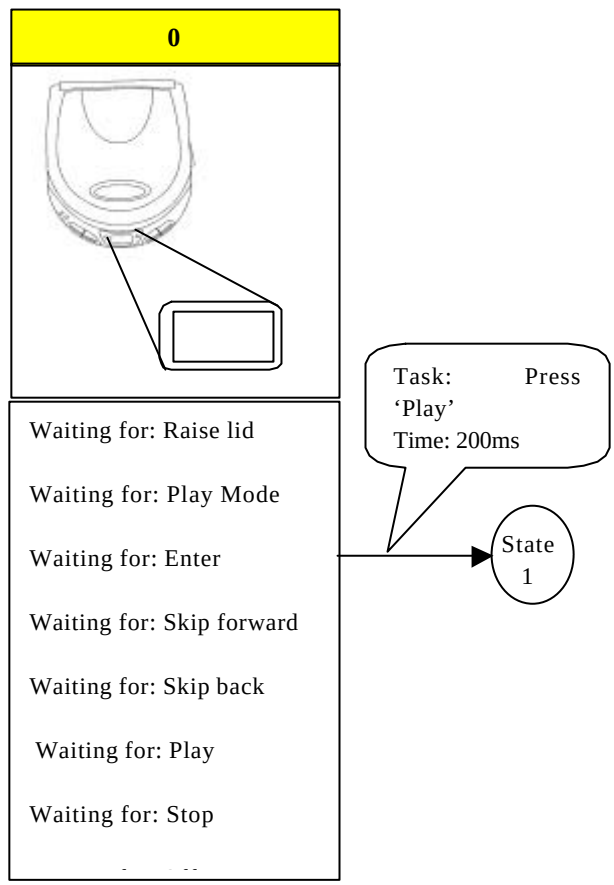

Figure 4: State 0 for CD TAFEI

[ii] MicroSAINT is a computer-simulation package, from Micro Analysis and Design Simulation Software Inc. that uses Monte Carlo routines to model a state-transition description of a system.

Unit-Times:	Probability of failure:	Novice	Expert
Recall plan: 1380ms	Shift system to new state on single attempt:	0.26	0.003
Select: 360ms	Routine task with low level of skill:	0.02	0.0004
Press: 200ms	Completely familiar task:	0.0004	0.0004
Read: 180ms	Complex task with scant attention:	0.16	0.09

Table 3: Data used in Model

We assume that the expert will be less likely to make errors than the novice, but that the times for unit-tasks will be fairly similar. Thus, novices will have a higher probability of failure when they first encounter the product. We ask will novice performance be worse than expert performance?

In order to arrive at the probabilities for actions for novice and expert users, we applied a simple heuristic: if there was a single, well-illustrated button, e.g., Play, that could be pressed to perform a task, then the probability of pressing that button was high, but if the choice of button was problematic, e.g., due to lack of marking, unclear instructions or ambiguous plan, then the probability of pressing the button was reduced. The degree of reduction was assumed to be greater for the novice than the expert. A number of assumptions can be explicitly stated for this interaction, e.g., if the user employs the wrong plan, this is a mistake, but if the user presses 'playmode' first, then this is a slip. The implications for these errors relate to the movement between states. It is also assumed that errors are detected and escaped from, e.g., playmode leads to cycle through menu back to 'play' then resume with press play; no slips on play / playmode / play (for space, but will have similar consequences to those shown, and are included in model).

One might disagree with these assumptions, but it is not difficult to modify the model in response to other assumptions. This model is then created using the MicroSaint software. The simulation was run 1000 times in order and the results are shown in figure 5.

The expert model assumed that tasks were performed in series and that there was a minimal probability of error between each task. The novice model used the same unit-times but assumed a higher level of error. One can see that the average times for the experts and novice vary: the mean time for the Expert is 1681.33 (±632) ms and the mean time for Novice was 2626.32 (±1743.4) ms. Thus, the model can differentiate between skill levels by making assumptions concerning time and error. By way of validation, the owner of the CD player performed this task 10 times and averaged just over 1.5s (which is comparable to the 'expert' time produced from the model).

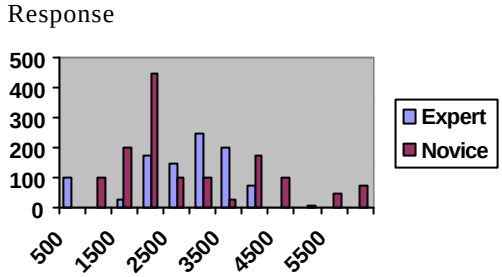

Figure 5: Results from MicroSaint model

6 Discussion

This paper has presented a theoretically-motivated approach to evaluating conceptual designs of products. The paper presents a complete approach to analytical prototyping of personal technologies. The approach can be applied during the initial stages of product design, e.g., when storyboards and initial interface concepts are being discussed. The theory of rewritable routines, drawing on mental models literature, combined with state-space diagrams as a means of representing human-product interaction represents a useful means of conducting functional and scenario-based analysis. Models of user performance provides us with a means of structural analysis of concepts. Time and error data allows us to very quickly generate data in order to compare design alternatives, and so inform early designs from a human factors point of view.

References

Baber, C. and Stanton, N.A., 1994, Task analysis for error identification, *Ergonomics 37* 1923-

1942

Baber, C. and Stanton, N.A., 1996, Human error identification techniques applied to public technology: predictions compared with observed use, *Applied Ergonomics, 27* 119-131

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

Gray, W.D., John, B.E. and Atwood, M.E. (1993) Project Ernestine: validating a GOMS analysis for predicting and explaining real-world performance *Human-Computer Interaction 8* 237-309

Green, T.R.G., Payne, S.J., Gilmore, D.J. and Mepham, M., 1985, Predicting expert slips, *Interact'85*, Amsterdam: North-Holland, 519-525

Hone, K.S. and Baber, C., 1999, Modelling the effect of constraint on speech-based human computer interaction *International Journal of Human Computer Studies 50 (1)* 85-113

Johnson-Laird, P.N., 1983, *Mental Models*, Cambridge: Cambridge University Press

Kiljander, H., 1999, User interface prototyping methods in designing mobile handsets, In M.A. Sasse and C. Johnson (eds) *Interact'99*, Amsterdam: IOS Press, 118-125

Kirwan, B.I., 1992, *Human Reliability Assessment* London: Taylor and Francis

Newell, A. and Card, S.K., 1985, The prospects for psychological science in human-computer interaction, *Human Computer Interaction, 1* 209-242

Norman, D.A., 1988, *The Psychology of Everyday Things*, New York: Basic Books

Olson, J.R. and Olson, G.M.,1990, The growth of cognitive modelling in human-computer interaction since GOMS *Human-Computer Interaction 3* 309-350

O'Malley, C. and Draper, S.,1992, Representation and interaction: are mental models all in the mind?, In Y. Rogers, A. Rutherford and P.A. Bibby (eds), *Models in the Mind: theory, perspective and application*, London: Academic Press, 73-91

Silfverberg, M., MacKenzie, I.S and Korhonen, P., 2000, Predicting text entry speed on mobile phones, *CHI'2000*, New York: ACM, 9-16

Wichansky, A.M., 2000, Usability testing in 2000

and beyond, *Ergonomics 43 (7)* 998-1006

Young, R.M. and Whittington, J., 1990, Using a knowledge analysis to predict conceptual errors in text-editor usage, *CHI'90*, New York: ACM, 91-96