0743-4618/97/1304-0239 $3.00/0, Volume 13, December 1997
AAC Augmentative and Alternative Communication

Keystroke-Level Models for User Performance with Word Prediction

Heidi Horstmann Koester and Simon P. Levine

Koester Performance Research, Holland, Michigan, USA (H.H.K.) and Rehabilitation Engineering Program, Graduate Bioengineering Program, Department of Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, Michigan, USA (S.P.L.)

Two modeling studies have been performed to develop quantitative models of user performance with word prediction and test their predictions against empirical data. In the first study, the model structure represented performance as a linear combination of two user parameters, keypress and list search time. Two types of simulations were performed using this structure: Model 1A, in which user parameter values were determined independently, and Model 1B, which used parameter values derived from subjects' data. Model simulations of overall session performance, word entry times, and item selection times were compared to actual performance of able-bodied and spinal cord injured subjects transcribing text with and without word prediction over the course of seven test sessions. The average errors for Models 1A and 1B in modeling subjects' word entry times were 27% and 16%, respectively. The second study used a revised model for list search time in an attempt to improve model accuracy and increase understanding of the list search process. The model revision led to only a small improvement in accuracy but did provide insight into how list search time depends on the context of the search. The results point out the need to understand a user's characteristics before applying a model, but they are an encouraging demonstration of the ability of analytical models to represent user performance with word prediction.

KEY WORDS: augmentative communication, group designs, rate enhancement, user modeling, word prediction

There is a strong need for greater understanding of user performance with augmentative and alternative communication (AAC) systems. Progress toward this understanding can be made through empirical studies of actual user performance under a range of conditions. However, given the impossibility of measuring performance under all conditions of interest, an important complement to empirical studies is the creation and validation of analytical models of user performance.

This paper focuses on word prediction systems and addresses the question of whether analytical models of user performance can be developed that accurately represent the performance of actual users (Horstmann & Levine, 1992; Newell, Arnott, & Waller, 1992). Three model implementations were developed. Models 1A and 1B shared a structure that 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. Each was tested against actual performance in a range of experimental conditions to determine which one yields better accuracy and how consistent accuracy is across different users and usage conditions. The short-term goal is an accurate model that will support simulations of user performance with word prediction systems and provide insight into the conditions under which word prediction does and does not lead to improved performance. The longer-term goal is to progress toward establishing modeling techniques that can be used to analyze and optimize a wide range of current AAC systems and the design of future systems for a particular user or user population.

BACKGROUND

Approaches to Modeling of AAC Systems

Quantitative models can be used to simulate and predict performance for different system and user characteristics, which makes them a potentially powerful tool for clinicians and designers. A number of researchers have recognized the significant potential benefits of the modeling approach, and several quantitative models of AAC systems have been developed as a result. Most of these evaluate systems based primarily on motor efficiency (Baker & Barry, 1990; Damper, 1984; Goodenough-Trepagnier & Rosen, 1988; Pollak, 1982; Pollak & Gallagher, 1989; Vanderheiden, 1988), so they are not well suited to represent systems like word prediction, in which cognition and perception have an important impact. The available validation data suggest that models that focus only on motor efficiency may make accurate quantitative predictions for simple letter-based systems (Damper, 1984), but they may seriously overestimate achievable speed when applied to more complex systems (Goodenough-Trepagnier et al., 1987; Goodenough-Trepagnier & Rosen, 1988).

The need to consider cognitive and perceptual abilities in models of user-system interaction is well recognized (Dabbagh & Damper, 1985; Gibler & Childress, 1982; Horstmann & Levine, 1990; Vanderheiden, 1988), but there are still shortcomings in current approaches. Gibler and Childress (1982) produced the most comprehensive attempt when they decomposed the user's selection time for their scanning word prediction system into several components. These were the visual search time (for word lists and letter matrices), cognitive time delay (to make decisions based on visual search or decide on the next word to type), and movement time (for the sequence of switch hits required to make a selection). Cognitive time was subsequently ignored, but visual search time and movement time for a given selection were modeled as linear functions of that selection's position. The resulting rate calculations were accurate to within 7% for the three able-bodied (AB) subjects tested. Model accuracy results were not reported for the single disabled subject tested. A parametric analysis predicted that a dictionary of 1000 words would result in poorer performance than a dictionary of 500 words because the increase in visual search and scanning times for longer word lists would offset the improvement in word prediction success. The validity of this prediction was not tested against empirical data.

The research reviewed above demonstrates the potential power of the modeling approach to address the numerous trade-off issues that face the AAC developer and clinician. However, two major gaps remain. First, very little work has been done to empirically test the predictions of the models, particularly with users who have disabilities. Although two studies suggest a high degree of model accuracy for a small number of AB subjects (Damper, 1984; Gibler & Childress, 1982), neither addresses how accuracy is affected by different system configurations, user characteristics, or model implementations, which are critical issues in gauging the usefulness of any model. Second, most models struggle with how to represent cognitive and perceptual time and simply include these times with the time for physical action. This study addresses these gaps by combining previous AAC modeling work with modeling techniques developed and validated in the field of human-computer interaction. The fundamental techniques employed are reviewed below.

Techniques for Modeling Human-Computer Interaction

The models developed in this work are based on the Keystroke-Level Model, a technique that originated in the field of human-computer interaction (Card, Moran, & Newell, 1983; Olson & Nilsen, 1988). This technique provides a means of identifying the component actions that a user must perform for a particular task. The time required for executing that task is then predicted by summing the times for each component action. In the case of text entry with a word prediction system, the unit task is the entry of a single word, accomplished through a series of letter and word list selections, each of which involves cognitive, perceptual, and motor component actions.

Within the Keystroke-Level Model technique, there are a variety of specific implementation methods. "Keystroke level" means that the model represents events in the range of 100 milliseconds to a few seconds. The way in which the unit task is broken down into these short-duration component actions is the model structure. Each model structure yields a parametric equation for task performance time. A second dimension on which models can vary is the source of user parameter values. *Independent* simulations use parameters drawn from independent sources, such as previous human performance studies, while *data-driven* simulations employ parameters derived directly from subjects' own performance data.

Independent simulations are easier to apply and potentially more generalizable than data-driven simulations, but they are typically less accurate. For tasks involved in text editing and spreadsheet use, errors reported for models with independent structure and parameters average 52.2%, with a range of 24% to 137% (Card et al., 1983; Gong, 1993; John, 1988; Olson & Nilsen, 1988). For models whose structure and/or parameters have been derived from subject data, the average error is 16.6%, with a range of 4% to 33% (Card et al., 1983; Gong, 1993; John, 1988; Olson & Nilsen, 1988).

SPECIFIC AIMS

Two modeling studies were performed to evaluate keystroke-level modeling of performance with a word prediction system. Four specific aims were addressed:

1. Determine the accuracy of *a priori* model predictions, made before data are collected and subject parameters are measured;
2. Assess the model's ability to account for performance under different usage conditions, including

different user strategies, user characteristics, and task configurations;
3. Compare the accuracy of model predictions for users with and without physical disabilities; and
4. Determine the relative accuracy of three different model implementations.

The first study tested a relatively simple model structure, while the second study revised that structure in an attempt to improve model accuracy. The methods and results for each study are described separately below, followed by an overall discussion.

MODELING STUDY #1: METHODS

The first modeling study tested the accuracy of Model 1A, in which structure and parameter values were independent of validation data, and Model 1B, which had the same structure as Model 1A but employed user parameters derived from validation data.

Empirical Data for Model Validation

Methods used to gather the empirical data for model validation are described briefly below. More detailed methods, as well as a report on the empirical results of the study, have been presented elsewhere (Koester & Levine, 1994, 1996).

Fourteen subjects transcribed text both with and without a word prediction feature for seven test sessions. Eight subjects were AB and used mouthstick typing, while six subjects had high-level spinal cord injuries (SCIs) and used their usual method of keyboard access.[1] Each subject was assigned to use one of two word prediction strategies to provide a basis for model structure and reflect a clinical situation in which a user may be given guidelines for when to search the list. 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 without searching the list, then search the list before each subsequent selection. The between-subjects factors of presence/absence of SCI (SCI or AB) and search strategy (1 or 2) were combined to form four subject groups: SCI1, SCI2, AB1, and AB2.

The two interfaces were developed by the investigators specifically for this project to provide sufficient control over the system configuration as well as the means of data collection. The "letters-only" system involved letter-by-letter spelling on a standard computer keyboard, and the "letters + word prediction (WP)" system used single-letter entry augmented by a word prediction feature. A six-word prediction list with a fixed word order was used and presented vertically in the top left corner of the screen. A fixed set of six words was displayed before the first letter of a word was entered, with subsequent predictions based on the letters entered by the user.

Text transcription blocks were presented visually, sentence by sentence. The keystroke savings provided by word prediction was fixed across Sessions 1 to 4, at a level corresponding to an "average" word prediction system (Higginbotham, 1992). Keystroke savings was varied across Sessions 5 to 7, with Session 5 at a higher level and Sessions 6 and 7 providing successively poorer-than-average levels.

The following dependent measures for model validation were recorded in each test session: text generation rates with and without word prediction, the percent improvement in text generation rate with word prediction relative to letters-only typing, and the times to enter each individual word and item (i.e., single letters or word list selections) during word prediction use. Data were collected in real time as subjects transcribed their sentences. Errors, words not entered according to the assigned strategy, and pauses due to the subject referring back to the text card during transcription were all removed from the data before comparison with model predictions (Koester & Levine, 1996).

The rationale for measuring performance at three levels—across an entire session, for each individual word, and for each specific item selected—was to examine model accuracy in successively finer detail. Each level has a specific role to play in understanding the theoretical accuracy and practical relevance of the model. Session-level simulations are of interest because overall session performance is the most clinically relevant measure. However, in summing the times for individual component actions over an entire session, the model may capitalize on the cancellation of positive and negative errors, leading to an overly optimistic estimate of model error. Analysis at the word level was therefore performed to reduce the effect of fortuitous error cancellation, while maintaining some clinical relevance. Analysis at the item selection level provides the strictest test of model accuracy because error cancellation cannot occur at that level of detail. From a clinical standpoint, high accuracy in modeling individual item selections is not necessary but would certainly be encouraging.

