Building predictive human performance models of skill acquisition in a data entry task

Wai-Tat Fu (wfu@uiuc.edu)
Human Factors Division and Beckman Institue
University of Illinois at Urbana-Champaign
1 Airport Road, Savoy, IL 61874

Cleotilde Gonzalez (conzalez@andrew.cmu.edu)
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Alice F. Healy, James A. Kole, and Lyle E. Bourne, Jr.
(ahealy@psych.colorado.edu)
Psychology Department, University of Colorado
Boulder, CO 80309-0345

Abstract

This paper presents a predictive model of a simple, but important, data entry task. The task requires participants to perceive and encode information on the screen, locate the corresponding keys for the information on different layouts of the keyboard, and enter the information. Since data entry is a central component in most human-machine interaction, a predictive model of performance will provide useful information that informs interface design and effectiveness of training. We created a cognitive model of the data entry task based on the ACT-R 5.0 architecture. The same model provided good fits to three existing data sets, which demonstrated the effects of fatigue with prolonged work, repetition priming, depth of processing, and the suppression of subvocal rehearsal. The model also makes predictions on how performance deteriorates with different delays after training, how different amounts of rehearsal during training affect retention, and how re-training helps retention of skills.

Introduction

In this paper we present cognitive models constructed to predict the effectiveness of specific training principles in perceptual, cognitive, and motor tasks. We used empirical data and theory developed at the University of Colorado in three main experiments. We constructed executable models that represent the theory including the cognitive processes and mechanisms we expect to be involved. From these models we obtained data that we compared to the human data obtained from the experiments. After our models were validated using the procedure just described, we generated a set of new predictions for future experiments.

The Data Entry Task

The data entry task has been studied extensively in the laboratory and is routinely used outside the laboratory in a variety of complex, naturalistic situations. In this task, subjects are typically shown sets of four-digit numbers, either as numerals (e.g., 4 8 2 6) or as words (four eight two six). Subjects respond by typing each of the four digits using either the keypad on the right-hand side of the computer keyboard or the number row on the top of the keyboard. In some cases, they respond instead by typing the initial letters of each word (e.g., f e t s). Typically, no feedback concerning the accuracy of the responses is provided to the subjects, and they do not see their typed responses. There are three major component-processing stages in the data-entry task: encoding, response preparation, and response execution. Encoding involves perceptual processes, response preparation involves the mental construction of a motor program for entering the sequence, and response execution involves the actual motoric button presses.

The Model

A general structure of the ACT-R (Anderson et al., 2004) models used in the next three data entry experiments is represented in Figure 1. We divided the task into the cognitive steps represented in this figure: encode each of the 4 digits; retrieve the location of the key for each of the digits; type each of the 4 digits and hit the “enter” key. Each of these steps involved at least one production in the ACT-R model.

Figure 1 shows that the task can be decomposed into three different components. The first is initiation time, which measures the time to enter the first digit. During the intiation time, participants had to encode the four digits, find the corresponding key locations on the keypad, and execute the first keypress. Thus, the initiation time includes both cognitve and motoric processes. The second is execution time, which measure the time to enter the second, third, and fourth digits. During the execution time, participants presumably engaged in less encoding, so the time represented primarily finding the corresponding key

locations on the keypad and executing the keypresses. The third is conclusion time, which measures the time to enter the concluding "Enter" keystroke. During the conclusion time, participants had to only find and press the key, which is largely a motoric process. We will show later, since different cognitive and motoric processes are involved in these component measures, this breakdown of response times will provide important information for the underlying cognitive and motoric processes in this task.

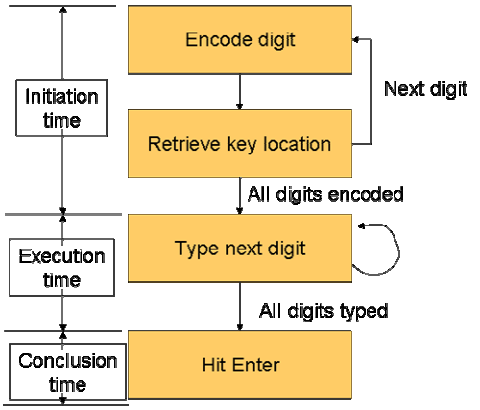

Figure 1: Steps in the ACT-R cognitive models of the data entry task

ACT-R is a computational cognitive architecture used to simulate the cognitive processes involved in human performance of a task. There are multiple representations of ACT-R modules and capabilities (see ACT-R 4.0 representation in Anderson & Lebiere, 1998, and ACT-R 5.0 representation in Anderson et al., 2004). There are two major kinds of knowledge representation in the ACT-R architecture: declarative and procedural knowledge. ACT-R's declarative knowledge involves a representation of basic units of knowledge, or chunks. These represent facts described by slots or attributes. Procedural knowledge in ACT-R consists of production rules represented as condition-actions pairs. We will describe the two major mechanisms in ACT-R that are relevant to the current task.

