Modeling the Speed of Text Entry with a Word Prediction Interface

Heidi Horstmann Koester and Simon P. Levine

Abstract—**This study analyzes user performance of text entry tasks with word prediction by applying modeling techniques developed in the field of human-computer interaction. Fourteen subjects transcribed text with and without a word prediction feature for seven test sessions. Eight subjects were able-bodied and used mouthstick typing, while six subjects had high-level spinal cord injuries and used their usual method of keyboard access. Use of word prediction decreased text generation rate for the spinal cord injured subjects and only modestly enhanced it for the able-bodied subjects. This suggests that the cognitive cost of using word prediction had a major impact on the performance of these subjects. Performance was analyzed in more detail by deriving subjects' times for keypress and list search actions during word prediction use. All subjects had slower keypress times during word prediction use as compared to letters-only typing, and spinal cord injured subjects had much slower list search times than able-bodied subjects. These parameter values were used in a two-parameter model to simulate subjects' word entry times during word prediction use, with an average model error of 16%. These simulation results are an encouraging first step toward demonstrating the ability of analytical models to represent user performance with word prediction.**

I. Background

COMPUTER-BASED augmentative and alternative communication (AAC) systems provide people who have severe disabilities with the opportunity to communicate independently in the areas of speech, writing, and computer applications. A major goal in the design and prescription of these systems is to provide the user with the fastest means of communication possible. A variety of techniques designed to enhance user performance are currently used in AAC systems, including word abbreviations [1], [2], message encoding [3], [4], and word prediction [5], [6]. There continues to be a need for greater understanding of the efficacy of these systems.

A primary aim in most rate enhancement approaches is to reduce the motor requirements placed on the user. This is clearly an important goal, since the vast majority of users have severe physical impairments. However, a frequent consequence of reducing motor requirements is to increase the cognitive and perceptual loads on the user [4], [7], [8]. The net balance of this trade-off determines whether the user's overall performance will be enhanced or inhibited with a system [9].

Manuscript received November 17, 1993; revised August 22, 1994. This work was supported in part by the National Science Foundation, in part by the University of Michigan Rackham School of Graduate Studies, and in part by the University of Michigan Rehabilitation Engineering Program.

The authors are with the Rehabilitation Engineering Program, Graduate Bioengineering Program, Department of Physical Medicine and Rehabilitation, University of Michigan, 1C335 University Hospital, Ann Arbor, MI 48109-0032 USA.

IEEE Log Number 9405747.

This paper focuses on user performance with word prediction systems in particular and how it is affected by the trade-off between decreased motor and increased cognitive loads. Word prediction systems attempt to predict the word intended by the user by presenting the user with a set of word choices. Word prediction choices are typically displayed in a short list and are refined as the user selects additional letters. Since many words can be completed by choosing from the list rather than through letter-by-letter spelling, the number of selections required per word can be substantially reduced. Keystroke savings provided by several commercial word prediction systems have been measured in the range of 37–47% [10], with clinical reports ranging from 23–58% [5], [11]-[13].[1]

Keystroke savings represents the extent to which word prediction reduces the motor requirements on the user relative to letter-by-letter spelling. This benefit comes at the cost of additional cognitive and perceptual activities required to use the system. These include the visual search of the word list and the subsequent decision about whether the list contains the desired word. An additional source of cognitive load may be the processing involved in planning use strategies (e.g., deciding when to search) and guiding overall activity [14]–[16].

Evidence that these additional cognitive loads can have a negative effect on user performance is shown in Fig. 1. The figure shows the improvements in text generation rate with word prediction as reported in the literature for 13 individuals, relative to the keystroke savings achieved by these individuals [5], [11], [12], [17], [18]. It also shows what the rate improvements *would be* if there were no time cost due to additional cognitive and perceptual activities [9]. All but two of these individuals achieved less than this ideal improvement, which provides indirect yet strong evidence that the additional cognitive and perceptual activities reduce the benefit of decreased motor requirements. More direct evidence comes from our recent study on able-bodied users of scanning systems, in which use of word prediction slowed the rate of selecting items (i.e., letters and/or words) by 30–40% compared to letters-only typing [17].

In addition to providing evidence of cognitive cost, these data also show a large diversity in the effect of word prediction on text generation rate. This diversity may be partially due to differences in methodologies between studies, but it also suggests that the effect of word prediction depends on the

[1] Keystroke saving is measured as 1-(keystrokes required/ characters generated). Keystrokes are broadly defined to include keypresses in a direct selection system, as well as items selected in other ways, such as through scanning or Morse code.

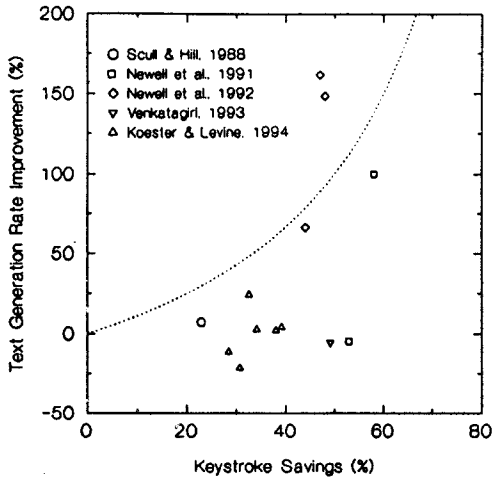

Fig. 1. Reported improvements in text generation rate with word prediction as a function of keystroke savings. Each point corresponds to the performance of a single individual. The dotted line shows what rate improvement *would be* if there were no cognitive time costs associated with use of word prediction.

specific characteristics of the individual user and the context of use. For some individuals, under some conditions, the benefit of keystroke savings seems to outweigh the cost of additional cognitive activities, resulting in a welcome improvement in overall speed of text generation, while for other individuals, the opposite is true. A goal of our research is ultimately to determine the conditions under which word prediction improves text generation rate and those under which it does not.

