Validating Quantitative Models of User Performance with Word Prediction Systems

Heidi Horstmann Koester and Simon P. Levine
Rehabilitation Engineering Program, Graduate Bioengineering Program
University of Michigan

Abstract
Three model implementations were developed to simulate user performance with a word prediction system. Model simulations were tested against the actual performance of able-bodied and spinal cord injured subjects. Models 1A and 1B shared a structure which represented performance as a linear combination of two user parameters, keypress and list search time, while Model 2 used a revised model for list search time. For Model 1A, user parameter values were determined independently, while Models 1B and 2 used parameter values derived from subjects' data. The average errors for Models 1A, 1B, and 2 in simulating subjects' word entry times were 27%, 16%, and 14%, respectively.

Introduction
Models of user performance with AAC and computer access systems have the potential to help practitioners design, configure, and recommend more appropriate systems to meet users' needs. A key to realizing this potential is understanding the accuracy obtainable with quantitative models. This paper reports on the development and validation of quantitative models of user performance with word prediction systems.

The models developed in this work are based on the Keystroke-Level Model, a technique which originated in the field of human-computer interaction [1,2]. This technique provides a means of identifying the component actions that a user must perform for a particular task. By attaching times to these activities, the overall task time can be predicted.

Within the Keystroke-Level Model technique, there are a variety of specific methods. "Keystroke-Level" means that the model represents events in the range of 100 ms to a few seconds. The way in which the overall task is broken down into these short duration component events is the *model structure*, and is a choice of the modeler. A second dimension on which models can vary is the source of *user parameter values*. Each model structure yields a parametric equation for task performance time. Parameters representing the user can be drawn from independent sources, such as previous human performance studies, or they can be derived directly from subjects' performance data.

Independent models are easier to apply and potentially more generalizable than data-driven models, but they are typically less accurate. For tasks such as text editing and spreadsheet use, errors reported for models with independent structure and parameters average 52.2%, with a range of 24 - 137%. For models whose structure and/or parameters have been derived from subject data, average error is 16.6%, with a range of 4 - 33% [3]. To begin to assess the suitability of models for AAC applications, we need to measure the accuracy for a representative task under different model implementations. The three model implementations used in this work are:

- Model 1A. Structure and parameters are independent of validation data.
- Model 1B. Same structure as Model 1A; user parameters are derived from validation data.
- Model 2. Structure and parameters are derived from validation data.

Methods
Empirical Data for Model Validation. Fourteen subjects transcribed text with and without a word prediction feature for seven test sessions. Text was unique in each session. Eight subjects were able-bodied and used mouthstick typing. Six subjects had high level spinal cord injuries and used their usual method of keyboard access. Two of the spinal cord injured subjects used a mouthstick for typing, and the other four used hand splints. Each subject was assigned to use one of two word prediction strategies, to provide a behavioral basis for the model structure. The rule for Strategy 1 was to search the list before every selection. The rule for Strategy 2 was to choose the first two letters of a word, then search the list before each subsequent selection.

Dependent measures of subject performance included overall text generation rate, time to enter each individual word in a session, and time to select each item. Model simulations of any of these measures are possible. Overall session performance is the most clinically relevant measure, but simulations across an entire session may capitalize on cancellation of positive and negative errors, leading to an overly optimistic estimate of model accuracy. This paper focuses on simulations of word entry times, to reduce the effect of fortuitous error cancellation while maintaining some clinical relevance.

Structure of Models 1A and 1B. The structure of Model 1 represented the task of text entry as a simple sequence of keypresses and list searches. The time for each item selected in a session was modeled as a keypress-only (t_k) or a list search-plus-keypress (t_s + t_k), depending on the strategy rules followed by the subject. Keypress time was assumed to be independent of the key's position, and list search time was assumed to be independent of the search context. An equation for each word's entry time was formed by summing the model times for each selection made for the word. This process was performed for each word entered by each subject in every test session.

Structure of Model 2. With Model 2, keypress time continued to be represented by a single parameter value, and the representation of list search time was refined. The general strategy used for the refinement was to identify candidate factors that might influence list search time, determine the relative strength of each of these factors using regression analysis, and create a new model of list search time to reflect the influence of the strongest factors [3].

Two factors examined for their influence on list search time were a serial search process and anticipation of list contents. Variables reflecting the context of the search were chosen to represent these factors. For example, variables related to serial search were the position of the target word and the length of the list, for successful and unsuccessful searches respectively. The strength of the candidate predictor variables was evaluated by a series of multivariate and bivariate regression analyses, performed for each subject's set of list search times in Session 4. (Only one session was examined due to the large number of analyses involved.)

User Parameters, Model 1A. For Model 1A, keypress and list search times were determined from previous studies reported in the literature. The best available source for keypress times was a study of able-bodied individuals who typed with mouthsticks for 30 typing tests [4]. The average typing speed for the 30 tests was fit to a Power Law curve, and the resulting fitted values used for keypress time ranged from 1.15 to 1.01 seconds across Session 1 through Session 7.

Independent values for list search time were derived from a composite of studies on list search [3,5]. These studies were used to establish an appropriate value for search times in the 1st and 7th test sessions. A Power Law curve was then fit between the estimates for Sessions 1 and 7 to determine values for the intermediate sessions. The resulting values ranged from 1.0 to 0.5 seconds across Sessions 1 to 7.