The first mechanism concerns the access of declarative knowledge, or chunks. ACT-R keeps track of the usefulness of the knowledge and the use of a chunk is determined by its activation level, as described in the following activation equation:

$$A_i = B_i + \sum_j W_j S_{ji}$$

where Bi is the base-level activation of the chunk i, the Wj are the attention weightings of the elements that are part of the current goal, and the Sji are the strengths of association from the elements j to chunk i. The major mechanism relevant to the current task is the base-level activation equation that governs the value of Bi:

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d})$$

where tj is the time since the jth practice of an item. This equation reflects the findings that access of declarative knowledge increases with each exposure as a power function (producing the power law of practice), otherwise access to knowledge decays exponentially.

At any point in time, the productions representing procedural knowledge detect the patterns in declarative knowledge that are actively being processed by the system (usually relevant to the current goal). The key idea is that at any point in time multiple production rules might apply, but because of the seriality in production rule execution, only one can be selected, and this is the one with the highest utility. Production rule utilities are like activations for chunks, and play a similar role as activations play in chunk selection.

The mechanism most relevant to the current task for procedural knowledge is called production compilation, which basically is a combination of composition and proceduralization as described in Anderson's (1983) theory of skill acquisition. Production compilation will try to take each successive pair of productions and build a single production that has the effect of both. After a production New is composed from productions Old1 and Old2, whenever New can apply, Old1 can also apply. The choice between New, Old1, and whatever other productions might apply will be determined by their utilities. However, the new production New has no prior experience, and so, its initial utilities will be determined by Bayesian priors. We describe how the prior θ is set for the utility value U. When New is first created, θ is set to be 0. Thus, there is no chance that the production will be selected. However, whenever it is recreated, its θ value is incremented according to the delta rule: $\Delta\theta = a(U - \theta)$, where U is the utility of Old1. Eventually, if the production rule New is repeatedly created, its priori θ will converge on U for the parent Old1. When it is actually superior, it will come to dominate its parent. Although our experience with this production rule learning mechanism is relatively limited, it seems to work well with the reinforcement-learning mechanism for production systems (Fu & Anderson, 2006).

We will next describe how our model produces performance of the data entry task in three different settings, as described in three different existing data sets. We will then describe how we can use our model to make predictions of performance in other novel situations.

Effects of prolonged work

When people work continuously over time on a task, two opposing processes might affect their performance. As predicted by the law of practice, performance may improve by becoming more accurate, faster, or both. On the other hand, performance may deteriorate as they suffer the effects of fatigue, boredom, and diminished attention over prolonged periods. In addition, it is possible that practice and fatigue may affect various measures of performance (e.g., speed, accuracy, and different components of a complex response) in different ways. The goal of the

experiments by Healy, Kole, Buck-Gengler, and Bourne (2004) was to encourage fatigue in a data entry task and to examine its effects on accuracy and on response time over a long practice period. As shown in Figure 2, Healy et al. (2004) found that prolonged work produced both learning and fatigue-like effects, depending on which measure, speed or accuracy, was used. With prolonged work, error rate increased (i.e., proportion correct decreased). This result suggested deterioration in performance as a result of fatigue. When an analysis was done for four different types of errors: missed trial, missed digit, extra digit, and wrong digit, it was found the trials with wrong or extra digits were more common than those involving missing digits or missed trials. These errors involving wrong or extra digits increased dramatically across the session.