Equations for Model 1

Item Selection Times

The first step in modeling item selection time was to determine the user actions most important to use of the Letters + WP system. This was done by analyzing the task of entering words for each of the word prediction strategies. As an example, a flow chart of hypothesized user activity for Strategy 2 is shown in Figure 1. The major activities for both strategies are keypresses (to select a letter or a word) and list

[1] Two subjects with SCI used mouthstick typing and four used hand splint typing.

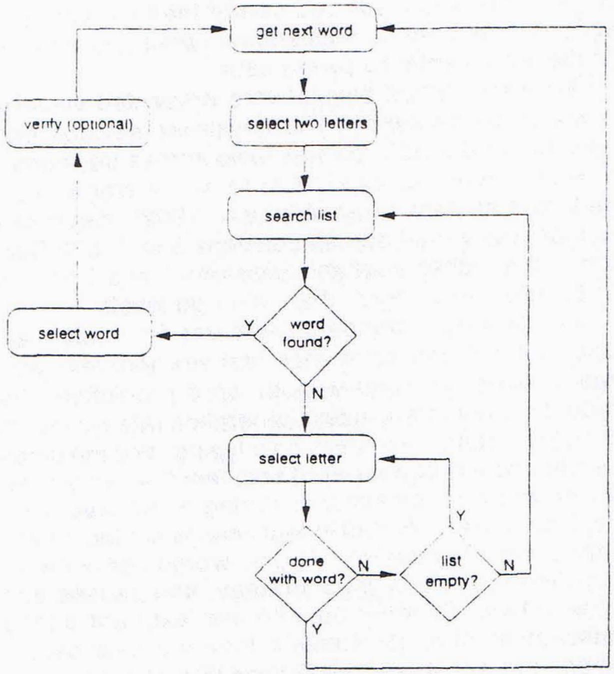

Figure 1. Flow chart representing activities required during use of Strategy 2 with the Letters + WP system.

searches, so these were chosen as the two user model parameters.

Each item selected during use of word prediction was modeled in terms of keypress and list search times, as follows. First, every item required one keypress, and it was assumed that the keypress time was the same for letters and words and was independent of the key's position. Second, some selections included a list search and keypress, depending on the rules of the strategy, while some involved a keypress only. It was assumed that the time required for the list search was independent of the list contents and the sequential position of the search in the series of searches performed for the word. Each selection time, then, was represented by a keypress only (t_k) or a list search plus keypress ($t_s + t_k$).

Word Entry Times

The representations of each item selection provided the building blocks for models of word entry performance with the Letters + WP system. An equation for each word's entry time was formed by summing the model times for each item selected for the word, assuming no overlap between item selection times. The exact equation for·each word was a function of the word prediction strategy employed by the user. For example, to enter the word "that," users of Strategy 1 would find the word in their first list search, before any letters had been entered, and select it. The equation for that word's entry time would then be $t_s + t_k$. Users of Strategy 2, in contrast, would first select the "t" and the "h," only then searching the list and completing the word with a list selection. The resulting equation in that case would be ($t_s + 3t_k$). Equations for each word entered in each test session were formulated in this way.

Overall Session Performance

A variety of specific representations for overall session performance is possible, but all involve a simple weighted average of the number of list searches and the number of keypresses across the session. The average time necessary to generate each character in a session is the sum of two components: the number of searches per character (S) multiplied by the search time (t_s) and the number of keypresses per character (1-ksav) multiplied by the keypress time (t_k).[2] The equation for text generation rate with word prediction, *TGR_{wp}*, can then be written as follows, after converting units of seconds per character to characters per minute:

$$TGR_{wp} = \frac{60}{(S)(t_s) + (1\text{-ksav})(t_k)} \text{ char/min} \qquad (1)$$

The percent improvement in text generation rate relative to Letters-only typing, *TGIMP*, was represented as:

$$TGIMP = \frac{TGR_{wp} - TGR_{lo}}{TGR_{lo}} * 100, \qquad (2)$$

where *TGR_{lo}* is the Letters-only typing rate, modeled as $60/t_k$.

Task Parameters

In Equation 1, the parameters that represent the task are S, the number of searches per character, and ksav, the keystroke savings provided by the Letters + WP system. The values for these are a function of the search strategy used, the specific characteristics of the prediction algorithm, and the transcription text for each session. Task parameters were computed by simulating the entry of each transcription test with each strategy, assuming error-free and strategy-compliant text transcription, and are shown in Table 1.

User Parameters

The two parameters that represent the user are t_k, for keypress time, and t_s, for list search time. Two methods were used to determine the values for these parameters. For Model 1A, user parameters were

[2] Ksav is the keystroke savings provided by the word prediction system, expressed as a proportion rather than a percent.

TABLE 1: **Letters + WP Task Parameters for Strategies 1 and 2**

Session	*Strategy 1*		*Strategy 2*	
	S	*1-ksav*	*S*	*1-ksav*
1	0.512	0.584	0.258	0.703
2	0.508	0.580	0.247	0.686
3	0.519	0.594	0.258	0.703
4	0.518	0.587	0.250	0.701
5	0.456	0.475	0.194	0.585
6	0.541	0.682	0.257	0.765
7	0.625	0.785	0.335	0.857

TABLE 2: **Independent User Parameter Values Used in Model 1A**

Session	*WP Keypress, t_k (sec)*	*List Search, t_s (sec)*
1	1.153	1.000
2	1.117	0.781
3	1.088	0.676
4	1.065	0.610
5	1.045	0.564
6	1.028	0.528
7	1.012	0.500

determined from previous studies reported in the literature. For Model 1B, user parameter values for each subject were derived directly from the subject's performance data. Details regarding each of these methods are provided below.

User Parameters for Model 1A

Human keypress times have been reported in a number of previous human-computer interaction studies (Card et al., 1983; Olson & Olson, 1990). However, studies of AB typists were not considered appropriate sources of mouthstick or hand splint typing speed, and measurements for physically disabled typists could not be found in the literature. Therefore, the performance of AB individuals who typed with mouthsticks for 30 typing tests was used as the basis for independent estimates of keypress time (Levine, Gauger, Bowers, & Khan, 1986). The average typing speed for the 30 tests was plotted and fit to a Power Law curve, which accounted for 90% of the variance in the data.[3] The fitted values were used to approximate what would be expected across the seven test sessions, as shown in Table 2.

Independent values for list search time were derived from a composite of studies on list search (Card, 1982; Landauer & Nachbar, 1985; Neisser, 1963; Somberg, 1987). None of these reports exactly matches the search task presented by this word prediction system or provides specific information on how list search time improves across sessions. Therefore, the following approach was used to estimate a list search time for each session: (1) establish a theoretically appropriate value for search times in the first and seventh test sessions; and (2) draw a Power Law curve between the estimates for Sessions 1 and 7 to determine values for the intermediate sessions.

In defining the list search value for the first session, one relevant result is Neisser's finding that visual scanning requires about 0.2 seconds per word (Neisser, 1963), or 1.2 seconds for a six-word prediction list, assuming each word is examined in series. However, other studies, employing longer lists, suggest that this estimate is probably too long (Card, 1982; Somberg, 1987). Therefore, the expected search time for the first session was set at 1.0 seconds.

There is little direct evidence to guide estimates of search time for the seventh test session, although improvements of several hundred milliseconds between blocks of search trials have been reported (Landauer & Nachbar, 1985; Somberg, 1987). Therefore, an analytical approach was used, which assumed that the time for practiced searches would be a logarithmic function of list length (Landauer & Nachbar, 1985) (i.e., $t_s = [m]\log_2[\text{list\_len} + 1]$, where the slope, m, is the time required to search a single item). The slope was modeled as a simple pattern match, taking an average of 0.17 seconds (Card et al., 1983). With a list length of six words, this yields an estimate of approximately 0.5 seconds for t_s in a practiced search, and this value was used as the estimate for search time in the seventh test session.

The remainder of the list search estimates were determined by defining a Power Law curve based on the endpoints of the first and seventh sessions. The resulting values are shown in Table 2.

User Parameters for Model 1B

In contrast to the independent parameter values used in Model 1A, Model 1B employed values derived directly from the empirical data for each of the 14 subjects. Times for the component actions of list search and keypress while using Letters + WP were determined using the subtractive methods of Card et al. (1983) and Olson

[3] The Power Law of Practice states that task time improves with practice at a rate approximately proportional to a power of the amount of practice (Card et al., 1983). Its mathematical expression s $T_n = T_1n^{-\alpha}$, where T_n is task time on the nth trial, T_1 is task time on the first trial, n is the number of trials, and α is a constant whose value depends on the task.

and Nilsen (1988). Based on the strategy used with Letters + WP, each item selection was labelled according to whether it involved a keypress preceded by a list search or a keypress with no list search. For example, for users of Strategy 2, the first two selections of every word were labelled as keypress only. The keypress time (t_k) during use of Letters + WP was then 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 remaining times. The keypress time during Letters-only typing was calculated by averaging all selection times in the Letters-only test of each session. A set of parameter values was derived for each subject and every test session. Table 3 shows the derived user parameter values for keypress times during use of Letters-only and Letters + WP as well as list search times.

Simulations with Model 1

The dependent variable for overall session performance was the percent improvement in text generation rate with Letters + WP relative to Letters-only typing (TGIMP). In Model 1A, model predictions of TGIMP at each session were generated by substituting the task and independent user parameter values for that session into the model equations (Equations 1 and 2). For Model 1B, each subject's parameter values from Session 3 were used to predict performance in Sessions 4, 5, 6, and 7. This method was employed because user parameter values were derived directly from a session's data, so they could not be used to meaningfully simulate overall performance from that same session.[4] Sessions 4 to 7 were chosen for validation to represent fairly well-practiced performance at four different levels of keystroke savings, leaving Session 3 as the best remaining source for the parameter values.

