

Learning and Performance of Able-Bodied Individuals Using Scanning Systems with and without Word Prediction

Heidi Horstmann Koester, M.S., and Simon P. Levine, Ph.D.

Rehabilitation Engineering Program, Department of Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, Michigan

This study examines how the cognitive and perceptual loads introduced by a word prediction feature impact learning and performance. Two groups of able-bodied subjects transcribed text using two row-column scanning systems for 10 consecutive trials each. The two systems differed only in that one system had a word prediction feature. Subject groups differed in their order of system use. The results show that, under the conditions of this study, the word prediction system was not substantially more difficult to learn, but it did not yield a statistically significant improvement in text generation rate. This suggests that the cost of using this word prediction system balanced the benefit of the keystroke savings achieved by these subjects. The relationship between keystroke savings, cost in item selection rate, and improvement in text generation rate is explored in order to provide insight into this outcome.

Key Words: Assistive technology—Augmentative communication—Rate enhancement—Word prediction—User performance modeling.

A wide range of assistive technology systems has been developed to facilitate function in a variety of areas, including powered mobility, environmental control, augmentative communication, and computer access. All of these systems include a user interface, which accepts some type of user input to control the system in the desired fashion. In many instances, the interface is designed with a primary

focus on utilizing the motor abilities of the intended user as efficiently as possible.

While the goal of improving motor efficiency is an important one, it has been recognized that this may also place increased cognitive and perceptual requirements on the user, leading to unknown effects on the user's ability to learn and use the system. This dilemma exists in almost every area of assistive technology (1), but it has been discussed most frequently in connection with computer access and augmentative communication (AAC) systems, especially those that employ a rate enhancement feature such as word abbreviations (2), message encoding (3,4), or word prediction (5-7).

This paper focuses on the trade-off between improved motor efficiency and increased cognitive-perceptual loads in the context of word prediction systems. These 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 refined as the user selects additional letters. Because many words can be completed by choosing from the list rather than single letter spelling, the number of selections required per word can be substantially reduced.

The motor efficiency of word prediction systems is often measured in percentage of keystrokes saved.¹ Experimental measurements on two different prediction systems show a range of 37-47% keystroke savings over several different types of text samples (8). Clinical data on actual users reveal a broader range of 23-58% keystroke savings (9-11). Many of the clinical reports are anecdotal, with little spe-

Address correspondence and reprint requests to Heidi Horstmann Koester, 1C335 University Hospital, Ann Arbor, MI 48109-0032 U.S.A.

¹ "Keystrokes" are broadly defined to include key presses in a direct selection system, as well as items selected in other ways, such as through scanning or Morse code.

cific information on the conditions under which keystroke savings were measured, but most of them are consistent with Higginbotham's experimentally determined range of 37–47% (8).

While word prediction systems can be successful in reducing the motor requirements for text generation, this alone does not always yield a significant improvement in rate. Figure 1 shows a scatter plot of improvements in text generation rate at different levels of keystroke savings for eight single case reports (9–12). While some users enjoyed substantial improvement relative to letter-by-letter spelling, others improved only marginally or even decreased in speed. The most dramatic example of this is seen in Newell et al. (10), in which one subject's rate doubled (for an improvement of 100%), with a keysaving of 58%, while another subject's rate decreased by approximately 5%, despite an average keysaving of 53%. Additionally, at least two clinical case studies report that while efficiency may improve substantially, text generation rate may not (13,14).

These data support the long-standing hypothesis that using a predictive system to decrease the number of necessary selections may increase the time required to make each selection, leading to unknown effects on overall performance (4,5,7,15). One way of conceptualizing this trade-off is through the performance model developed by Rosen and Goodenough-Trepagnier (16). In this model, the average time per word, τ , is expressed as $\tau = C \cdot L \cdot T$, where C is the linguistic cost, or the average number of selection units per word, L is the average number of acts required per selection unit, and T is the average time per act. When word prediction is added to a letter-by-letter spelling system, the keystroke savings yields a decrease in C, while L is unchanged. However, T may increase because the act of making a selection has increased in complexity. The net impact on τ , the average time per word, depends on the relative magnitude of these changes in C and T.

Empirical information on just how much, and under what conditions, the time per selection may increase has not been reported in previous studies of performance with word prediction. However, several investigators have attempted to estimate the cost of using word prediction by analyzing the component processes involved, such as searching the word list, or deciding whether to search in the first place (6,7,12,17). One such analysis estimated the extra time per selection to be 1.22 seconds (7). These analyses help explain why time per selection may increase when using word prediction, but their quantitative accuracy has not been verified.