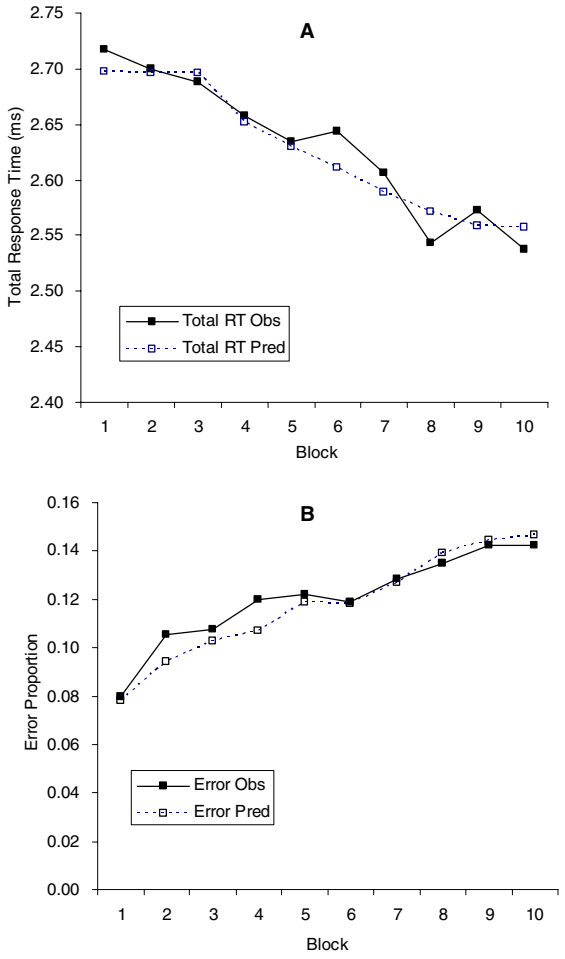

Figure 2. The observed (Obs) and predicted (Pred) total response time (A) and error proportion (B) in Experiment 1 of Healy et al. (2004).

As shown in Figure 2, total response time decreased over time. This decrease contrasts with the increase in error rate and points to a speed-accuracy tradeoff. The model provided good fits to the data. We obtained a fit of R^2=.96, RMSE= 0.23, and R^2=.91, RMSE=0.14 for the RT and accuracy respectively. The model produced these results as follows. Production compilation, which replaces the encoding and retrieving of the key locations by a "macro" production, results in faster execution. Fatigue corresponds to the increase of activation noise in retrievals, which leads to more errors (i.e., a decrease in the proportion of correct responses).

Repetition priming and depth of processing

Buck-Gengler and Healy (2001) asked subjects to enter the 4-digit numbers displayed as either words or numerals. At test one week later, half of the old numbers from each group were presented in the same format as at training, and the other half were presented in the alternate format. They expected that the abstract concept would contribute to repetition priming. To separate the effect of learning in the motoric component, subjects were trained with one key configuration (keypad or row) and tested with a different one. Buck-Gengler and Healy found that old numbers were typed faster than new numbers (there was repetition priming) independently from the motor component. Also, numbers in word format at training were entered faster (as words or numbers) at test than numbers coded as numerals at training (see Figure 3).

In our model, because old numbers had a stronger base level activation (see the base-level activation equation), the retrieval was faster, thus creating the difference between old and new numbers. When numbers were presented as words, during the encoding process, the episodic representation of the word-format numbers invoked more phonological rehearsals of the numbers represented by the words. Both the episodic and semantic representations were encoded together with the key locations. During retrieval, both representations acted as sources of activation that speeded up the retrieval of the key locations. On the other hand, when numbers were presented as numerals, semantic concepts were not encoded and thus led to a lower level of activation due to the episodic representation. The difference in source activation created the faster response for numbers presented in word format.

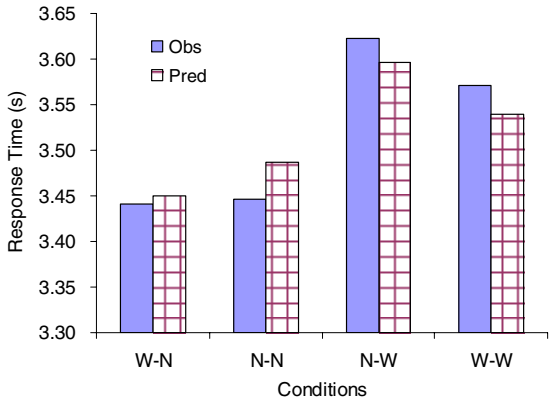

Figure 3. The observed (Obs) and predicted (Pred) response time in different conditions. N-W = Numeral at training, Word at testing; N-N = Numeral at training and testing, etc.

Figure 3 shows the reaction times for different conditions of Buck-Gengler's experiment. The model fits the data well (R^2=.92, RMSE=0.89).

Suppression of subvocal rehearsal

An explanation for the results from the second data set that was not answered by Buck-Gengler and Healy (2001) is that words may be more likely than numerals to elicit phonological processing. That is, it may not be that abstract meaning is retained better for numbers displayed as words than for numbers displayed as numerals. Rather, it is possible that individuals may simply activate the phonological loop of working memory. If this hypothesis is true, articulatory suppression would disrupt this means of coding and thus alter the performance in the task.