Methods for performing word- and item-level simulations were identical for both models. For word-level simulations, the time to enter each word during each session of Letters + WP use was predicted by substituting the user parameter values for that session into the word's model equation. Similarly, model-predicted times for each item selected in a session were calculated as either one keypress time (t_k) or a list search-plus-keypress (t_s + t_k).

Measures of Accuracy for Model 1

Model error for session-level simulations was measured as ($TGIMP_{actual}$ - $TGIMP_{model}$). For word- and item-level simulations, model error scores for each subject in a given session were computed as:

$$\text{Error (\%)} = \frac{(T_{actual} - T_{model})}{T_{actual}} * 100,$$

where T_{actual} is the actual time for the word or item and T_{model} is the model-simulated time (John, 1988; Olson & Nilsen, 1988). Additionally, the accuracy of each model in predicting differences between users of different strategies was assessed by comparing the model-predicted difference to the observed difference.

Statistical analyses of the error scores were performed using a repeated measures ANOVA, with between-subjects factors of strategy and presence/absence of SCI and within-subjects factor of test session. To compare the accuracy of Models 1A and 1B, the repeated measures analysis was performed on all error scores, with the additional within-subjects factor of model type. Statistical significance was judged at a family-wise p value of .05, using the Bonferroni procedure to divide by the number of effects examined within

TABLE 3: **Empirically Derived User Parameter Values Used in Model 1B**

	LO Keypress (sec)		*WP Keypress (sec)*		*List Search (sec)*	
Session	*AB*	*SCI*	*AB*	*SCI*	*AB*	*SCI*
1	0.906	0.596	1.046	0.966	0.660	1.170
2	0.884	0.554	0.993	0.910	0.627	1.084
3	0.893	0.556	0.986	0.847	0.593	1.114
4	0.865	0.544	0.904	0.738	0.616	1.178
5	0.884	0.560	0.971	0.831	0.553	1.098
6	0.840	0.562	0.896	0.753	0.508	1.176
7	0.818	0.558	0.889	0.758	0.472	1.138
Mean	0.870	0.561	0.955	0.829	0.577	1.137
95% CI	(0.81, 0.93)	(0.36, 0.76)	(0.91, 1.01)	(0.50, 1.15)	(0.47, 0.69)	(0.87, 1.41)

[4] When parameter values derived from a session's data are used to simulate overall performance in that same session, the model error is precisely zero.

he test (Girden, 1992). For example, a test analyzing hree factors examines seven different effects (three nain effects and four interactions), so the critical p 'alue used for any one of these seven would be 05/7 = .007. Additionally, all p values were those ıdjusted based on the Greenhouse-Geisser epsilon as ı further precaution against Type I errors (i.e., mistakenly judging a difference to be significant when it truly s not) (Girden, 1992).

MODELING STUDY #1: RESULTS

Session Level Simulations

Figure 2 illustrates the Model 1A predictions for each strategy as compared to the average improvement with word prediction relative to letters only (TGIMP) achieved by each subject group.[5] Actual mprovements were lower than those predicted by the nodel, and the difference was particularly large for subjects with SCI.

Table 4 illustrates the average accuracy across sessions for both models. Results were collapsed across strategy because it did not have a significant effect on model accuracy for either model. Both models were fairly accurate at representing the performance of AB subjects. For Model 1A, accuracy was significantly worse for subjects with SCIs, but this was not the case for Model 1B. Session was a statistically significant effect in Model 1A, but not in Model 1B. Accuracy with Model 1B was significantly better than Model 1A across all subjects, especially for subjects with SCIs, for whom accuracy improved by almost 50 percentage points, as compared to about 3 percentage points for AB subjects.

TABLE 4: **Session-Level Errors of Models 1A and 1B**

		Model 1A		*Model 1B*	
Subjects	*N*	*Mean*	*95% CI*	*Mean*	*95% CI*
AB	8	11.14	(6.5, 15.7)	7.94	(4.2, 11.6)
SCI	6	53.07	(49.5, 56.7)	4.04	(2.0, 6.1)
All	14	29.11	(16.4, 41.8)	6.27	(4.0, 8.6)

Error scores were averaged across sessions and the resulting subject error scores were used to compute these statistics.

The ability of Model 1A to predict differences between strategies was mixed. The model predicted that subjects who used Strategy 2 would have text generation rate improvements an average of 7.5 percentage points higher than subjects who used Strategy 1. This proved to be fairly accurate for subjects with SCI, as the average TGIMP for Group SCI2 was 10.9 percentage points higher than that of Group SCI1. However, the relative accuracy of Model 1A was poor for AB subjects, as subjects in Group AB2 had text generation rate improvements that were 4 percentage points lower than users of Group AB1.

Model 1B was more accurate at predicting differences between strategies. The projected differences in text generation rate improvements between Strategy 2

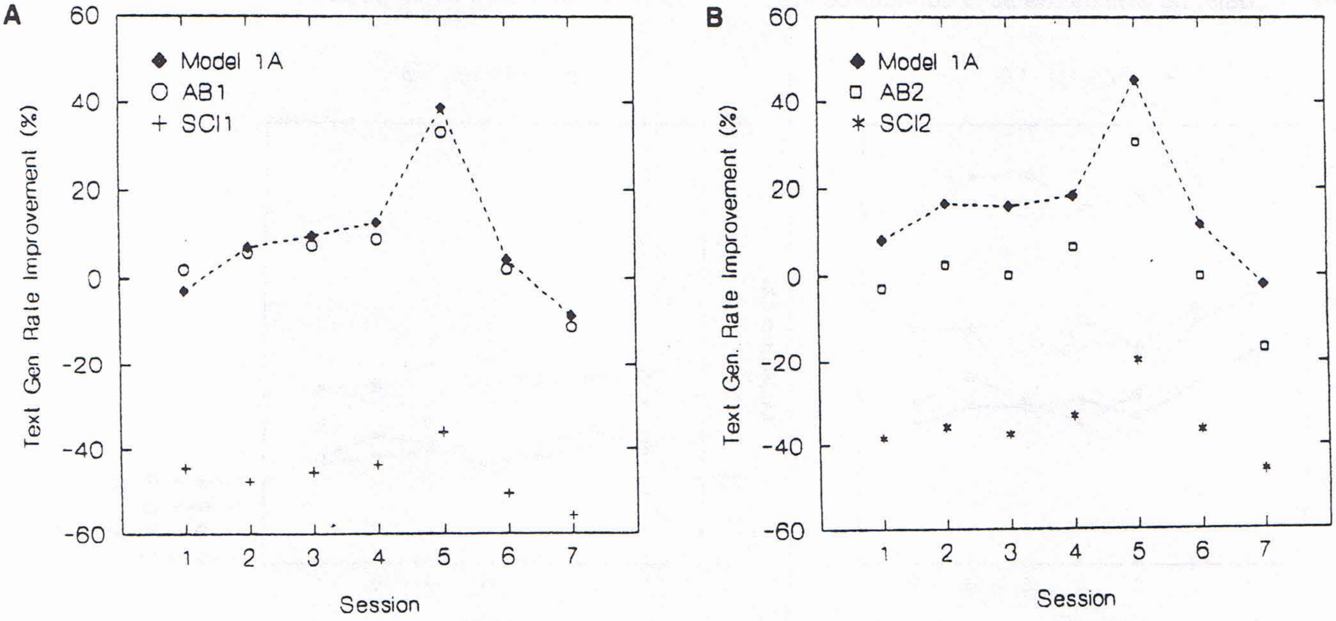

Figure 2. *A*, Comparisons of Model 1A predictions of TGIMP to observed averages of Groups AB1 and SCI1, across sessions. *B*, Comparisons of Model 1A predictions of TGIMP to observed averages of Groups AB2 and SCI2, across sessions.

[5] A similar graphic representation is not possible for Model 1B since model simulations and actual performance were compared individually for each subject.

and Strategy 1 were −2.9 and 11.9 percentage points for AB and SCI subjects, respectively, as compared to actual differences of −3.2 and 12.6 percentage points.

Word Entry Simulations

Figure 3 illustrates the average error for Models 1A and 1B simulations of word entry time. Model 1A error averaged 20.7% for the AB subjects and 35.0% for the SCI subjects with no statistically significant differences between subject groups, either on the basis of SCI or strategy. Session was a significant factor, as the model error was lower for later test sessions.

For Model 1B, model accuracy was not significantly different for the two strategies of Letters + WP use or for any of the seven test sessions. The average Model 1B error was 13.3% for AB subjects and 20.7% for SCI subjects. While that difference approaches statistical significance, at p = .019, it does not meet the Bonferroni criterion p value of .007.

As expected, the Model 1B word entry simulations were more accurate than Model 1A. The average improvement in accuracy was 10.5 percentage points across all subject groups and sessions, with the first two sessions showing the most improvement. There was no difference in the improvements for users of different Letters + WP strategies, but improvements for AB and SCI subjects were different. For AB subjects, the average improvement in accuracy was 7.5 percentage points, which was statistically significant. For subjects with SCI, Model 1B improved accuracy by 14.3 percentage points, but this improvement was not statistically significant, due to high variation in improvements for different subjects with SCI.

Table 5 presents the average error across sessions for both models, as well as the 95% confidence intervals (CIs). Variation between AB subjects was quite small for both models, as illustrated by the narrow CIs. Variation in model error was larger for the SCI than the AB subjects, particularly for Model 1A. Model errors for some of the subjects with SCI were very similar to those of the AB, but this was not true for all.