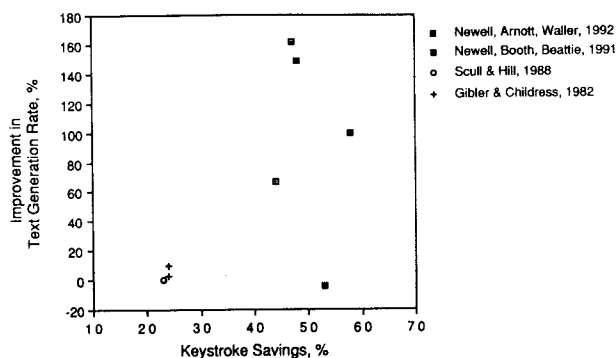


FIG. 1. Improvements in text generation rate reported for various levels of keystroke savings. Each point in the scatter plot corresponds to the performance of a single individual. Note the lack of a clear relationship between rate improvement and keystroke savings.

RESEARCH QUESTIONS

The general goal of our research is to improve understanding of the trade-off between increased cognitive-perceptual requirements and decreased motor loads in assistive technology systems, with a current focus on word prediction systems. In particular, we would like to eventually define the conditions under which word prediction improves text generation rate and those under which it does not. These conditions involve characteristics of the user, the specific implementation of the system, and the particular way in which the user employs the system. Ultimately this understanding may provide a means of simulating the effect of different conditions on overall performance, which would be a potentially powerful tool for designers as well as clinicians (6,7).

This paper reports on recent empirical and theoretical progress toward these goals. The empirical data come from a study in which able-bodied subjects used scanning systems with and without word prediction. Theoretical concepts are presented to explore the underlying reasons behind the results obtained.

METHODS

Subjects

Six able-bodied subjects were employed. All subjects were graduate students who had no cognitive, perceptual, or linguistic impairments. Each had some conceptual familiarity with assistive technology, but none of the subjects had direct prior experience with the systems studied.

-	E	A	R	D	U
T	N	S	F	W	B
O	H	C	P	V	J
I	M	Y	K	Q	,
L	G	X	Z	.	'
<	r	q			

FIG. 2. Arrangement of letter items in the Letters-only system. Letters are arranged in order of frequency of occurrence in English, with the upper left corner as the most frequently chosen character. The three special items on the bottom row correspond to backspace, carriage return, and quit, respectively.

Interfaces

The two interfaces under study were developed specifically for research purposes, to gain sufficient control over the system configuration as well as the means of data collection. Both interfaces used single switch row-column scanning as the basic input method. The first interface, referred to as "Letters-only," required letter-by-letter spelling, using a fixed, frequency-based, letter matrix, shown in Figure 2. Scanning proceeded continuously row-by-row until the switch was pressed to choose a particular row; each column in that row was then scanned until the switch was pressed a second time to choose the desired item. Scanning then resumed from the top row. The scan speed (i.e., the length of time that a row or column remained highlighted) was fully adjustable, as was an extra "row delay" for the first row and column in the matrix. These parameters were initially set at 750 milliseconds for the scan speed and 250 milliseconds for the row delay.

The second interface, referred to as "Letters + WP," used the same letter matrix augmented by a word prediction feature, shown in Figure 3. Characteristics of the word prediction feature included a six-word list, fixed prediction dictionary, and fixed order of words in the list. The method of item selection was similar to that of the Letters-only system, except that a "half-and-half" scanning pattern was used, in which scanning first alternated between the letter matrix and the word list. With this method, the first switch hit chose the desired half. If the matrix was chosen, two more switch hits were required to choose a letter. If the word list was chosen, one more switch hit was required to choose a word. After an item was selected, scanning re-

-	E	A	R	D	U	1: the
T	N	S	F	W	B	2: to
O	H	C	P	V	J	3: is
I	M	Y	K	Q	,	4: that
L	G	X	Z	.	'	5: of
<	bw	r	q			6: you

FIG. 3. Arrangement of letter and word items in the Letters + WP system. The letter matrix is identical to that of the Letters-only system. The four special items on the bottom row per form are for backspace, list selection correction, carriage return, and quit, respectively.

sumed on the matrix half. The timing parameters were fully adjustable; these included the scan speed, the row delay, and an extra delay on the matrix half. (Note that the extra pause on the matrix half was only added on the first scanning cycle.) These parameters were initially set at 750 milliseconds for the scan speed, 250 milliseconds for the row delay, and 500 milliseconds for the half delay.