In the study by Kole, Healy, and Buck-Gengler (2005) half of the subjects were in an articulatory suppression condition. Subjects in this group were required to repeat the word "the" continuously, starting before the first trial. The subjects in the other group were silent while they entered the digits. The main findings from their experiment were that old numbers were typed faster than new numbers (repetition priming holds), but the advantage of numbers presented as words at training was significant only for the silent group.

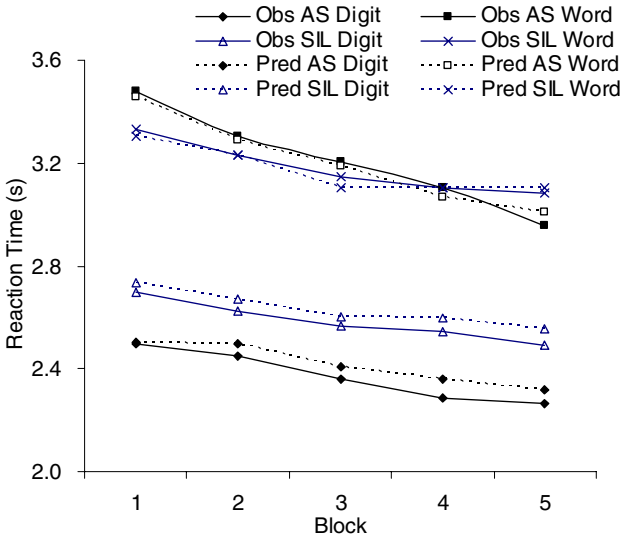

Figure 4. The observed (Obs) and predicted (Pred) response time for the different conditions in Kole et al. (2005) during training. AS=articulatory suppression, SIL=silence.

The ACT-R model helped extend the explanation of the results and understand what might be happening at the level of cognitive processing. First, through modeling we found that the long-term repetition priming can be explained by the spreading activation mechanism in ACT-R. The results from the digit and letter conditions in the Buck-Gengler and Healy (2001) study suggest that it is not the amount but the type of processing that leads to the long-term repetition priming effect. We modified the model so that in the articulatory suppression condition, no encoding of the key location together with the numbers was done. During testing, since there was no chunk representing both the key locations and the numbers, there was no spread of activation to the key location and thus the retrieval was slower, compared to the silence condition. The results from the cognitive model and fit to the training data are shown in Figure 4. We obtained a fit of R^2=0.88, RMSE=1.4.

Model Predictions

We aimed to answer three questions of general interest that had not been answered through empirical data collection in the data entry task. These questions are: 1) How would performance deteriorate with different delays after training?; 2) How would different amounts of immediate rehearsal during training affect the retention of skills?; and 3) How would re-training help retention of skills?

The delay was manipulated as the number of days between the end of training and the beginning of the testing phase. In the original experiments there was a delay of 7 days. Figure 5 shows the predictions of the depth of processing and repetition priming effects when the duration between training and test varies from 1 to 16 days. The RT difference is the difference between the numeral and word RT for the depth of processing effect and the difference between the new number and the old number RT for the repetition priming effect. These predictions indicate that the benefit of training in words disappears with a longer delay between training and testing. Also, the repetition priming effect decays exponentially after a long delay.

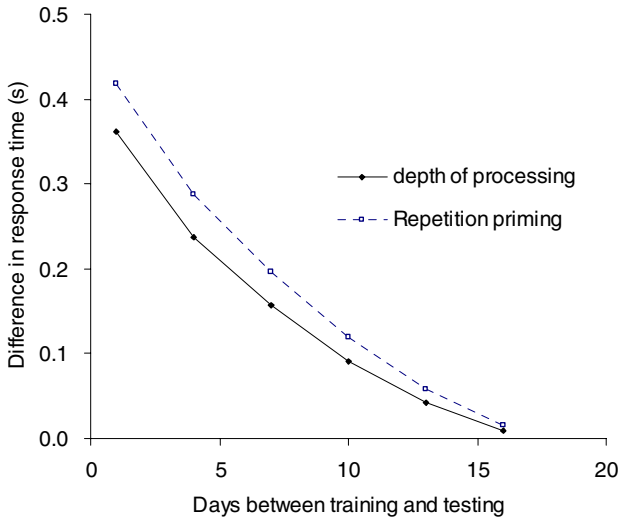

Figure 5. The predictions of the differences in response time for the depth of processing and repetition priming effects

We are also interested in predicting how the depth of processing and repetition priming effects will change across time with different initial amounts of training. We therefore manipulated the number of rehearsals of a word from 1 to 3 times. Figure 6 presents these predictions. Compared to the predictions in Figure 5, we see that more initial training leads to higher retention of skills, but the effect decays

rapidly. However, the benefit of initial training stays even after 16 days.