TABLE 5: **Word Entry Errors of Models 1A and 1B**

		Model 1A		*Model 1B*	
Subjects	*N*	*Mean*	*95% CI*	*Mean*	*95% CI*
AB	8	20.7	(18.52, 22.88)	13.3	(12.38, 14.22)
SCI	6	35.0	(11.23, 58.77)	20.7	(13.59, 27.82)
All	14	26.9	(17.68, 36.12)	16.5	(13.20, 19.80)

Item Selection Simulations

Figure 4 shows the average Model 1A and Model 1B errors for item selection times for each subject group.[6] These were computed for Session 4 only, to measure the accuracy that might be expected after some practice. The errors were larger than those seen for word entry times since there was no longer any opportunity for errors to cancel each other out. For both models, the difference in accuracy between the average error for AB subjects and SCI subjects is striking, suggesting that for at least some subjects with SCI, the models' representation of their item-by-item activity was quite flawed. This large difference was not quite statistically

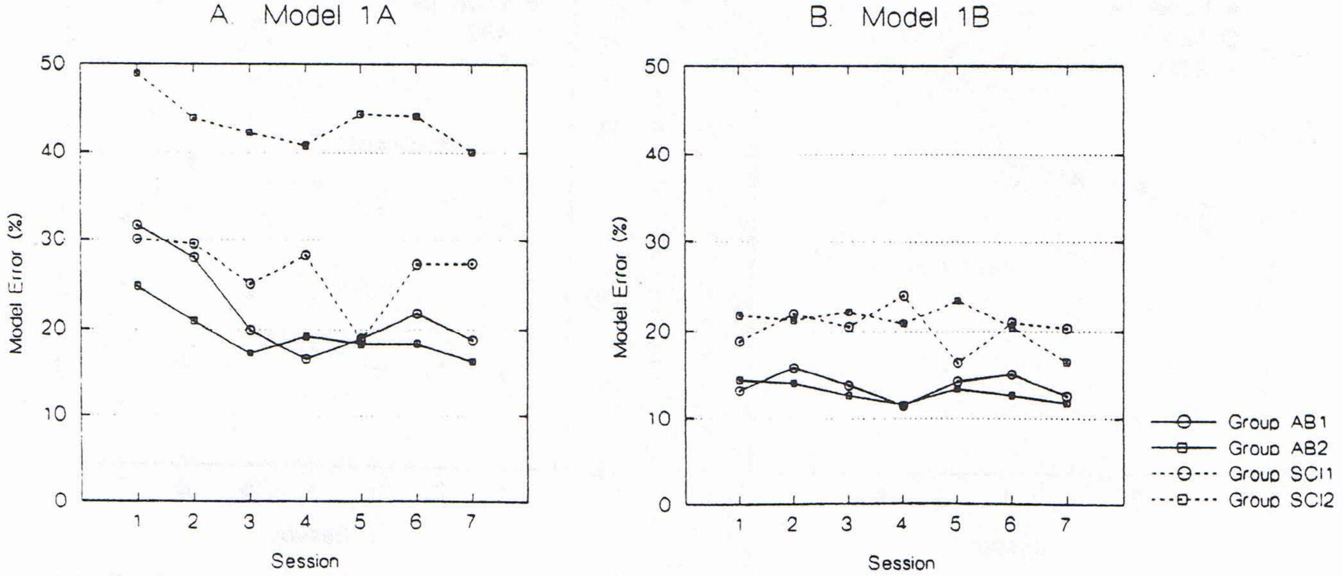

Figure 3. Error of Models 1A and 1B for word entry times, averaged for each subject group.

[6] Numeric data are also shown in Table 8 below in the Results section for the second modeling study.

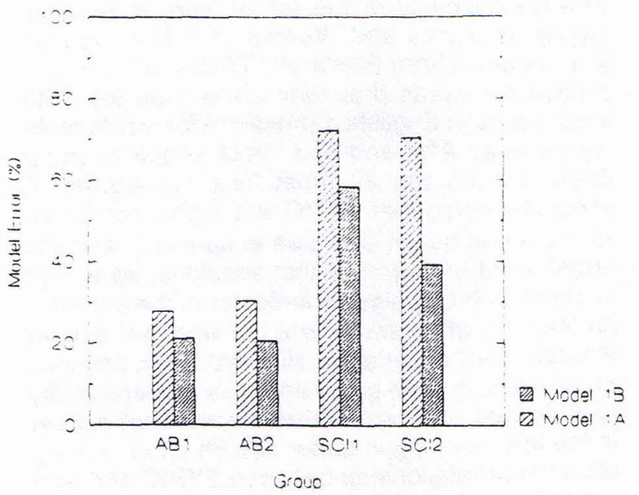

Figure 4. Errors of Models 1A and 1B for item selection time simulations, averaged for each subject group.

significant, however. For both models, there was no significant difference between errors for users of different Letters + WP strategies.

Use of derived parameter values improved model accuracy over Model 1A by an average of 15.4 percentage points for all subject groups, which was statistically significant. The smallest improvement was 6.4 percentage points for Group AB1, with the largest at 31.2 percentage points for Group SCI2. Although improvement was larger for the subjects with SCIs, the interaction between model type and SCI was not significant.

For both models, the variation in errors between individual AB subjects was quite low, while errors for the subjects with SCIs had a higher variance. This large variance is the primary reason why the difference in accuracy between AB and SCI subjects was not statistically significant. For Model 1B, when the SCI subject with the highest error was excluded from the analysis, accuracy for the AB subjects was 11.7 percentage points greater than for subjects with SCIs, and this smaller difference was statistically significant.

MODELING STUDY #2: INTRODUCTION

While the results with the simple structure of Model 1 were encouraging, its simplicity undoubtedly led to some mismatch between the model representation and subjects' actual behavior. This mismatch was revealed in the relatively high model errors observed for selection times of individual items, which averaged almost 50% for subjects with SCIs, even when empirically derived user parameter values were used. One assumption in the Model 1 simulations was that a user's list search time could be represented by a single value, regardless of the conditions of the search. The flaws in this assumption are one possible source of error in the original model.

The hypothesis of this second modeling study is that revising the model structure to refine the representation of list search time may improve model accuracy. Additionally, modeling list search in greater detail is expected to provide further understanding of the list search process during use of word prediction. The second modeling study develops a revised model structure, referred to as Model 2, and compares the accuracy of its simulations to the previously developed Models 1A and 1B.

MODELING STUDY #2: METHODS

Empirical Data for Model Development and Validation

Data from the same experiment used to validate Model 1 were used for the development and validation of Model 2. Due to the amount of analysis involved, only data from Session 4 were used to develop and test Model 2. The list search times for each subject in Session 4 were analyzed to determine the new structure for Model 2, as discussed in more detail below. Model 2 simulations were tested for accuracy against the actual word entry and item selection times of each subject.

General Approach to Structure Revision

The general strategy used to develop Model 2 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. Keypress time was not analyzed for possible structure revision and remained as a single parameter value. The goal was to find predictor variables that worked well for a majority of subjects using a particular word prediction strategy, rather than finding a unique set of predictor variables for each subject.

Identifying Predictor Variables for List Search Time

Based on the list search literature (Card, 1982; Landauer & Nachbar, 1985; Neisser, 1963; Somberg, 1987), as well as the observed performance of these subjects, two hypotheses were generated regarding the factors that may systematically explain variation in list search time. Because it is possible that a user's search processes might be different for successful (i.e., the target word was found and selected) and unsuccessful searches, the implications of each hypothesis for both types of searches were considered.

Hypothesis #1: Serial Search

The serial search hypothesis states that list searches are performed serially, with each word examined one by one. Specifically, the time for successful searches is a linear function of the word's position in the list, while the time for unsuccessful searches is a linear function of the list length. Gibler and Childress (1982) employed the serial model, with the assumption that the linear slope would be 200 milliseconds/word, but they did not empirically verify it. The serial search hypothesis contrasts with Landauer and Nachbar's (1985) findings that logarithmic searches are possible.

Hypothesis #2: Anticipation

The anticipation hypothesis states that search time is affected by the extent to which the user can accurately anticipate the list contents. Search time should be faster at higher levels of accurate anticipation. For successful searches, anticipation of the word's position in the list would provide more direct and therefore faster identification of the word. For unsuccessful searches, anticipation that the word will not be in the list for this search would support a less thorough and therefore faster scan through the list. Anticipation may be influenced by the general typicality of the target word in English or by prior experience with the word's appearance in the list.

These two hypotheses are not mutually exclusive. Serial search and anticipation could both be employed in a given session or even for a given word. The question to be addressed in evaluating these hypotheses is not which one is true, but rather to what extent either of them is true.

Based on these hypotheses, a set of candidate predictor variables was defined. The first variable, SRCH_TYPE, was used to distinguish between successful (SRCH_TYPE = 1) and unsuccessful (SRCH_TYPE = 0) searches, in order to determine if search time depends on whether or not the target word was found in the list. If serial search is dominant, unsuccessful searches should take longer than successful ones, since each word in the list would be examined. If anticipation is dominant, unsuccessful searches might be faster, due to skimming of the list when success is perceived to be unlikely. Successful and unsuccessful search times were also analyzed separately using different predictor variables, as described below.

Successful Searches

The following predictor variables were tested for successful searches:

1. LPOS—the position of the target word in the list, ranging from LPOS = 1 at the top of the list through LPOS = 6 at the bottom of the list. To the extent that the serial search hypothesis is accurate, there should be a positive linear relationship between successful search times and LPOS.
2. TYPIC—typicality of the target word in English, based on Jones and Wepman's (1966) spoken word count. Three levels of TYPIC were defined: 0 (low) for words that rank lower than the 85th most typical in English, 1 (medium) for words ranking between 21st and 85th most frequent, and 2 (high) for the top 20 most frequent words.[7] It should be noted that TYPIC was highly correlated with the number of previous encounters with the target word during prior test sessions, so a high value of TYPIC is also an indicator of the potential for learning when and where the word will appear through direct experience. High typicality, previous experience, or both could enhance a user's ability to anticipate when and where the word will appear in the list, resulting in faster search times. A negative linear relationship between TYPIC and successful search times would be consistent with the anticipation hypothesis.