An empirical approach to pursuing this goal involves measuring the performance of a variety of users under a range of conditions and attempting to deduce the underlying principles of user performance from the resulting data. There is a great need for further empirical information gathered under well-defined conditions, particularly in light of the relatively sparse and diverse data reported to date, as seen in Fig. 1. However, a limitation of an empirical focus is that only a small subset of user-system combinations can be studied, so it is difficult to extrapolate to conditions that have not been empirically examined.

One way to move beyond the limitations of a purely empirical approach is through the development of analytical models that integrate information about the system and user to predict the user's performance [13], [14], [19], [20]. These build on empirical data and support the simulation of user performance across a range of conditions. Accurate analytical models could provide AAC system developers with a means of evaluating the consequences of design decisions, to support the development of an optimal design. For clinicians, models could aid system prescription and configuration by estimating user performance with a range of candidate systems.

The prognosis for user performance modeling in AAC is somewhat controversial. Questions exist about whether accurate models of user performance can be developed, due to variation in the abilities of the user population and the different approaches users may take toward a particular system [12]. The potential benefits of the modeling approach have also been recognized, and several research efforts have been aimed at quantitative model development [13], [19], [20]–[23]. Many of the models developed to date evaluate systems based primarily on motor efficiency [20], [22], [23], so they are not well-suited to represent systems like word prediction, in which cognition and perception have an important impact. Models which have explicitly included cognitive and perceptual processes have provided important conceptual frameworks, but very little work has been done to compare their quantitative predictions to actual user performance [13], [19], [21]. These limitations in model structure and/or the extent of model validation have hampered the success of previous efforts, so the question of whether an accurate model of user performance can be developed in AAC remains open.

Outside of AAC, a great deal of research attention has focused on the development of user interface modeling techniques [15], [16], [24]. The research presented here is based on one such technique, called the Keystroke Level Model (KLM) [15]. This technique provides a means of identifying the cognitive, perceptual, and motor activities 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 activity. 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.

A primary feature of the KLM is its ability to accurately account for user performance using only a small number of parameters. For example, [16] was able to predict the time to enter spreadsheet commands with a 26% error, using only two user parameters, one for keypress time and one for general mental time. The use of the aggregate mental operator illustrates the emphasis of the model on providing useful approximations to cognitive costs, rather than precise psychological models for each cognitive process. It is recognized that this feature is a potential source of model error, but it is important to determine whether useful accuracy can be obtained despite the simplifying assumptions used. Knowledge of the limitations of these assumptions provide guidance for subsequent revisions of the model if necessary. The flexibility in the KLM technique, as well as its proven accuracy in modeling similar tasks, were the reasons for choosing the KLM as the basis for this work.

II. Purpose

This research is part of a long-term program to gain greater empirical understanding of user performance with AAC systems and to develop analytical models that can accurately simulate expected performance. The current study focuses on word prediction and addresses the following specific issues:

1) Available data on user performance with word prediction suggests that the time required for additional cognitive and perceptual processes involved in the use of word prediction will at least partially offset the benefit of decreased motor requirements. This study is intended to extend the available data base by employing both able-bodied and physically disabled subjects across a multi-

session protocol. We hypothesize that text generation rates for all subjects with word prediction will be lower than those expected based solely on consideration of keystroke savings. Further, we propose to examine the source of the cognitive loads in greater detail than has been reported previously, through the derivation of subjects' list search and keypress parameter times.

2) The potential for analytical user performance models in AAC has not yet been realized, and it is not clear if this is due to limitations in previous modeling work or to more serious theoretical problems. A main goal of this study is to address this controversy. We hypothesize that a two-parameter model of user performance can be developed using KLM techniques which will simulate word entry time with an accuracy at least as good as that reported for other applications of the KLM (below 25–30% error). In this initial effort, the two parameter values for keypress and list search times will be derived for individual subjects based on their performance data.
3) Within the general issue of modeling feasibility, a main question is how well a model can accommodate differences between individual users or groups of users. This study addresses part of this broad issue by comparing model accuracy for able-bodied subjects to a group of spinal cord injured subjects. We hypothesize that model simulations will be equally accurate for able-bodied and physically disabled subjects. Any differences in observed performance between these two groups will be accounted for in the model by using different user parameter values within the same two-parameter model structure [25]. We expect the major difference in user parameter values to be in those that represent motor activities, rather than cognitive activities, since these subjects will have only physical disabilities.
4) A fourth goal is to explore the potential of using an analytic model to identify optimal strategies for system use. As a first step toward this long-term goal, we hypothesize that the two-parameter model will be equally successful in simulating performance under different strategies of use. Additionally, we expect user parameter values to be independent of strategy used, even if overall performance is not, since the parameters are intended to represent fairly low-level building blocks of overall performance.
5) Finally, the study addresses the accuracy of model simulations across a range of usage conditions, with the hypothesis that model accuracy will not change as subjects gain experience with word prediction during the experiment or as the keystroke savings of the system is varied. As in #3 above, any differences in performance with practice will be accounted for by different user parameter values within the same two-parameter model structure.

To test these hypotheses, an experiment was conducted to measure user performance with and without word prediction, as a source of model validation data as well as a contribution to empirical understanding. User performance was modeled using KLM techniques with parameter values derived from the subject data. Actual subject performance was then compared to modeled performance. The modeling results reported here represent one step toward a thorough assessment of the model's accuracy, as a necessary prerequisite to applying the model in clinical and design situations.

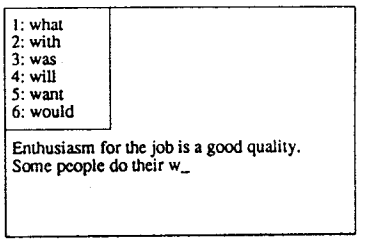

Fig. 2. Schematic representation of the Letters+WP system display. The six-word list is fixed in the upper left corner, with the transcribed text displayed below the list. A sample of actual text transcribed by subjects is shown, and the list contents are those that follow selection of the letter *w*.