User Parameters, Model 1B. For the Model 1B simulations, user parameter values for each subject were derived directly from the subjects' performance data, using the subtractive methods employed in other Keystroke-Level Model studies [1,2]. The keypress time during word prediction use (t_k) was calculated by averaging the times for all keypress-only selections in the session. The list search time (t_s) was derived by subtracting one t_k from the time recorded for each list search-plus-keypress selection, then averaging the resulting time differences. Average keypress times from Session 1 to Session 7 ranged from 1.05 to 0.89 seconds for able-bodied subjects, and from 0.97 to 0.76 seconds for spinal cord injured subjects. Average list search times from Session 1 to Session 7 ranged from 0.66 to 0.47 seconds for able-bodied subjects and from 1.17 to 1.14 seconds for spinal cord injured subjects. (An analysis of the differences between able-bodied and spinal cord injured subjects has been presented elsewhere [5].)

User Parameters, Model 2. The keypress times used for Model 1B were also used for Model 2. New values for list search time were calculated for each list search-plus-keypress selection using the regression equations derived in the revision to the model structure.

Simulations and Validation. Simulations of word entry times were performed for each subject-session combination by substituting the user parameter values into each word's model equation. For Model 2, simulations were performed for Session 4 only. Model error scores were computed for each subject-session combination by averaging the absolute value of the percent difference between the actual and simulated word entry times. Statistical analyses were performed using a repeated measures ANOVA, with between-subjects factors of strategy and presence/absence of spinal cord injury and within-subjects factor of test session. Significance of all statistical tests was judged at a familywise p-value of 0.05.

Results

Figure 1 summarizes the error of each Model in simulating subjects' word entry times, collapsed across the seven test sessions. Average error for able-bodied (AB) and spinal cord injured (SCI) subjects are shown separately, in addition to the average error across all subjects (ALL). Bars indicate one standard deviation.

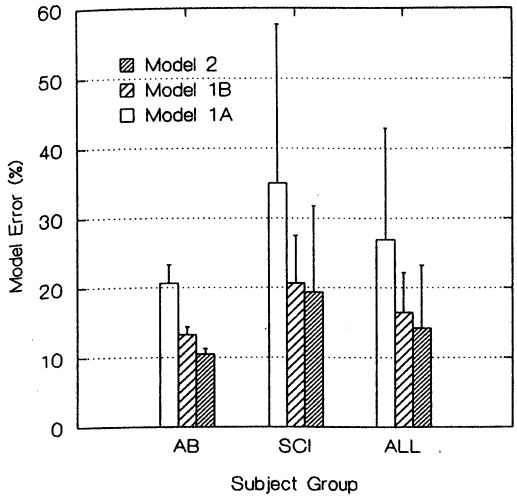

Figure 1. Average error of Models 1A, 1B, and 2, for simulations of word entry times.

For Model 1A, whose simulations were completely independent from subject data, the average error was 26.9% across all subjects. Model error decreased significantly when empirically-derived user parameters were used, as illustrated by the 16.5% average error obtained with Model 1B. Revising the model representation of list search time yielded a small but statistically significant improvement in accuracy, as Model 2 had an average error of 14.3%.

The trends in the results were similar for able-bodied and spinal cord injured groups, but model error was generally higher and more variable for the subjects with SCI as compared to the able-bodied subjects. The difference in model error between subject groups, however, was not statistically significant.

Discussion and Conclusions

The results of these simulations compare very favorably to accuracy reported in previous applications of Keystroke-Level Models, which suggests that the modeling technique is as applicable to AAC and computer access systems as it is to other human-computer interfaces. A companion paper [6] illustrates how such a model can be used to examine performance across a range of usage conditions.

Manipulating the modeling dimensions of structure and user parameters had the expected effects. With respect to model structure, the refined representation of search time in Model 2 did improve model accuracy. However, the amount of improvement was quite small relative to the effort involved in the structural revisions. It is possible that a revised representation of keypress time may have led to additional gains in model accuracy.

With respect to user parameters, empirically-derived values yielded fairly large improvements in model accuracy, particularly for spinal cord injured subjects. This reinforces the intuition that variation between user parameter times must be accounted for in an accurate user performance model, particularly for the AAC user population.

While the overall results were encouraging, the model has been validated on only a relatively small subset of possible users and systems that may be of interest. It will be important in future work to assess the accuracy of the model for users who are word prediction experts and those who have more variable motor control and/or cognitive impairments. Additionally, imposing particular strategy rules on subjects influenced the accuracy obtained, which suggests that these results are most applicable to a clinical situation in which the user consistently employs a particular method of using the list. Research is currently being performed to determine the suitability of this approach for modeling users' natural strategies.

References

1. S. Card, T. Moran, A. Newell. *The Psychology of Human-Computer Interaction.* Hillsdale, NJ: Erlbaum Associates, 1983.
2. J.S. Olson, E. Nilsen. "Analysis of the cognition involved in spreadsheet software interaction," *HCI*, 3:4, 309-350, 1988.
3. H.H. Koester. "User Performance with Augmentative Communication Systems: Measurements and Models," Doctoral Dissertation, University of Michigan, 1994.
4. S.P. Levine et al. "A comparison of mouthstick and Morse code text inputs," *AAC*, 2:2, 51-55, 1986.
5. H.H. Koester, S.P. Levine. "Text entry performance with a word prediction interface," *IEEE Trans. on Rehab. Engin.*, 2:3, 177-187.
6. H.H. Koester, S.P. Levine. "Simulations of user performance with word prediction," submitted to *RESNA '95.*

Acknowledgments

This research was supported by the National Science Foundation, Rackham School of Graduate Studies, and the U-M Rehabilitation Engineering Program.

Heidi Horstmann Koester, Ph.D.
1C335 University Hospital
Ann Arbor MI 48109-0032
Internet: hhk@umich.edu