Experimental Design

The six subjects were evenly divided into two groups. The order of system use was as follows:

Group A

1. Letters-only, training session
2. Letters-only, test sessions
3. Letters + WP, training session
4. Letters + WP, test sessions

Group B

1. Letters + WP, training session
2. Letters + WP, test sessions
3. Letters-only, training session
4. Letters-only, test sessions

Training required one session and combined verbal instruction and practice. Subjects were given the goal of achieving their maximum possible rate with each interface while keeping to less than 10% timing errors (i.e., item was correct but not selected at first opportunity) and less than 5% incorrect selections (i.e., both corrected and uncorrected errors). During Letters + WP training, the rationale behind word prediction was explained, but subjects were not given specific guidelines or strategies for when to use the feature. Subjects practiced using

the system at the relatively slow initial speed on a text sample until they could select text with 95% accuracy. This criterion was generally reached within two sentences.

Testing sessions occurred twice a week and involved two text transcription tests, each preceded by a warm-up period. All subjects began the study with the initial timing parameters. Before the first test, they had three 2-minute warm-up periods in which to tune the timing parameters to match their skill, with an experimenter available to provide assistance as needed. Parameters could be adjusted in increments of 25, 50, or 75 milliseconds, as long as errors did not exceed the established error criteria. Timing errors were detected and tracked in real time by the software, while selection errors were noted by the experimenter during each session. The limitation on errors helped ensure that the timing parameters were neither too fast nor too slow for each subject. Subsequent tests contained a single warm-up period, and parameters could be modified before or after the warm-up under these same guidelines.

Subjects transcribed 10 unique five-sentence blocks of text for each system. Text blocks were drawn from published typing tests, already matched with respect to average word length, average syllables per word, and percentage of words with a high frequency of occurrence (18). Slight revisions to these tests were made to match overall length and average scan steps across blocks.² Individual text blocks were not precisely matched with respect to the word prediction characteristics (e.g., keystroke savings, proportion of words in the dictionary); the specific values for these at each transcription trial are provided in Table 1. Overall, the dictionary contained 83% of the words in the Letters + WP text blocks, providing an average keystroke savings of 42%.

Subjects read the sentences from index cards, containing one sentence per card. They had 20 seconds to flip to a card and read the sentence. During this period, no selections could be made. Scanning resumed automatically at the end of the "freeze" period, and subjects then transcribed the sentence using the assigned interface. Errors could be corrected by selecting special items for backspacing single letters as well as word list selections. The sentence card remained in view for reference throughout transcription.

² The number of scan steps associated with an item on the letter matrix is the number of rows and columns that must be scanned to reach the item.

TABLE 1. Characteristics of text blocks used in Letters + WP trials

Letters + WP trial	Keystroke savings (%)	Prediction success (%)
1	42	79
2	36	70
3	44	81
4	41	84
5	45	85
6	43	89
7	45	90
8	38	77
9	47	92
10	40	79

Data Analysis

All items selected by subjects were timed and stored by the software in real time. The configuration parameters used during a session were also recorded with the item data. An experimenter was present throughout each session to record observations of subject behavior.

The two primary dependent measures were text generation rate and item selection rate. Text generation rate was measured in characters per minute (cpm), to be independent of word length, and incorporated all characters generated in a trial, including punctuation, timing errors, selection errors, and error corrections. However, the trends described below remain consistent whether or not these factors are included in the text generation rates. Item selection rate for each trial was defined as the number of items that were selected per unit time. Note that for the Letters-only system, the text generation rate and item selection rate were necessarily identical, since each item selection generated only one character. For the Letters + WP system, the text generation rate was necessarily faster than the item selection rate, because each selection generated more than one character, on average, due to use of the word list.

Statistical differences in these rates across groups, systems and trials were determined using a repeated measures ANOVA model, with group or system as the between-subjects factor, and trial as the repeated measures (within-subjects) factor. Interactions between the system and trial factors were examined to identify differences in learning rates.

RESULTS

Analysis of Text Generation Rate

Figure 4 shows the average text generation rate achieved by each group over the 20 trials. Across