III. Methods

A. Subjects

Fourteen subjects were employed. All subjects shared the following characteristics: at least some post-secondary education; frequent computer use and high familiarity with the standard keyboard layout; no significant prior experience with word prediction; and no reported cognitive, perceptual, or linguistic disabilities. Eight of the subjects were able-bodied, and the remaining six had spinal cord injuries at levels ranging from C4–C6.

B. Interfaces

The two interfaces used in the study were developed by the investigators specifically for research purposes, to provide sufficient control over the means of data collection. Both interfaces used direct selection on the standard computer keyboard as the basic input method. Able-bodied subjects used mouthstick typing to access the keyboard, while subjects with spinal cord injuries used their usual method of keyboard access, which was mouthstick typing for two of the subjects and hand splint typing for the other four. The first interface, referred to as "Letters-only," simply involved letter-by-letter spelling. The second interface, referred to as "Letters+WP," 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, as shown in Fig. 2.

C. Experimental Design

The protocol involved a three-session training phase and a seven-session testing phase. The testing phase employed an alternating treatments design, in which subjects' text transcription performance with and without word prediction was recorded in each test session. The keystroke savings provided by word prediction was fixed across Sessions 1–4 and varied in Sessions 5–7 (as discussed in more detail below). Each subject was randomly assigned to one of two different strategies with

TABLE I
THE FOUR SUBJECT GROUPS

	SCI No	SCI Yes
Strategy 1	AB1 (n=4)	SCI1 (n=3)
Strategy 2	AB2 (n=4)	SCI2 (n=3)

which they were to use the word prediction feature. These strategies are discussed in more detail below. The assignment of subjects to the four groups is shown in Table I.

D. Procedures

Subjects were tested in individual sessions which were conducted in laboratory space for eleven subjects and at subjects' homes for the three spinal cord injured subjects who had difficulty arranging travel to the university. Subjects took an average of 21 days to complete the protocol.

Training began with two sessions of practice using the Letters-only system. Each able-bodied subject was provided with a 17″ anodized aluminum mouthstick to use for the duration of the study,[2] while the spinal cord injured subjects used their own mouthstick or typing splints. For able-bodied subjects, the keyboard was placed at standard desk height and tilted at an angle of 45 degrees relative to the desk surface. For spinal cord injured subjects, the keyboard was placed to match their normal set-up; all used a flat keyboard. In the first training session, subjects were instructed in the transcription task and proper use of the mouthstick was demonstrated for able-bodied subjects. Subjects were given the goal of typing as quickly as possible, while keeping mistakes to a minimum. They then practiced for six blocks of text (four sentences each) over two sessions. After each block of text, subjects were asked to rate the difficulty of the task on a continuous scale ranging from "Very Easy" to "Very Difficult."

The third training session introduced subjects to the word prediction feature and their assigned strategy for its use. The rules for the two strategies were defined as follows:

Strategy 1. Search the list before every selection.

Strategy 2. Choose the first two letters of a word without searching the list, then search the list before each subsequent selection.

For both strategies, an exception to these rules occurred when the word list was empty, in which case a list search was not required. These strategies were chosen to be realistic enough to represent at least a subset of actual user approaches, simple enough to be learned in a single training session, and distinct enough to yield measurable performance differences. Subjects were asked to follow the rules as closely as possible. All subjects practiced using their strategy for four blocks of text (4 sentences each), which was sufficient for each to use the strategy correctly without prompting.

Each of the seven test sessions involved four sentences of warm-up using word prediction, an eight-sentence test with word prediction, then a two-sentence typing test. Text blocks were drawn from published typing tests, matched with respect to syllable intensity, average word length, and percent of words that occur with high frequency [26]. The texts were carefully revised to provide the same level of keystroke savings across Sessions 1–4 and systematic variation of keystroke savings in Sessions 5–7. Fig. 3 shows the specific keystroke savings across sessions for each strategy of word prediction use. (Note that the text transcribed was identical for both strategies; the difference in keystroke savings was caused only by differences in the strategies.)

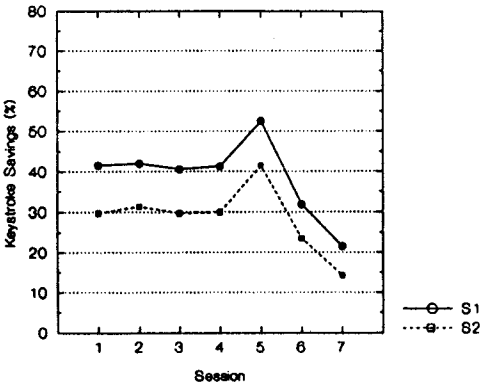

Fig. 3. Keystroke savings provided by the Letters+WP system at each session for each strategy.

Sentences were presented singly on index cards. Subjects were given twenty seconds to read the sentence before an audio cue signalled them to begin transcription. Errors could be corrected by selecting the "Backspace" key as well as a special key for correcting word list selections. The sentence card remained in view for reference throughout transcription.

E. Data Collection

All items selected by subjects were timed and stored by the software in real time. Entries were also encoded to store various information such as the type of selection made (i.e., a single letter or a word list selection) and the number of words in the list when the item was selected. The raw data was used to produce an entry log, in which each line shows the selected item, its various characteristics, and the time at which it was selected.

All sessions were videotaped, with the camera focused on the subject's face, close enough to determine easily the direction of eye gaze. Keypresses were recorded on the video using a mirror, placed behind the subject to reflect a view of the keyboard into the camera, and a speech synthesizer, which echoed the selected item onto the audio track (without being audible to the subject). The camera's clock was synchronized with that of the computer, so the times on the videotape matched those on the entry log.

An experimenter was present throughout each session to record observations of subject behavior. In addition to the difficulty rating described above, subject comments were solicited after each session. Subjects were also given immediate feedback on their text generation rate with each system.

[2] AdLib Incorporated, 5142 Bolsa Avenue, Suite 106, Huntington Beach, CA 92649.