Unsuccessful Searches

The following predictor variables were tested for unsuccessful searches:

1. LLEN—the length of the word prediction list (ranging from 0–6). Most often, the list contained a full six words, but it was shorter when fewer than six dictionary words matched a subject's current input. If serial search was employed during unsuccessful searches, there should be a positive linear relationship between unsuccessful search times and LLEN.
2. SRCHNUM—the ordinal position of the search in the series of searches across the entire word (ranging from 1–N). For example, when the word "never" was entered using Strategy 1, the "n" was coded as SRCHNUM = 1, as the first search for the word, the "e" as SRCHNUM = 2, and so on until the word was found in the list. When the user can anticipate that the word will probably appear only after several letter selections, less time may be spent on the early searches. Therefore, to be consistent with the anticipation hypothesis, unsuccessful list search times should have a positive linear relationship to SRCHNUM.

Regression Methods for Evaluating Predictor Variables

All analyses were performed separately for users of different word prediction strategies because the specific strengths of the predictor variables were not necessarily expected to be the same due to differences in the strategy search rules. To test the relationship

[7] The choice of the 85th most frequent word as the cut-off for low typicality was made such that all low typicality words appeared at most once across the transcription texts used in this study.

between search times and the success or failure of the search, each subject's list search times in Session 4 were regressed on the SRCH_TYPE variable.

The strength of the remaining candidate predictor variables was evaluated by a series of multivariate and bivariate regression analyses of list search times in Session 4, performed separately for successful and unsuccessful searches. The general procedure was to test four different regression models: the full multiple regression model with both predictor variables and their interaction, the full model without interaction, and the two bivariate models (one for each predictor).

The significance of the regression coefficients and the variance explained by the predictor variables were evaluated to determine the relative strength of each predictor variable. Those predictors that explained at least 10% of the variance and were statistically significant at the 0.05 level for the majority of subjects in each strategy group were selected for Model 2. The results of these analyses are summarized below.

MODELING STUDY #2: LIST SEARCH ANALYSIS RESULTS

Observed List Search Times

The average observed list search times in Session 4 were 0.62 seconds for AB subjects, with a 95% CI of (0.51, 0.72), and 1.18 seconds for subjects with SCI, with a 95% CI of (0.90, 1.40). These and other empirical results from the study have been discussed previously (Koester & Levine, 1994, 1996) but are provided here to give background to the regression results reported below.

Regression Results for Successful vs. Unsuccessful Searches

Strategy 1 Subjects

Unsuccessful searches were an average of 240 milliseconds (msec) faster than successful searches, suggesting that Strategy 1 subjects skimmed the list when success was not likely. The regression coefficient for SRCH_TYPE was significant for five of seven subjects, but it explained more than 10% of the variance in search times for only two subjects. Therefore, SRCH_TYPE was not considered to be a strong predictor of list search times.

Strategy 2 Subjects

On average, unsuccessful searches were 50 milliseconds faster than successful ones; the difference was not statistically significant for any of the seven subjects. SRCH_TYPE explained only 1.6% of the variance in search times for this group, and the variance explained did not exceed 10% for any subject. Therefore, SRCH_TYPE was not considered to be a strong predictor of list search times.

Regression Results for Successful Searches

Strategy 1 Subjects

The combination of LPOS (word's position in list) and TYPIC (word's typicality) explained an average of 12.6% of the variance in successful search times. The coefficient for LPOS was significant for five of the seven subjects, while the TYPIC coefficient was significant for only one subject. The bivariate analyses also showed TYPIC to be a weak predictor, as TYPIC by itself explained an average of only 3.2% of the variance. The bivariate relationship between successful search times and LPOS alone explained an average of 10.1% of the variance and was significant for five of seven subjects.

Strategy 2 Subjects

Regressing successful search times on the linear combination of LPOS (word's position in list) and TYPIC (word's typicality) for Strategy 2 subjects explained 37.8% of the variance in list search times. Additionally, the coefficients for LPOS and TYPIC were each significant for a majority of subjects, suggesting that both predictor variables had a strong contribution to the fit of the multiple regression model. The interaction between LPOS and TYPIC did not improve regression fit.

Regression Results for Unsuccessful Searches

Strategy 1 Subjects

Each subject's list search times were fit to a full regression model in LLEN (list length) and SRCHNUM (search number). The model explained an average of 23.8% of the variance in unsuccessful list search times, and the LLEN * SRCHNUM interaction was significant for six of seven subjects.

To determine the source of the interaction, the bivariate relationship between list length and unsuccessful search time was examined. For all Strategy 1 subjects, there was a positive linear trend across partially full lists (LLEN < 6) and a downturn when the list was full (LLEN = 6). This suggests that a linear model in LLEN was most appropriate only for partially full lists. When the bivariate influence of LLEN was examined for partially full lists only, a significant linear relationship was observed for six of seven subjects, and LLEN explained an average of 45.1% of the variance in times for unsuccessful searches of partially full lists. Furthermore, when SRCHNUM was added to this model, the amount of variance explained improved only slightly, to an average of 47%.

The weak influence of SRCHNUM on search times for partially full lists, combined with the relatively fast search times seen for full lists, suggests that the interaction between LLEN and SRCHNUM was due largely to SRCHNUM's stronger effect on full lists. This was supported by the bivariate analysis of SRCHNUM and the unsuccessful search times for full lists, which revealed a significant positive linear relationship for six of seven subjects. SRCHNUM explained an average of 18.7% of the variance in these search times, and, on average, each successive search took 240 milliseconds longer than the previous one, providing strong evidence that Strategy 1 subjects did make guesses as to when the word would appear in the list.

Strategy 2 Subjects

For Strategy 2 subjects, regressing unsuccessful list search times on LLEN (list length) and SRCHNUM (search number) explained an average of 35.0% of the variance in unsuccessful search times. The interaction between LLEN and SRCHNUM was not significant and did not improve regression fit. The LLEN coefficient was significant for four of seven subjects, while the SRCHNUM coefficient was significant for only two subjects. The bivariate relationship between unsuccessful search times and LLEN explained an average of 30.7% of the variance in list search times. The decrease in search time for full lists that was observed for Strategy 1 subjects was not observed for Strategy 2.

Summary of Strong Predictor Variables

Table 6 summarizes the predictor variables that had the strongest influence on list search times for users of each word prediction strategy. These variables were chosen for use in the revised structure of Model 2.

MODELING STUDY #2: MODEL SIMULATION RESULTS

Calculation of Revised List Search Values

Based on the strong predictors identified above, new values for list search times were calculated for each subject to better reflect how search time varies with the conditions of the search. For example, to represent successful searches for a Strategy 1 subject, search times at each level of LPOS were calculated from the bivariate regression equation for that subject. Similarly, new search times for other contexts were calculated using each subject's regression equation for the strong predictor variables in Table 6. The keypress times used were the same as those derived in Model 1B.

Model Simulations with Revised Model Structure

Model simulations for word entry and item selection times were performed for each subject and compared to the actual times observed in Session 4, following the same methods used in Models 1A and 1B, described above. Session model errors were not computed since only one session was analyzed. The percent error of Model 2 was compared to Models 1A and 1B using a repeated measures ANOVA technique, with the between-subject factors of strategy and SCI and within-subject factor of model type. Statistical comparisons of Model 2 error for the between-subject factors of strategy and SCI were made using standard ANOVA tests. For all statistical tests, significance was judged at a family-wise p value of .05.

Word Entry Simulations

Table 7 shows the average error of Model 2 simulations of word entry times for each subject group, along with the corresponding errors for Models 1A and 1B in Session 4. Model 2 results are discussed relative to Model 1B primarily, since that was the more accurate of the Model 1 implementations. For word entry simulations, the average error for Model 2 across all subjects was 14.3%, which was a small but statistically significant improvement over Model 1B.

Like Models 1A and 1B, the average error of Model 2 was not significantly different for Strategy 2 subjects as compared to Strategy 1. Also consistent with Model 1 results, Model 2 error was greater for subjects with SCIs, averaging 19.4%, as compared to AB subjects at 10.5%. This difference was not statistically signifi-

TABLE 6: **Summary of Strongest Predictor Variables for Each Strategy**

Search Type	*Strategy*	*Coefficient*	*Mean (msec)*	*SD (msec)*	*N Significant*	*Average $\bar{R}^2$ (%)*
Successful	1	LPOS	88	64	5/7	10.1
	2	LPOS	172	102	5/7	37.8
		TYPIC	−193	82	4/7	
Unsuccessful	1	SRCHNUM*	237	164	6/7	18.7
		LLEN†	192	68	6/7	45.1
	2	LLEN	144	97	6/7	30.7

Coefficient means were averaged across subjects and are in msec units.
*Regressed on searches of full lists; † regressed on searches of partially full lists.

TABLE 7: **Word-Level Simulation Errors for Models 1A, 1B, and 2 in Session 4**

		Model 1A		*Model 1B*		*Model 2*	
Subjects	*N*	*Mean*	*95% CI*	*Mean*	*95% CI*	*Mean*	*95% CI*
AB	8	17.9	(14.9, 20.9)	11.5	(10.6, 12.4)	10.5	(9.8, 11.1)
SCI	6	34.5	(14.7, 54.3)	22.5	(9.0, 35.9)	19.4	(6.4, 32.3)
All	14	25.0	(15.5, 34.6)	16.2	(10.6, 21.8)	14.3	(9.1, 19.4)

cant, however, largely because high error for one subject (BG) led to high variance within the subjects with SCIs.[8] The average improvement in accuracy with Model 2 as compared to Model 1B was significantly larger for subjects with SCIs, at 3.1 percentage points, as compared to AB subjects at 1.0 percentage point.

The variation in Model 2 error between individual subjects was similar to that seen in Model 1B simulations of word entry times. Ten of the 14 subjects had very similar errors, ranging from 9.5% to 12.2%. The subjects in Group SCI2 had slightly higher errors, from 14.5% to 19.6%, and subject BG (of Group SCI1) continued to have the largest error at 43.8%. The persistently high error for subject BG even after this fairly extensive structural revision suggests a large amount of unsystematic variation in his performance.