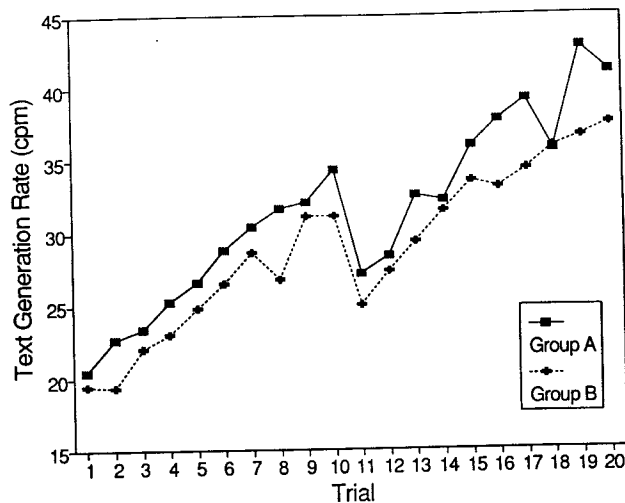


FIG. 4. Average text generation rates achieved by each group at each trial. The discontinuity at trial 10 corresponds to the point at which groups switched to a different system.

all trials, Group A is somewhat faster than Group B. The results can be examined more closely by considering the two halves of the study: the first half where Group A used Letters-only and Group B used Letters + WP for 10 trials each, and the second half where groups switched systems.

Over the 10 trials of the first half, the Letters + WP subjects made an average of 34.8% fewer selections than Letters-only subjects, yet their text generation rate was an average of 8.2% slower. The left half of Figure 4 shows this difference graphically, while Table 2 provides the average and standard deviation for subjects' rates at each trial. The 8% rate difference was not statistically significant, based on an ANOVA test for main effect of system, with repeated measures on trials ($p = 0.442$). From the first to tenth trials, the average rate for subjects using Letters-only ranged from 20.4 cpm (3.6 wpm) to 34.3 cpm (6.0 wpm), with the fastest user achieving 38.0 cpm (6.7 wpm).³ For subjects using Letters + WP, this range was 19.5 cpm (3.4 wpm) to 31.1 cpm (5.5 wpm), with the fastest user attaining 35.7 cpm (6.3 wpm).

Over the second half of the experiment, subjects who used Letters + WP (Group A) made an average of 36.4% fewer selections and had a text generation rate that was an average of 8.7% faster than those who finished with Letters-only (Group B). Table 2 shows the average and standard deviation for subjects' rates over the second half, and the right half of Figure 4 illustrates the differences graphically.

³ Words per minute calculations assume 5.7 letters per word.

TABLE 2. Text generation rate data for each trial

Group	Trial	First half		Second half	
		\bar{x}	SD	\bar{x}	SD
A	1	20.42	2.70	27.12	3.33
	2	22.68	3.57	28.39	3.50
	3	23.41	4.71	32.50	3.97
	4	25.22	4.40	32.12	2.78
	5	26.62	5.35	35.94	3.57
	6	28.82	5.15	37.66	4.84
	7	30.38	4.55	39.07	2.57
	8	31.66	4.45	35.64	3.37
	9	32.11	3.86	42.76	4.54
	10	34.31	3.40	41.06	4.72
B	1	19.47	1.76	24.97	2.89
	2	19.42	0.86	27.27	3.87
	3	22.05	1.53	29.34	4.47
	4	23.03	1.26	31.45	3.90
	5	24.82	1.84	33.46	6.05
	6	26.53	3.19	33.08	5.00
	7	28.59	2.90	34.30	4.65
	8	26.78	3.31	35.82	5.57
	9	31.09	3.97	36.57	3.44
	10	31.07	4.37	37.43	2.12

Means and standard deviations are in characters per minute.

As in the first half, an ANOVA test for main effect of system, with repeated measures for trials, showed that this difference was not statistically significant ($p = 0.418$). From the first to tenth trials, the average rate for subjects using Letters-only ranged from 25.0 cpm (4.4 wpm) to 37.4 cpm (6.6 wpm), with the fastest user at 42.3 cpm (7.4 wpm). For subjects using Letters + WP, this range was 27.1 cpm (4.8 wpm) to 41.1 cpm (7.2 wpm), with the fastest subject achieving 48.0 cpm (8.4 wpm).

There did not appear to be a notable difference in the learning rates between systems in either half of the study. In the first half, average text generation rate improved by 59.6% for Letters + WP subjects and 68.0% for Letters-only subjects. For both systems, then, practice was a major factor in determining text generation rate, and the repeated measures ANOVA showed the effect of trial to be highly significant ($p < 0.001$). Graphically, the rates of improvement for both systems look similar (Fig. 4), which is statistically supported by the lack of a significant interaction between trial and system ($p = 0.553$).