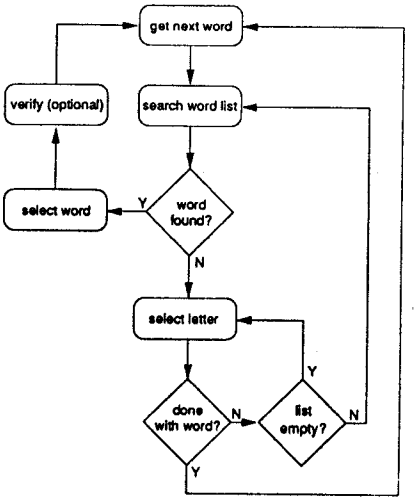

Fig. 4. Flow chart representing activities required during use of Strategy 1 with the Letters+WP system.

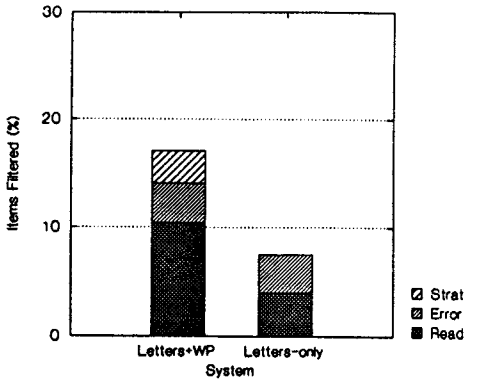

Fig. 5. Amount of data filtered for each system as a percent of total number of items selected, averaged across subjects and sessions. Also shown is the relative contribution of the three filtering categories: "Strat" for items not consistent with the assigned strategy, "Error" for erroneous selections, and "Read" for items following subjects' referral to the text card.

F. Data Filtering

The raw data were filtered to remove events judged to be in any of the following three categories. The first was a general category of text errors and error corrections, including typographical as well as transcription errors. The second included all words that were not entered in a manner consistent with the assigned strategy. Events in these two categories were identified by comparing the subject's generated text to an error-free template. The final category consisted of "card reads," or times when the subject referred back to the text card during transcription, as identified through analysis of the videotape records. Carriage returns and periods were filtered out as well.

The process of identifying and coding events to be filtered was performed by the first author and a trained assistant. Interrater reliability was measured at 99.4%, based on a sample of four sessions analyzed by both raters. A point-to-point reliability measure was used, in which the total number of agreements between raters is divided by the total number of agreements and disagreements [27].

G. Dependent Measures of User Performance

Text generation rates for the Letters+WP and Letters-only systems were measured for each subject at each test session by dividing the number of characters generated during the test by the total time required to generate those characters. Items that were filtered out were not counted either in the number of characters generated nor in the total time.

H. Measurement of User Parameters:

The first step in measuring user parameters was to determine the parameters most important to task execution time. This was done by analyzing the task of entering words for each of the word prediction strategies. As an example, the flow chart of hypothesized user activity for Strategy 1 is shown in Fig. 4. The major activities for both strategies are keypresses (to select a letter or a word) and list searches, so these were chosen as the two user model parameters. While additional parameters could have been defined, only two were used, in order to test the accuracy of a more parsimonious model. A consequence of this choice is that although each parameter primarily represents the activity for which it is named, it may include other components as well. For example, the keypress parameter reflects the motor component of selecting an item, but it may also incorporate more subtle activities such as verifying accuracy, retrieving from memory the next word to be typed, or retrieving the rule that guides the next action to be made. Similarly, the list search parameter may also include these extra activities, in addition to the list search itself. Fortunately, such integration is consistent with the spirit of previous KLM studies, in which a single mental time parameter has often been used for all cognitive actions [15].

Durations for the component actions of list search and keypress while using Letters+WP were derived from the filtered data for each subject. The technique used for this followed the subtractive methods of [15], [16]. Based on the strategy used with Letters+WP, each selection was labelled according to whether it involved a keypress preceded by a list search or a keypress with no list search. For example, when using Strategy 2, the first two letters of every word involved no list searches, so they were labelled as keypress-only. The third letter, however, did include a list search, so it was labelled as a list search-plus-keypress. 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. In all, 98 pairs of parameter values were derived in this way (14 subjects $\times$ 7 sessions).

I. Model Simulations

Using these parameter values, simulations of the time to enter each word during use of Letters+WP were performed as follows. A model value for each item selection was calculated based on whether that selection involved a keypress only (t_k) or a list search-plus-keypress ($t_s + t_k$). The values were summed for each item in a word to yield an entry time for

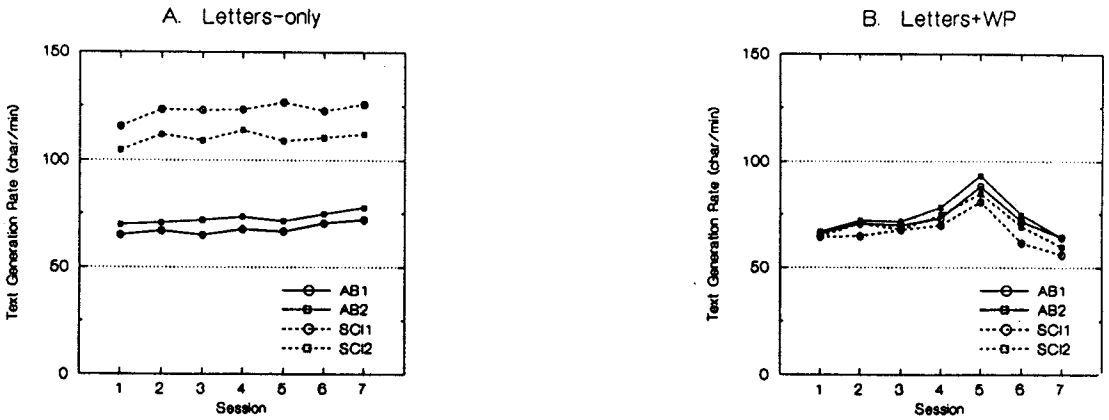

Fig. 6. Average text generation rates achieved by each subject group for each session. (a) shows rates achieved with the Letters-only system, with spinal cord injured subjects significantly faster than able-bodied subjects. (b) shows rates for the Letters+WP system, with no significant differences between subject groups.

that word. This process was performed for each word entered by each subject in every test session, using the parameter values determined for that subject-session combination. Model error was measured for each subject-session combination by averaging the absolute value of percent error across all words in that session.