Item Selection Simulations

Table 8 shows the average error of Model 2 simulations of item selection times for each subject group, along with the corresponding errors for Models 1A and 1B in Session 4. Across all subjects, Model 2 yielded a statistically significant improvement in accuracy over Model 1B. Model 2 error for subjects with SCIs averaged 42.6% and 17.6% for AB subjects. This large difference was not statistically significant, however. Word prediction strategy had no significant effect on Model 2's accuracy or improvement in accuracy relative to Model 1B.

Example: Comparison of Fit for Models 1A, 1B, and 2

One way to concretely illustrate the fit of the three model implementations tested is to plot actual and predicted selection times for representative words. Figure 5 illustrates the entry of the word "choice" by subjects SJ, of Group AB1, and SD, of Group AB2, comparing their actual selection times to the times simulated by Models 1A, 1B, and 2. This word was not in the prediction dictionary so each of its six letters needed to be entered.

Since Models 1A and 1B share the same structure, the pattern of their predictions is the same. Each item predicted to include a list search (e.g., the first four letters for subject SJ) is simulated at one fixed time (t_s + t_k), and each item predicted to include only a keypress (e.g., the first two letters for subject SD) is simulated at another fixed time (t_k). While this predicted pattern does not precisely fit the observed pattern, particularly for subject SJ, Model 1B provides a closer fit since its data-driven parameter values better reflect each subject's characteristics.

In revising the model structure, the fit for subject SJ improved noticeably with Model 2 as compared to Model 1B, although the "peak" in model-predicted times is not as large as the actual peak. Use of the SRCHNUM predictor succeeded in modeling the gradual rise in search time across the letters "c," "h," and "o," while the decrease in search time across "o," "i," and "c" was accounted for by considering the shrinking list length (LLEN). For subject SD, it is hard to improve on the close fit achieved by Model 1B. However, the revised structure of Model 2 did succeed in modeling the slight decrease in search time

TABLE 8: **Item-Level Simulation Errors for Models 1A, 1B, and 2 in Session 4**

		Model 1A		*Model 1B*		*Model 2*	
Subjects	*N*	*Mean*	*95% CI*	*Mean*	*95% CI*	*Mean*	*95% CI*
AB	8	28.9	(26.0, 31.9)	20.8	(19.3, 22.3)	17.6	(16.5, 18.8)
SCI	6	71.2	(12.5, 130.0)	48.7	(6.3, 91.1)	42.6	(5.1, 80.1)
All	14	47.1	(23.6, 70.8)	32.7	(16.1, 49.4)	28.3	(13.5, 43.1)

[8] When this subject was removed from the analysis, the AB-SCI difference shrunk to 3.5 percentage points, but it was statistically significant.

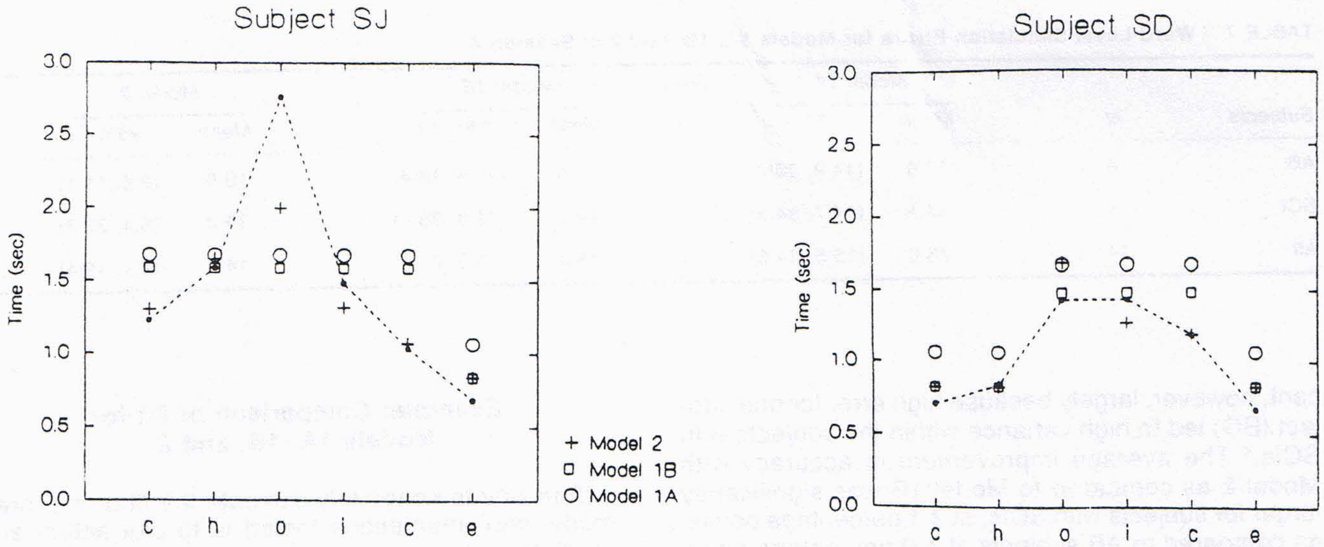

Figure 5. Moment-by-moment graphs for SJ (Group AB1) and SD (Group AB2), entering the word "choice." Actual item selection times are connected by dotted line. Times predicted by Models 1B and 2 are shown for comparison.

across "i" and "c," due to the shrinking list length, which Model 1B did not account for.

Quantitatively, the Model 2 error for SJ's entry of "choice" was 6.9%, which was actually worse than the Model 1B error of 0.11%. This was due to advantageous cancelling of errors under Model 1B. In contrast, the item selection time error was much improved with the revised structure of Model 2, at 12.5%, compared to the 25.8% error under Model 1B. For SD, the error for "choice" overall improved from 10.4% under Model 1B to 5.8% with Model 2, while the item selection time error was unchanged at 11.6% for this word.

DISCUSSION

The following discussion summarizes the major findings of this modeling research and discusses the progress made toward the specific aims outlined at the beginning of the paper. Limitations and future directions for this work are also presented.

Aim 1: Accuracy of *a priori* Model Predictions

Aim 1 concerns the accuracy of *a priori* performance predictions, made with independent subject parameter values. This was measured by comparing Model 1A predictions to subjects' actual performance. *A priori* accuracy for the AB subjects, with an error of 11% for overall session performance and approximately 20% for word entry times, was quite good relative to the 52% error reported in previous studies (Card et al., 1983; Gong, 1993; John, 1988; Olson & Nilsen, 1988). The success of the independent parameter values for these subjects provides some confidence that parameters measured for similar individuals in different studies can be used to represent performance in new situations. Model 1A's accuracy was consistently worse for subjects with SCIs, with an error of 35% for word entry times, which argues strongly for the use of caution when transferring parameter values between studies. While Model 1A was quite accurate for some of the SCI subjects, it was extremely inaccurate for others. In contrast, model error was quite consistent across the AB subjects. The success of Model 1A in predicting differences between strategies was mixed, since the predicted advantage for Strategy 2 occurred only for the SCI subjects.

Aims 2 and 3: Model Accuracy across Different Conditions

To be maximally useful, performance models should be equally accurate for different conditions of interest, such as strategy of use, task characteristics, and user characteristics. With respect to strategy of use, this was the case for Models 1A, 1B, and 2, as strategy of word prediction use had no effect on accuracy for any of the models. Model accuracy for Models 1A and 1B was consistent across the changes in task characteristics that occurred in the last four sessions.[9] Model 1A had mixed success at representing within-subject changes that occurred across sessions. Accuracy in the first two sessions was worse than in subsequent sessions because the user parameter estimates for the early sessions were too slow compared to subjects' actual values. Fortunately, a fast improvement in parameter values was also predicted, which accounted for the improved accuracy in subsequent sessions. For Model 1B, accuracy was consistent across all sessions, which

[9] The sensitivity of Model 2 to the session-dependent factors of task characteristics and user skill was not assessed for Model 2, as its accuracy was measured only in Session 4.

demonstrates that changes in user characteristics were successfully accommodated by changes in the user parameter values.

All three model implementations were not quite as successful at representing differences *between* subjects, since accuracy was almost always worse for the SCI as compared to AB subjects. The magnitude of the difference was largely a function of poor accuracy for one SCI subject, BG, but even when this individual was removed from the analyses, model accuracy remained worse for SCI performance. However, the fact that the models were less accurate for subjects with SCIs does not mean that they are not applicable for those subjects. In absolute terms, accuracy was quite good for subjects with SCI when empirically derived parameter values were used in the models. For example, even when subject BG was included, model error for SCI word entry times averaged 21% and 19% for Models 1B and 2, respectively. In addition, accuracy of Model 1B in simulating session-level performance measures was equally accurate across all subjects, with errors averaging well under 10%.

Aim 4: Relative Accuracy of Different Model Implementations

The relative accuracy of different model implementations was addressed by comparing the accuracy for Models 1A, 1B, and 2. The accuracy of Model 1A vs. 1B is discussed first to illustrate the difference between using independent and data-driven user parameter values in the same model structure. Then Model 2 is discussed relative to Model 1B, to highlight the effect of revising the model structure while using data-driven parameter values.

Deriving user parameter values for each subject, as in Model 1B, led to significantly better accuracy than the independent parameter values used in Model 1A. The sources of this difference can be traced to several implicit assumptions involved in the use of independent parameter values. One assumption of Model 1A is that AB and SCI subjects would have the same values for keypress time. This is partially supported by the lack of a significant difference between the measured keypress times during word prediction use for these groups. However, keypress times during Letters-only typing were faster for SCI than for AB subjects. An extension of this assumption is that every subject has the same keypress time. This is a surprisingly accurate assumption in the case of the AB subjects, as the 95% CI were quite narrow for both Letters-only and Letters + WP keypress times. For subjects with SCIs, however, variance between subjects was higher, as illustrated by the wider confidence intervals (see Table 3).

The independent estimates also assumed that keypress time would be the same for the Letters-only and Letters + WP systems, since the keypress parameter is intended to reflect the motor component of item selection. The empirical data showed this assumption to be false, particularly for subjects with SCIs. Keypress times during use of Letters + WP were significantly slower than during Letters-only typing, by an average of 23% across all subjects (Koester & Levine, 1994, 1996). This difference has been attributed to additional cognitive load involved in the use of word prediction (Koester & Levine, 1994, 1996).