In the second half, practice also significantly improved text generation rate for both groups ($p < 0.001$), with rate improving by 51.4% for Letters + WP subjects and 49.9% for Letters-only subjects. Statistically, the repeated measures ANOVA did

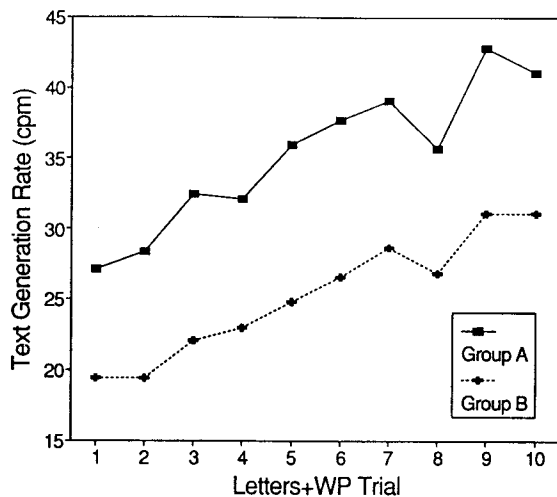


FIG. 5. Average text generation rates achieved at each trial, when each group used the Letters + WP system. Note that this shows Group A's performance for the second half of the study and Group B's performance for the first half.

show a significant interaction between trial and system ($p = 0.005$). This may indicate some difference in learning rates, since the Letters-only group improved more slowly in trials 6–10 than in trials 1–5. However, a clear comparison to the improvement of the Letters + WP group is difficult to make due to the variation in the keystroke savings offered in each text block (see Table 1).

A second set of analyses compared the groups' performances when they used the same system, in an attempt to understand how prior experience with one system might affect subsequent performance on the other system. For example, in comparing the performance of both groups on the Letters + WP system, Group A has already had 10 trials of experience with Letters-only, while Group B has had no prior experience. Therefore, it is not surprising that when Group A subjects switched to Letters + WP, their performance with that system was an average of 39.9% faster than that of Group B (Fig. 5). An ANOVA test for main effect of group, with repeated measures on trials, showed this difference to be significant, at $p = 0.015$.

This experience effect was not quite so pronounced, however, when groups were compared using Letters-only. In this case, Group B subjects have had 10 trials of prior experience with the Letters + WP system, while Group A subjects have had none. When Group B subjects switched to Letters-only, their performance was an average of 18.2% better than that of Group A (Fig. 6), which was not statistically significant ($p = 0.224$). The gap between groups began to close slightly during the last five trials, ending at a 9.1% difference for trial 10.

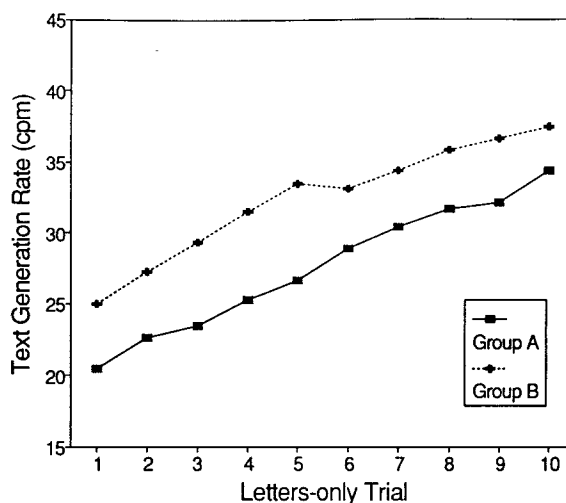


FIG. 6. Average text generation rates achieved at each trial, when each group used the Letters-only system. Note that this shows Group A's performance for the first half of the study and Group B's performance for the second half.

In both of these “same system” comparisons, 10 trials of experience with one system improved the subsequent performance with the other system. Learning rate, however, did not seem to be significantly affected by prior experience. Graphically, the slopes of the lines in each of Figures 5 and 6 are roughly the same, suggesting that the learning rates for groups were similar when each used the same system. Furthermore, neither of the ANOVA tests for differences between groups showed a significant interaction between trial and group ($p = 0.428$ for Letters + WP; $p = 0.463$ for Letters-only).

Analysis of Item Selection Rate

Analysis of the item selection rate data provides some insight into why the use of word prediction did not provide a significant enhancement of text generation rate. For both halves of the study, the item selection rate for Letters + WP users was significantly slower than that of Letters-only users, as shown graphically in Figure 7 and numerically in Table 3. In the first half, the selection rate for Letters + WP users (Group B) was 40.5% slower than that of Letters-only users (Group A). This difference was analyzed using an ANOVA for main effect of system, with repeated measures on trials, and found to be significant ($p = 0.011$). The second half of the study also showed a large difference between selection rates for the two systems, since the rate for Letters + WP selections was an average of 30.9% slower than Letters-only selections (significant at $p = 0.029$).

This decrease in item selection rate means that