J. Statistical Analyses

Statistical differences in all dependent measures across subject groups, systems, and test sessions were determined using a repeated measures ANOVA technique. The four experimental factors were strategy and presence/absence of spinal cord injury (SCI) as the between-subjects factors, with system and session as the repeated measures (within-subjects) factors. Statistical significance for each effect was judged at a familywise p-value of 0.05, using the Bonferroni procedure to divide by the number of effects examined within the test [28]. For example, a test analyzing all four experimental factors examines fourteen different effects (four main effects and ten interactions), so the critical p-value used for any one of these fourteen would be $0.05/14 = 0.003$. The corresponding critical values for three- and two-factor tests were 0.007 and 0.017, respectively. Additionally, all p-values examined and reported were those adjusted based on the Greenhouse-Geisser epsilon as an additional precaution against Type I errors (i.e., judging as significant differences that are truly nonsignificant) [28].

IV. Results

A. Filtering

The percentage of data removed from analysis was 16.3% of all Letters+WP selections and 7.3% of all Letters-only selections, averaged across all subjects and sessions. Fig. 5 shows the relative contribution of the three filtering categories to each of these percentages. A four-factor ANOVA showed that the amount of data filtered was independent of subject SCI ($p = 0.314$), strategy used ($p = 0.702$), or session ($p = 0.383$). Significantly more data was filtered from Letters+WP selections than from Letters-only selections ($p < 0.0005$), because subjects referred to the text card more often with Letters+WP ($p < 0.0005$) and because Letters+WP had the additional category of strategy compliance.

B. Simple Empirical Results

A full analysis of all empirical results is beyond the scope and specific goals of this paper, but the major results are presented to provide a context for the parameter derivation and model simulation results.[3]

Fig. 6(a) shows the average text generation rate for the Letters-only system for the four subject groups. Subjects with spinal cord injuries typed an average of 65.8% faster than those without (significant at $p = 0.005$). That they were faster is not surprising given their prior experience with their keyboard access method, while the able-bodied subjects were new to mouthstick typing. The large magnitude of the difference is somewhat surprising, however; the subjects with SCI were clearly quite skilled at Letters-only typing.

Fig. 6(b) shows the average text generation rate with Letters+WP for all four subject groups. The consistency between the groups is striking, and a three-factor ANOVA on strategy, SCI, and session confirmed that there were no differences between any of the groups due to strategy ($p = 0.677$) or SCI ($p = 0.519$). Session was a significant effect ($p < 0.0005$) because the higher keystroke savings provided in Session 5 increased Letters+WP performance for all subjects, while the lower keystroke savings in Session 7 decreased it. Over the first four sessions, in which keystroke savings was fixed, there was also a significant ($p < 0.0005$), but moderate (13%) increase in text generation rate with Letters+WP.

The difference between spinal cord injured and able-bodied subjects re-emerged when the net change in text generation rate using Letters+WP relative to Letters-only was examined (Fig. 7). Analysis of rate improvements using a three-factor ANOVA

[3] While all reported analyses used filtered data, text generation rates based on unfiltered data were also examined, and it was found that filtering did not change the pattern or significance of the empirical results.

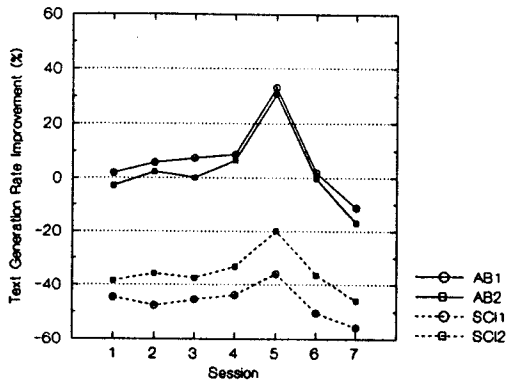

Fig. 7. Average improvement in text generation rate during use of Letters+WP system, relative to Letters-only. A comparison to the shape of Fig. 3 illustrates a generally positive relationship between keystroke savings and rate improvement. However, rate improvement was significantly lower than would have been expected based solely on consideration of keystroke savings. Note also that use of Letters+WP had a strongly negative impact on rate for spinal cord injured subjects.

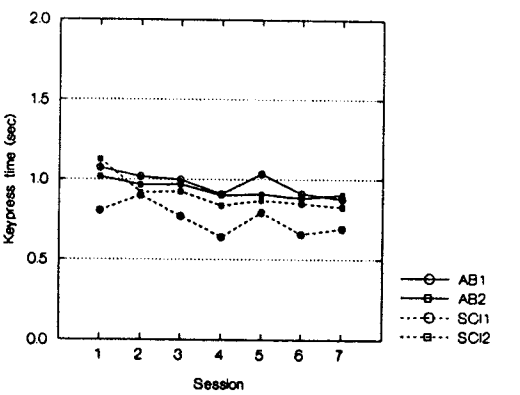

Fig. 8. Average keypress times during use of Letters+WP for each subject group, showing no significant differences between the groups. The small improvement seen with practice was statistically significant.

showed a strong main effect of subject SCI ($p < 0.0005$), so data from able-bodied and spinal cord injured subjects were analyzed separately. For spinal cord injured subjects, use of word prediction had a strongly negative impact on text generation rate; on average, rate decreased by 40.7% when word prediction was used. For the able-bodied subjects, text generation rate was unaffected by the use of word prediction, except during Session 5, which had the highest level of keystroke savings and improved rate by 31.9%, and Session 7, which had the lowest keystroke savings and inhibited rate by 14.0%. Strategy of using Letters+WP had no effect on rate improvement for the able-bodied subjects, while disabled subjects who used Strategy 2 had a significant advantage over those who used Strategy 1 ($p = 0.014$).