For list search times, the independent estimates were much more accurate for AB than for SCI subjects. Actual list search times for AB subjects in the first two sessions were faster than the independent estimates, but for Sessions 3 to 7, the values were remarkably similar. The 27% improvement of AB subjects' list search times across sessions was less than the 50% expected based on the independent estimates. The list search times of the subjects with SCI refute the assumption that list search times would be the same for AB and SCI subjects. Subjects with SCIs had much slower search times than the independent estimates, and their 2.7% improvement was much smaller than the rapid improvement suggested by the independent estimates.

The model accuracy obtained with Model 1B should be considered the best accuracy possible for this simple model structure since the parameter values were derived directly from the performance data. While Model 1B simulations used parameters specific to each subject, it is also possible to average parameters across subjects before performing simulations, which would generally reduce model accuracy. The use of user-specific parameter values is important for the subjects with SCI in this study, given the individual differences seen in their parameter values.

Concerning the future usefulness of Model 1B, which depends on empirical measurements of user parameter values, it is encouraging to note that the parameter values derived for these subjects did not depend on the strategy used with word prediction. This helps provide confidence that a set of parameter values measured for an individual during use of one strategy could be used to simulate expected performance with an alternative strategy. However, there is still a great deal left to learn about user parameter values during use of word prediction. In particular, the source of differences between AB and SCI subjects, as well as expected changes in search time with practice, must be better understood.

In Model 2, the model structure was revised to provide a more specific account of the list search action, yielding small improvements in accuracy of 1.9 and 4.4 percentage points over Model 1B for word entry and item selection times, respectively. While these were statistically significant improvements, the corresponding improvements in Model 1B accuracy relative to Model 1A were much larger, at 8.8 and 14.3 percentage points. In both cases, improvements were larger at the item selection level and were diluted at the word level due to cancellation of positive and negative errors.

Part of the reason why Model 2 did not result in greater improvement was that accuracy for Model 1B was already quite good, with less than 14% error at the word level for 10 of the 14 subjects. While this is a comforting consideration, it does not account for the other four subjects whose Model 1B errors were around 20% or more. Accuracy for those subjects improved only slightly more under Model 2 than did those whose accuracy was already high with Model 1B.

In developing Model 2, there were several limitations in the ability of the tested predictors to explain variance in list search time. This study was not specifically designed as an experiment to isolate potential factors in list search, so confounding between some factors was an issue and cases were not evenly distributed across all levels of each factor. For example, only a relatively small percentage of searches were performed on partially full lists. While Model 2's sensitivity to list length better accounted for those cases, the overall effect on accuracy was low due to their small influence on the total number of cases. However, the search conditions of this study were realistic representations of actual word prediction use, so conditions in which there were few cases reflect what would be expected with actual use.

A second possibility for the relatively small improvement in accuracy with Model 2 is that the regression analyses did not examine all possible factors that may have influence over search time. For example, the nature of the distractor words in the list may have some systematic effect. It may be more difficult (and therefore slower) to find a target from among a list of words that all have the same length, perhaps, or that share the first one or two letters. Additionally, the presence of morphological variants of the target word may distract and even confuse the search process, whether the target word is present in the list or not.

It is likely, however, that even the most comprehensive set of predictors would be limited, as there may simply be a certain amount of variance in list search time that cannot be easily explained. The range of search conditions during use of word prediction may have led to more complex search performance than could be captured by the tested predictors. While the predictors did reveal some regularities in search performance, their influence on search time may be more specific and localized than could be modeled within the broad partitions of Model 2 (e.g., all unsuccessful searches), and there may be other factors that were not even examined. In other words, the unexplained variance in list search time may not simply be entirely random; it could be meaningfully connected to subjects' goals and strategies in ways that are not easily discovered.

Finally, Model 2 was revised only relative to list search time; keypress time remained a single parameter as it was in Models 1A and 1B. An attempt to explain variance in keypress time, perhaps through a Fitts' Law model, might lead to greater improvement in model accuracy (Olson & Olson, 1990).

Limitations

Although the model validation methods were designed to assess model accuracy under a range of conditions, it was not possible to examine every condition of interest in a single study. In particular, these validation results are most applicable to individuals who have developed skill but not long-term expertise with word prediction, have relatively low variation in the motor abilities relevant to use of their system, do not have significant cognitive impairments, and follow particular strategies in using a system. Future work validating the modeling techniques with other populations and with different systems is necessary in order to confidently apply models to a truly broad range of user-system combinations.

The model structures tested in this work represented user performance as a linear combination of component actions at the keystroke level with the assumption that component actions are performed serially rather than in parallel. This is a fairly simple structure, yet its validity is supported by the accuracy obtained with Model 1B (as well as Model 1A for AB subjects). However, other types of modeling techniques could have been used. A more complex modeling technique, such as a GOMS model or a production system model, would represent performance at a finer level of detail, down to individual cognitive cycles taking 70 to 100 milliseconds (Card et al., 1983). A detailed model of this type may provide more insight into the source of cognitive overhead with word prediction. For example, it may suggest the specific cognitive processes that account for the slower keypress time observed with word prediction relative to Letters-only typing. This is a potential advantage of finer-grained models, and the extent of this advantage could be evaluated in future work. However, a more complex structure has distinct disadvantages in that it is more cumbersome to work with and more difficult to apply. The same is true for modeling techniques that can cope with parallel execution of processes, such as critical path analysis (John, 1988). Examining the word prediction task using critical paths may lead to new insights into user behavior as well as opportunities for enhancing performance, but its complexity makes it unsuitable as a starting point for model development.

While the Keystroke-Level Model used here was reasonably accurate in simulating users' performance time with word prediction, it is not designed to address other important aspects of user-system interaction. In particular, the technique has difficulty accounting for problem-solving aspects of user behavior, the effect of errors on performance, user fatigue, and user acceptance of a system (Olson & Olson, 1990). This means that the Keystroke-Level Model should not be expected to address every question related to AAC and computer access systems, but this limitation does not diminish its usefulness as a means of simulating performance time.

One limitation that is not often explicitly mentioned is that Keystroke-Level Models are tied to assumptions about users' methods or strategies. While this has the advantage of supporting comparative analyses of different strategies, it limits the ability to make general conclusions about word prediction independent of the strategy used or in instances where a user does not follow a particular strategy. John (1988) has addressed this limitation by averaging predicted times for all plausible strategies, which has practical value but detracts from the behavioral accuracy of the model.

Insights into List Search Performance

While the revision in model structure did not improve model accuracy a great deal, the revision process did provide some insight into the factors that influence searching the word list. The results suggest that list search does have serial elements, as Gibler and Childress (1982) hypothesized. Evidence of serial list search was observed in successful as well as unsuccessful searches, through the significance of the word's position in the list (LPOS) and the list length (LLEN), respectively. One or both of these predictor variables had a statistically significant influence on search time for all 14 subjects. The average coefficient for the LPOS and LLEN variables was 150 milliseconds, which can be interpreted as the time required to process one word in the list. This is similar to, but a little faster than, the rate of 200 milliseconds/word used by Gibler and Childress (1982) and observed by Neisser (1963).

These results also show that list search is not purely serial. Serial search appeared to be complemented by a knowledge-driven search, in which anticipation of success or failure affects how one searches the list, consistent with Somberg's (1987) findings. Evidence of anticipation was observed for successful searches in Strategy 2; these subjects found highly typical words in the list an average of 380 milliseconds faster than atypical words. Additional support for anticipatory search was found for unsuccessful searches of full lists for Strategy 1 subjects; results strongly suggest that subjects did not thoroughly search the list word by word when they had a high expectation that it would not contain the target word. While anticipation was an important factor, it cannot be determined from the conditions of this study whether subjects' knowledge of list contents was learned through specific experience with the word prediction system or inferred through knowledge of English.

Practical Application of Results

The primary goal of this paper was to demonstrate model accuracy in preparation for applying the models in future work (Koester & Levine, in press). However, two points relevant to the clinical application of modeling can be briefly discussed here.

One way to apply the models developed in this research is to generate specific predictions of performance time using the model equations. The validation results suggest that Model 1 should be preferred for this application over Model 2, since its two-parameter structure is much simpler and almost as accurate as Model 2. Making performance predictions requires the specification of the user parameter values to be used in the equations, and as the validation results showed, values measured directly from a specific individual give notably better results than independent parameter values. The use of empirically derived user parameters in the model leads to questions regarding the practical application of the model. First, how feasible is parameter measurement in real-world settings? Fortunately, the individualized approach used in AAC and computer access makes it reasonable to think that a clinician could measure an individual's parameters as part of a clinical evaluation. Second, since the technique used to measure the parameters involves measurement of user performance, what is the purpose of subsequent model simulations of that same performance? This concern can be addressed by considering that the parameters measured are more fundamental building blocks of user performance, so they can be used to simulate performance in conditions beyond the particular ones from which they were derived. The method and application of model simulations to address clinical questions about performance will be the subject of a subsequent paper (Koester & Levine, 1997).

In addition to specific answers regarding expected performance with word prediction, a second and often under-appreciated use of modeling is to provide help in framing the problem. Modeling provides "tools for thought" (Newell & Card, 1985), alerting clinicians and designers to the important factors in determining performance with word prediction. For example, Equation 1 shows that there are four primary factors that determine performance with word prediction: the average number of searches required per character, the keystroke savings, the user's keypress time, and the user's list search time. The list search analyses performed for Model 2 were able to concretely demonstrate how factors such as the length of the prediction list and the typicality of the target word can influence the user's search time.

The modeling framework is also a valuable research tool that complements empirical methods, even when model validation is not the primary research goal. The structure provided by the models and the strategy assignment supported the measurement of individual keypress and list search time during use of word prediction, neither of which has been reported previously.