Finally, we wanted to predict the effect of repetition: when individuals are trained and retrained after a particular time period. We varied the training from zero to two times. We found (Figure 7) that re-training may be more efficient than extensive initial training for retention of skills.

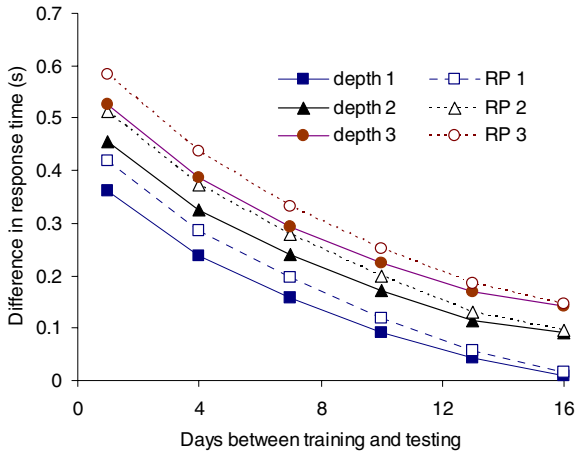

Figure 6. The predictions of the differences in response time for the depth of processing and repetition priming effects. Depth 1 = depth of processing effect with 1 rehearsal; RP1 = repetition priming effect with 1 rehearsal, etc.

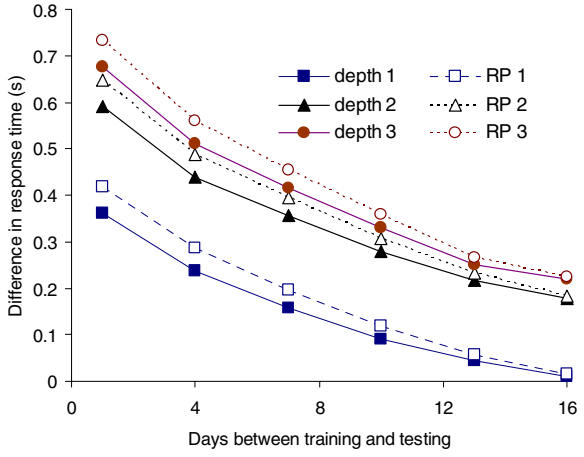

Figure 7. The predictions of the differences in response time for the depth of processing and repetition priming effects. Depth 1 = depth of processing effect with 1 retraining, RP1 = repetition priming effect with 1 retraining, etc.

Discussions

For decades, human factors practitioners have been calling for predictive models of human performance. One of the most prominent calls was made by Newell and Card (1985), who warned the human factors community it was not through advocating the empirical testing of endless design alternatives but, rather, through the use of predictive and reliable quantitative techniques. As the scope and scale of the issues that the human factors community was asked to consider expanded, the tool chest of quantitative methods seemed to diminish. With the recent advance of architectures such as ACT-R, or SOAR, engineering quantitative models of human performance is the wave of the present and represents an important part of the future of the human factors profession.

The fact that we used the same model to fit the data is interesting in itself, because it demonstrates the flexibility and accuracy of the human cognition represented in the models. Through this exercise we were able to extend the explanations that researchers offered in the original papers. Most importantly, we were also able to make predictions of general interest using the model. The assessment of retention following different delay intervals and of transfer after changing the contexts are two general goals of effective training. Although this is only a first step towards our overriding goal of developing a theoretical framework for predicting the effectiveness of different training methods, we feel this is an important step. Given the good data fits in the three experimental data sets, we feel confident that the predictions offered here are reasonably close to actual human performance.

Acknowledgments

This research was supported by the Multidisciplinary University Research Initiative grant from the Army Research Office (W911NF-05-1-0153). We thank Carolyn Buck-Gengler for providing us the raw data reported in the Buck-Gengler and Healy (2001) study.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111*, 1036-1060.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates

Buck-Gengler, C. J., & Healy, A. F. (2001). Proccesses underlying long-term repetition priming in digit data entry. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 879-888.

Fu, W.-T. & Anderson, J. R. (2006). From recurrent choice to skilled learning: A reinforcement learning model. *Journal of Experimental Psychology: General*, 2.

Healy, A. F., Kole, J. A., Buck-Gengler, C. J., & Bourne, L. E. J. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied, 10*, 188-199.

Kole, J. A., Healy, A. F., & Buck-Gengler, C. J. (2005). Does number data entry rely on the phonological loop. *Memory, 13*, 388-394.

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