Actual rate improvements were also compared to those that would have been expected based solely on consideration of keystroke savings, in order to directly address Hypothesis #1. Actual improvement averaged 70 percentage points below ideal improvement (significant at $p < 0.0005$). This large gap indicates the impact of the cognitive loads experienced by these subjects during use of word prediction, relative to the ideal situation in which cognitive loads have no effect.

C. User Parameter Values

Fig. 8 shows the average keypress times during use of Letters+WP at each session for each of the four subject groups. A three-factor ANOVA on strategy, SCI, and session factors showed that there were no differences between the groups, either on the basis of Letters+WP strategy or spinal cord injury. The one significant difference that did emerge was a main effect of session ($p = 0.001$), as keypress time improved an average of 17.7% from Session 1 to Session 7.

An unexpected result was that keypress times during use of Letters+WP were significantly slower than during Letters-only typing (main effect of system significant at $p < 0.0005$, in a four-factor ANOVA). Fig. 9 shows keypress times with and without word prediction for able-bodied and disabled subjects. (The times are collapsed across strategy of Letters+WP use for greater clarity, since it had no significant effect ($p = 0.632$).) Across both subject groups, keypress times with Letters+WP averaged 23% (170 msec) slower than during Letters-only typing. Visual analysis of the figure suggests that the keypress slow-down was more pronounced for subjects with spinal cord injuries than those without; quantitatively, the amount of slow-down was 48% (270 msec) for SCI subjects, and 10.8% (94 msec) for able-bodied subjects. Statistically, however, the interaction between system and SCI was not quite low enough to be judged significant ($p = 0.009$ vs. criterion of 0.003).

Fig. 10 shows the average list search times at each session for each of the four subject groups. As with keypress time, strategy of use did not significantly affect list search time ($p = 0.058$ in a three-factor ANOVA). Spinal cord injury, however, did have a significant effect ($p < 0.0005$), as the list search times of subjects with SCI were an average of 96.4% (560 msec) slower than the able-bodied subjects. Because of this large effect, within-subjects effects for these two groups were examined separately, using two-factor ANOVA tests. For able-bodied subjects, session had a significant main effect ($p < 0.0005$), with list search time improving by an average of 27.3% (180 msec) from Session 1 to Session 7. For spinal cord injured subjects, however, average list search time improved only 2.7% over these sessions, which was not significant ($p = 0.395$). So in addition to having slower list search times overall, subjects with spinal cord injuries did not improve their search time with practice.

D. Model Simulations

Average model error for each session and subject group is shown in Fig. 11. As discussed above, the model error was measured for each subject by calculating the percent difference between actual and modeled time for each word in a session, then averaging their absolute values. The errors shown in Fig. 11 are averages for the subjects in each group.

As can be seen from the figure, and confirmed statistically through a three-factor ANOVA, model accuracy was not significantly different for the two strategies of Letters+WP use ($p = 0.949$) or for any of the seven test sessions ($p = 0.257$). Because neither strategy nor session had a significant effect, a clearer view of model error was achieved by pooling across

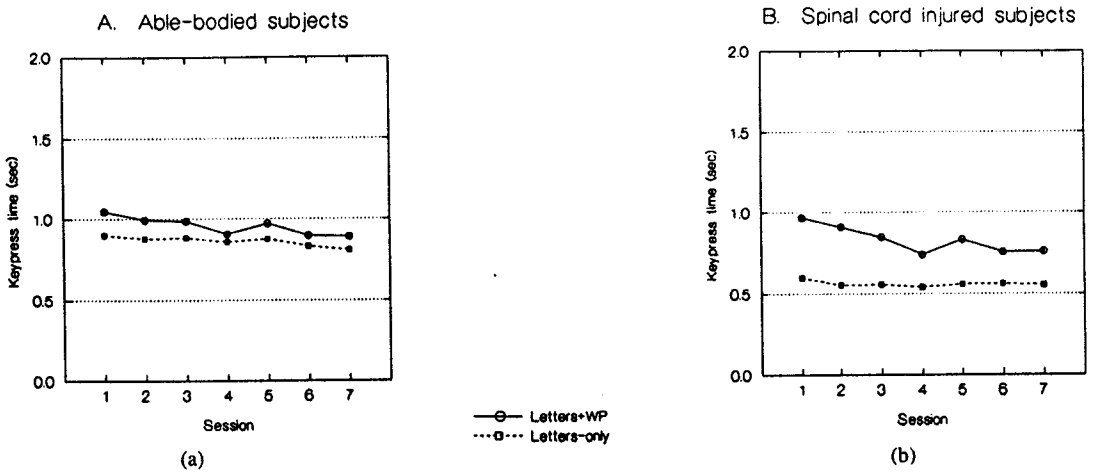

Fig. 9. Average keypress times during use of Letters+WP compared to those during use of Letters-only for (a) able-bodied subjects and (b) spinal cord injured subjects, collapsed across strategy of Letters+WP use. For all subjects, keypress time was significantly slower during use of Letters+WP, and this effect was more pronounced for spinal cord injured subjects.

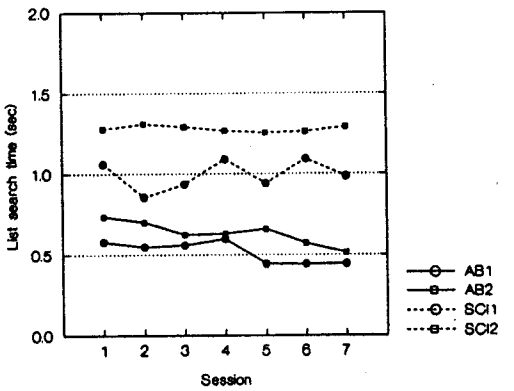

Fig. 10. Average list search times during use of Letters+WP for each subject group. Spinal cord injured subjects had significantly slower list search times than able-bodied subjects.