CONCLUSIONS

These results support the hypothesis that user performance with word prediction systems can be successfully modeled using a relatively simple model that considers only keypress and list search actions (Model 1). The accuracy of this very parsimonious model is encouraging, as Model 1B yielded an average error of 16% for word entry times and less than 10% for overall text generation rate. It is probably unrealistic to expect model simulations to do much better than that.

The revisions made to the model structure with Model 2 illustrate that list search during use of word prediction is a complex process influenced by the particular conditions of the search. Both serial search and anticipation of the list contents significantly affected subjects' list search times. While modeling list search in more detail raised some interesting hypotheses about the list search process, it did not greatly improve the accuracy of modeling users' performance. Model 2 is best suited for addressing specific questions regarding list search time, while Model 1 is more practical for generating quick simulations of overall performance. A thorough exploration of how these models can be used is presented in a subsequent paper (Koester & Levine, in press).

ACKNOWLEDGMENTS

Many thanks to the subjects who participated in this study for their generous contributions of time and energy. Support was provided by the National Science Foundation, The University of Michigan Rackham School of Graduate Studies, and the U-M Rehabilitation Engineering Program. This paper is based in part on a doctoral dissertation entitled "User Performance with Augmentative Communication Systems: Measurements and Models," written by the first author in partial fulfillment of the requirements for a Ph.D. degree in bioengineering at the University of Michigan. Simon P. Levine was the chair of the dissertation committee.

Address reprint requests to: Heidi Horstmann Koester, Koester Performance Research, 368 Oak Harbor Court, Holland, MI 49424, USA; E-mail: hhk@umich.edu.

REFERENCES

Baker, B. R., Barry, R. F. (1990). A mathematical model of Minspeak. *Second annual European Minspeak Conference.*

Card, S., Moran, T., Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Card, S. K. (1982). User perceptual mechanisms in the search of computer command menus. In *Human factors in computer systems proceedings* (pp. 190–196). Gaithersburg, MD: National Bureau of Standards.

Dabbagh, H. H., Damper, R. I. (1985). Average selection length and time as predictors of communication rate. In *Proceeding of the 8th Annual RESNA Conference* (pp. 404–406). Washington, DC: RESNA.

Damper, R. I. (1984). Text composition by the physically disabled: A rate prediction model for scanning input. *Applied Ergonomics, 15,* 289–296.

Gibler, C. D., Childress, D. S. (1982). Language anticipation with a computer-based scanning communication aid. In *Proceedings of the IEEE Computer Society Workshop on Computing to Aid the Handicapped* (pp. 11–15). Charlottesville, VA: IEEE.

Girden, E. R. (1992). *ANOVA: Repeated measures.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-084. Newbury Park, CA: Sage.

Gong, R. (1993). *Validating and refining the GOMS model methodology for software user interface design and evaluation.* Doctoral dissertation, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan.

Goodenough-Trepagnier, C, Rosen, M. J., Jandura L., Getschow, C. O., Genoese-Zerbi, F., & Felts, T. (1987). Preliminary validation of prescription guide for selection of communication aids. In *Proceedings of the 10th Annual RESNA Conference* (pp. 100–102). Washington, DC: RESNA.

Goodenough-Trepagnier, C., Rosen, M. J. (1988). Predictive assessment for communication aid prescription: Motor-determined maximum communication rate. In L. Bernstein (Ed.), *The vocally impaired* (pp. 167–185). Philadelphia: Grune and Stratton.

Higginbotham, D. J. (1992). Evaluation of keystroke savings across five assistive communication technologies. *Augmentative and Alternative Communication, 8,* 258–272.

Horstmann, H. M., Levine, S. P. (1990). Modeling of user performance with computer access and augmentative communication systems for handicapped people. *Augmentative and Alternative Communication, 6,* 231–241.

Horstmann, H. M., Levine, S. P. (1992). Modeling AAC user performance: Response to Newell, Arnott, and Waller. *Augmentative and Alternative Communication, 8,* 92–97.

John, B. E. (1988). *Contributions to engineering models of human-computer interaction.* Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Jones, L., Wepman, E. (1966). *A spoken word count.* Chicago: Language Research Associates.

Koester, H. H., Levine, S. P. (1994). Modeling the speed of text entry with a word prediction interface. *IEEE Transactions on Rehabilitation Engineering, 2,* 177–187.

Koester, H. H., Levine, S. P. (1996). The effect of a word prediction feature on user performance. *Augmentative and Alternative Communication, 12,* 155–168.

Koester, H. H., Levine, S. P. (in press). Model simulations of user performance with word prediction. *Augmentative and Alternative Communication, 13,* 000–000.

Landauer, T. K., Nachbar, D. W. (1985). Selection from alphabetic and numeric menu trees using a touch screen. In *CHI '85 Proceedings* (pp. 73–78). New York: Association for Computing Machinery.

Levine, S. P., Gauger, J. R. D., Bowers, L. D., Khan, K. J. (1986). A comparison of mouthstick and Morse code text inputs. *Augmentative and Alternative Communication, 2,* 51–55.

Neisser, U. (1963). Decision time without reaction time: Experiments in visual scanning. *American Journal of Psychology, 76,* 376–385.

Newell, A. F., Arnott, J., Waller, A. (1992). On the validity of user-modeling in AAC: Comments on Horstmann and Levine (1990). *Augmentative and Alternative Communication, 8,* 89–92.

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

Olson, J.R., Nilsen, E. (1988). Analysis of the cognition involved in spreadsheet software interaction. *Human Computer Interaction, 3*, 309–350.

Olson, J. R., Olson, G. M. (1990). The growth of cognitive modeling in human-computer interaction since GOMS. *Human Computer Interaction, 5*, 221–265.

Pollak, I. V. (1982). Microprocessor-based communications system for the non-verbal person with serious motor handicaps: A preliminary report. *Bulletin of Prosthetics Research, 19*, 7–17.

Pollak, I. V., Gallagher, B. (1989). A fast communication aid for non-verbal subjects with severe motor handicaps. *Journal of Medical Engineering and Technology, 13*(1/2), 23–27.

Somberg, B. L. (1987). A comparison of rule-based and positionally constant arrangements of computer menu items. In *Proceedings of CHI'87* (pp. 255–260). New York: Association for Computing Machinery.

Vanderheiden, G. C. (1988). A unified quantitative modeling approach for selection-based augmentative communication systems. In L. Bernstein (Ed.), *The vocally impaired.* Philadelphia: Grune and Stratton.

ISAAC

ISAAC (International Society for Augmentative and Alternative Communication) is a multidisciplinary organization devoted to the field of Augmentative and Augmentative Communication (AAC). ISAAC has 2,300 members in more than 47 countries, including 11 national/regional chapters. Membership is international and includes all persons who are interested in AAC.

The purpose of ISAAC is to advance the transdisciplinary field of AAC; facilitate information exchange; and focus attention on work in the field.

ISAAC MEMBERSHIP INFORMATION

An ISAAC chapter is a national, regional, or language group of members who address ISAAC's mission at the local level by acting as advocates for the development of AAC within their nation or region. Activities of the chapters include local conferences, newsletters and other publications in the local language(s), and involvement in national policy-making issues. ISAAC strongly encourages and promotes membership in its chapters. **Chapter membership includes membership in ISAAC.** If you live in a country or region *where there is an ISAAC chapter,* please contact the chapter in your area to receive further membership information and current membership rates. Chapter addresses are listed below.

ISAAC-CANADA
c/o Liz Baer, Secretariat
49 The Donway West, Suite 308
Toronto, Ontario M3C 3M9 Canada
Tel. 416-385-0351
Fax 416-385-0352
E-mail isaac_mail@mail.cepp.org

ISAAC-SUOMI-FINLAND
c/o Seppo Haataja
VTT Tietotekniikka, PL 1206
Tampere FIN-33101 Finland
Tel. +358 31 316 3334
Fax +358 31 317 4102

ISAAC-IRELAND
c/o Martine Smith
School of Clinical Speech &
Language Studies
Trinity College, 184 Pearse Street
Dublin 2, Ireland
Tel. +353 1 702 1496
Fax +353 1 671 2152

ISAAC NETHERLANDS-FLANDERS
c/o Margriet Heim
Instituut Algemene Taalwetenschap
Spuistraat 210
1012 VT Amsterdam, The Netherlands
Tel. +31 20 525 3851
Fax +31 20 525 3052

ISAAC-SVERIGE
c/o Bodil Andersson
Amu Hadar
Box 8166
Malmo S-200 41 Sweden
Tel. +46 40 32 19 68
Fax +46 40 32 19 90

USSAAC
c/o James F. Neils
Conferences Inc.
516-26 Davis Street
Suites 211-212
Evanston, IL 60201-4644 USA
Tel. (708) 869-2122
Fax (708) 869-2161

ISAAC-DANMARK
c/o Mogens Hygum Jensen
DLH Esbjerg, Skolebakken 171
DK-6705 Esbjerg 0, Denmark
Tel. +45 75 14 17 22
Fax +45 75 14 31 68

ISAAC-GSC
c/o Paul Andres
Nordfeld 8
31832 Springe-Bennigsen, Germany
Tel. +49/5045/1331
Fax +49/5045/8265

ISAAC-ISRAEL
c/o Judy Seligman-Wine
P.O. Box 1567
Ramat Motza
Jerusalem, Israel
Tel. +972 2 346078
Fax +972 2 340581

ISAAC-NORWAY
c/o Britta Nilsson
Smagruppesenteret-Rikshopspitalet
Pilestredet 32
Oslo N-0027 Norway
Tel. +47 22 868467
Fax +47 22 868458

ISAAC-UK
c/o Caroline Gray
ACE Centre, Ormerod School
Waynflete Road
Headington, Oxford OX3 8DD England
Tel. +44 1865 63508
Fax +44 1865 750188

If you live in a country or region *where there is not an ISAAC chapter,* please contact the ISAAC Secretariat for current application form and membership rates. **Nancy Christie, Executive Director, ISAAC Secretariat, 49 The Donway West, Suite 308, Toronto, Ontario, Canada M3C 3M9; Tel. 416-385-0351; Fax 416-385-0352, E-mail isaac_mail@mail.cepp.org.**