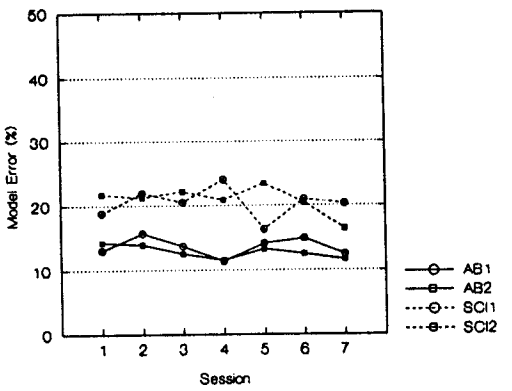

Fig. 11. Average error of the two-parameter model for each subject group. Model accuracy was not significantly different for different sessions or Letters+WP strategies. Accuracy appears to be somewhat better for able-bodied subjects, although the difference was not statistically significant.

these factors to obtain an average error of 13.3% for able-bodied subjects and 20.7% for spinal cord injured subjects. While that difference approaches statistical significance, at $p = 0.019$, it does not meet the criterion p-value of 0.007. In summary, the average model error was 16.5%, with a 95% confidence interval of [13.2, 19.8], and was independent of strategy of Letters+WP use, test session, and subject SCI.

V. Discussion

A. General Empirical Results

The text generation results provide additional support for the hypothesis that increased cognitive and perceptual loads have a major impact on performance with word prediction. For all subjects, any improvements in rate with word prediction relative to letters-only typing were much less than would be expected based on keystroke savings alone. Additionally, a statistically significant improvement was seen only for the able-bodied subjects, and only for the test session that provided the highest keystroke savings. In all other sessions, able-bodied performance with word prediction was not significantly faster than without, while for spinal cord injured subjects, performance with Letters+WP was significantly worse than for Letters-only typing. Finally, the result that subjects referred to the text card significantly more often during use of Letters+WP is also consistent with the hypothesis that word prediction demands a greater cognitive load from its users.

In attributing the additional selection time in word prediction to additional cognitive and perceptual loads, the implicit assumption is that the motor components of the transcription task were the same with and without word prediction. It should be acknowledged, however, that part of the extra selection time may be due to differences in motor activities. For example, searching the word prediction list requires the user to shift eye gaze and occasionally head position. The time required for the motor component of searching may partially depend on the level of the user's physical disability and whether typing was performed with hand splints or mouthstick, but for the spinal cord injured subjects of this study, at least, there was no association between search times and physical method of keyboard access. In addition, the motor aspect of pressing

the keys during use of word prediction may be affected by the need to select from both the number and alphabetic keys. This could be a significant effect for individuals with highly variable motor control ability. However, based on observations of the ability and skill of these subjects, there was no obvious reason to believe that in this study the difficulty of pressing the keys was different for number and alphabetic keys. A more thorough analysis of these motor factors would be a good focus for future work.

Certainly features of the experimental conditions limit the generalizability of these results. Subjects were constrained in what strategy they were to use with Letters+WP, the text they were to generate, and the number of sessions in which they used the systems. The limited time course is perhaps the most important of these factors, since none of the subjects could be considered true word prediction experts by the end of the experiment. Additionally, the spinal cord injured subjects represent only one sub-group of the actual user population, which includes individuals with more variable motor skills as well as those with cognitive impairments. Future work should focus on the performance of users with different abilities and levels of expertise than the subjects studied here, to either corroborate these results or reveal conditions under which word prediction does provide a large improvement in rate.

Hopefully it is clear that our position is not that keystroke savings is irrelevant to text generation rate performance. It *is* an important factor, but the net effect on performance can only be determined by considering keystroke savings in combination with the cognitive cost of using the system. For example, use of Strategy 2 in this study provided lower keystroke savings but yielded performance at least as good as Strategy 1, because the cognitive cost of Strategy 2 was lower, due to the fewer list searches required. Another important example of this phenomenon can be seen in comparing the text generation rate results of the spinal cord injured and able-bodied subjects. The keystroke savings achieved by both groups was the same, but the spinal cord injured subjects did much worse with word prediction, relative to letters-only typing, than did the able-bodied subjects. This suggests that the cost of word prediction was higher for the spinal cord injured subjects. Possible reasons for this are explored below.

B. User Parameter Results

An important benefit of the modeling process is that by partitioning a user's performance into component actions, it is possible to analyze that performance in greater detail. Examination of the values derived for subjects' keypress and list search times shows a similar pattern to the overall text generation rate results. The cognitive cost of using word prediction can be seen in both able-bodied and spinal cord injured subjects, as all subjects spent at least several hundred milliseconds searching the word lists, and all spent longer on keypresses during word prediction use as compared to letter-by-letter spelling. While these effects existed for all subjects, they were much larger for spinal cord injured subjects. Their list search time was almost twice as long on average as that for able-bodied subjects, and their keypress time slowed down by almost 50% during use of word prediction.

We did not expect such a large disparity in word prediction's effect on spinal cord injured as compared to able-bodied subjects. A major source of the disparity may be the difference in the groups' expertise in the Letters-only condition. All subjects had roughly equal familiarity with the keyboard layout. However, the able-bodied subjects had no prior experience in mouthstick typing, while the spinal cord injured subjects had extensive experience with their particular method of keyboard access and had highly developed skills in letter-by-letter typing using that method. When the SCI subjects were asked to use word prediction, additional cognitive effort may have been required to stop themselves from typing the word as they normally would, above and beyond the loads associated with word prediction itself. The able-bodied subjects would not incur these additional loads, since they did not have highly developed motor patterns for letter-by-letter mouthstick typing. Indirect support for this hypothesis is provided by several occasions in which spinal cord injured subjects typed an entire word letter-by-letter, forgetting to attend to the word list.[4]

A general explanation for this effect is that it may be due to negative transfer in switching between systems, in which the skills learned for one system interfere with developing new skills necessary for a second system. This has been observed when people trained on one form of text editor are forced to switch to another one [29]. However, the effect in that case was relatively mild and decreased rapidly with practice, while in this study the effect was large and durable, at least for the spinal cord injured subjects.

One hypothesis for why the negative transfer was so strong in the spinal cord injured subjects is that they may have experienced a qualitative shift in their general mode of information processing in moving from the highly practiced skill of typing to the new task of using word prediction. It is possible that for at least some of these subjects, single letter typing was largely an automatic process, requiring a minimum of cognitive effort, not unlike ten-finger touch typing. In contrast, use of word prediction may have required a mode known as control processing, which is slower and more effortful than automatic processing [30]. Able-bodied subjects would not have experienced this shift, since the relative novelty of both mouthstick typing and word prediction would suggest that control processing would be employed in both cases.

A natural extension to this hypothesis is to consider whether, or under what conditions, use of word prediction could become an automatic process. While it has been suggested that cognitive load may decrease as user familiarity and expertise with the system grows [31], there is very little information on the extent to which this actually occurs. Over the seven test sessions in this study, we observed at most only modest decreases in cognitive load. For example, the slow-down in keypress time during use of word prediction decreased somewhat in the early sessions but then was stable for later sessions (see Fig. 9). Additionally, while list search time did improve by almost 30% for the able-bodied subjects, it didn't improve at all for the spinal cord injured subjects, even though

[4] Such words would be removed from data analysis in the filtering protocol, since they would be judged in noncompliance with the assigned strategy.

the word lists were fixed throughout the experiment (see Fig. 10).

It is certainly probable that more practice would bring greater decrements in cognitive load. However, [30] suggests that performance on a visual search task of the sort required in word prediction is unlikely to become truly automatic even with extended practice, because the words in the list serve as both targets and distractors. On the other hand, even if visual search performance did not improve, a user may improve in the ability to anticipate the list contents, i.e, in deciding when to search the list. It is possible, therefore, that the key to automaticity lies in developing anticipation skills, rather than visual search skills. In this study, the strategy rules limited the amount of anticipation that could be employed by subjects, so these data cannot directly address that issue. Clearly, more empirical work is necessary to address the complex and critical question of automaticity.

C. Model Accuracy

An average error of 16% in modeling word entry times with the Letters+WP system is encouraging and is lower than errors found in other applications of the KLM technique. Model accuracy was no different for different strategies of use, levels of subject experience, or keystroke savings provided by the system. Accuracy appeared to be somewhat greater for able-bodied subjects as compared to spinal cord injured subjects (13% vs. 20% error), but this difference was not statistically significant. These results support our view that user performance with word prediction systems can be successfully modeled using a relatively simple model that considers only keypress and list search actions. Limitations to the particular methodology used here and the KLM technique in general are considered below to provide a balanced interpretation to this result.

Due to the simulation method used, the model accuracy obtained should be considered the best accuracy possible for word entry times with a model structure based on keypress and list search parameters. It is important to recall that the parameter values were derived directly from the performance data. While this is a standard method used in KLM studies, it does yield better accuracy than simulations based on independent parameter values (e.g., taken from another study, or measured from a separate subject group). Additionally, simulations for this study used parameters specific to each subject, which would generally yield better accuracy than using values averaged across subjects. This is particularly true in this case, given the important differences seen in parameter values between able-bodied and spinal cord injured subjects, as discussed above. In this initial study, our goal was to determine if even the best case accuracy was acceptable, so we used the best available estimates for parameter values. Future work is planned to assess the effect of different sources for parameter values, which will be critical in determining the ultimate usefulness of the modeling technique.

A general issue in any model of this kind is that model accuracy depends on the level of detail examined in the performance data. It would be possible, for example, to use the same model structure to simulate the time required to enter all the text in a given session, simply by adding up all the model times for the individual words in that text. This whole-session simulation would necessarily have greater accuracy than the model times for individual words because positive and negative errors for particular words would offset each other in the summing process.[5] The reverse holds true when examining the model's accuracy at predicting the time for each item selection in a given word. Word entry times were used as the focus here because they provide a more stringent test of model accuracy than overall session times and they form a coherent unit task within the overall task of text entry.

A second general issue is that although two user parameters were sufficient to successfully model performance time, they may not exclusively represent only the two processes of a pure keypress and list search. An inexact match between model parameters and underlying processes is not unexpected in a KLM-based model, given its emphasis on useful approximations. The empirical data suggest that such a mismatch may exist in this case, and in particular that there may be additional cognitive processes executed by the spinal cord injured individuals. The slow-down in keypress time seen in all subjects during use of word prediction suggests that there may be some general cognitive overhead that is not required during letters-only typing. That this slow-down was more pronounced in spinal cord injured subjects may mean that the cognitive overhead was greater for them, or it may be related to some specific process such as verification which the able-bodied subjects did not perform. For the list search parameter, the large difference between the spinal cord injured and able-bodied subjects suggests that the underlying processes may be somewhat different for these two groups. Certainly all subjects executed some process to identify whether the target word was in the list, but in addition to this, the time measured as "list search time" could include such activities as retrieval of the strategy rules, verification of selection accuracy, and/or movements of the eyes and head, as has been discussed. Future work is necessary to gain more understanding of what these processes are and the conditions under which they are executed.

VI. Conclusion

Under the experimental conditions studied here, the cognitive cost of using word prediction largely overwhelmed the benefit provided by keystroke savings. Spinal cord injured subjects appeared to incur higher cognitive costs than able-bodied subjects, possibly due to their prior expertise in typing without word prediction.

A two-parameter model, based on a linear combination of keypress and list search actions, was shown to account for subjects' word entry times with an average error of 16%. The accuracy of model simulations was not significantly different for subjects with and without physical disabilities and for two different strategies of word prediction use. However, user parameter values, particularly the list search parameter, were

[5] In fact, with the simulation methods used here, in which model parameters were derived directly from the performance data, the error across an entire session would be precisely zero.