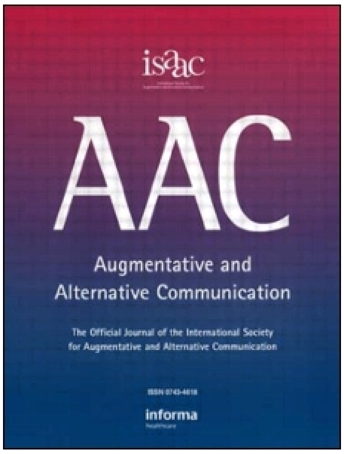


Augmentative and Alternative Communication
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713692248

The effect of context priming and task type on augmentative communication performance

D. Jeffery Higginbotham [a]; Ann M. Bisantz [a]; Michelle Sunm [a]; Kim Adams [a]; Fen Yik [a]
[a] University at Buffalo, New York, USA

First Published:March2009

To link to this Article: DOI: 10.1080/07434610802131869
URL: http://dx.doi.org/10.1080/07434610802131869

Augmentative and Alternative Communication, March 2009 VOL. 25 (1), pp. 19–31

The Effect of Context Priming and Task Type on Augmentative Communication Performance

D. JEFFERY HIGGINBOTHAM[a]*, ANN M. BISANTZ[a], MICHELLE SUNM[a], KIM ADAMS[b] and FEN YIK[c]

[a]*University at Buffalo, New York, USA,* [b]*University of Alberta, Canada, and* [c]*Info-Tech Research Group, Ontario, Canada*

Augmentative and Alternative Communication (AAC) devices include special purpose electronic devices that generate speech output and are used by individuals to augment or replace vocal communication. Word prediction, including context specific prediction, has been proposed to help overcome barriers to the use of these devices (e.g., slow communication rates and limited access to situation-related vocabulary), but has not been tested in terms of effects during actual task performance. In this study, we compared AAC device use, task performance, and user perceptions across three tasks, in conditions where the AAC device used either was, or was not, primed with task specific vocabularies. The participants in this study were adults with normal physical, cognitive, and communication abilities. Context priming had a marginally significant effect on AAC device use as measured by keystroke savings; however, these advantages did not translate into higher level measures of rate, task performance, or user perceptions. In contrast, there were various statistically significant process and performance differences across task type. Additionally, results for two different emulations of human performance showed significant keystroke savings across context conditions. However, these effects were mitigated in actual performance and did not translate into keystroke savings. This indicates to AAC device designers and users that keystroke-based measures of device use may not be predictive of high level performance.

Keywords: Keystroke Savings; Communication Rate; Augmentative and Alternative Communication; Human Factors; Word Prediction

INTRODUCTION

The enhancement of communication rate in augmentative and alternative communication (AAC) has traditionally been regarded as one of the most important areas for research and development (Beukelman & Mirenda, 2005). Communication rates associated with AAC range from 5 to 15 wpm for most direct selection techniques and 2 to 7 wpm for scanning, and are anywhere from 15 to 25 times slower than spoken speech rates. These slow speech rates are frequently noted as a major factor limiting educational and employment opportunities for individuals with complex communication needs (McNaughton & Bryen, 2007).

Over the years, a variety of rate enhancement techniques have been developed to facilitate faster communication rates. One of the most common rate enhancement technologies is word prediction. Word prediction software works by providing the communicator with a list of words and/or phrases predicted to be useful in the ongoing utterance construction. By choosing a word from the list, time and keystrokes may be saved. The word choices may be based on a variety of efficiency factors, including word frequency, recency of word use, and the linguistic context. Lesher and Rinkus (2002) have shown that word prediction systems using probabilistic models of inter-word dependencies accompanied by a large dictionary provide keystroke savings approaching 60%.

*Corresponding author. Center for Excellence in Augmented Communication, Department of Communicative Disorders and Sciences, 122 Cary Hall, University at Buffalo, Buffalo, NY 14216, USA. Tel: +1 716 829 2797. Fax: +1 716 608 1349. E-mail: cdsjeff@buffalo.edu

ISSN 0743-4618 print/ISSN 1477-3848 online
DOI: 10.1080/07434610802131869

However, despite the high efficiency levels associated with word prediction systems, research has shown that the cognitive processing requirements associated with this technique may limit the communication rates that can be achieved with direct selection (Koester & Levine, 1996, 1997). In a study involving individuals with spinal cord injury and persons with normal abilities, Koester and Levine (1996) found that the cognitive costs of using a word prediction system overshadowed any potential benefit associated with the method. Participants with spinal cord injury showed a small but statistically significant difference in text generation rate using different word prediction strategies; however, this effect was not found with the persons with normal abilities. Utilizing user-generated data to inform a variety a cognitive models of word prediction, Koester and Levine (1997, 1998) showed that text production rates varied according to the particular values obtained for list search time, search strategy, key press time, delay between the search and starting a key press, and the average keystroke savings of the word prediction system. Optimal combinations of these human and device factors resulted in fast communication rates (i.e., exceeding 90 characters per minute). Other, less optimal, human factor and device configurations resulted in text generation rates below simple keyboard typing. It should be noted that Koester and Levine's results were obtained using a single task type (text transcription) that did not involve social interaction. In addition, they did not examine the effect of different prediction vocabularies (e.g., specific to the communication topic) on text production performance.

Recent research has focused on ways to supply AAC systems with just-in-time topic-relevant vocabularies using a variety of novel techniques that include: global positioning systems that correlate vocabulary usage with specific locations (Dominowska, Roy, & Patel, 2002); processing the speech of the communication partner using speech recognition and natural language processing, to provide the user with relevant word choices on his or her communication display (Wisenburn, 2005; Wisenburn & Higginbotham, in press), and automated internet searches that prime word prediction dictionaries for topic specific vocabulary (Luo, Higginbotham & Lesher, 2007). Each of these techniques has been designed to improve the availability of contextually relevant vocabulary in order to minimize prediction list search times, reduce keystrokes, and increase communication rate. Thus, contextually relevant vocabularies should, in theory, increase the speed and efficiency of AAC device use. In practice, however, cognitive and interaction demands of the communication task may mitigate the potential enhancements offered by word prediction and other rate enhancement techniques. For example, research by Higginbotham, Kim, and Scally (2007) has shown that interaction communication demands may conflict with the actions necessary to use the AAC device. The researchers found that, in order for the communicators to successfully coordinate turn taking and manage meaning within the timeframe associated with a face-to-face direction giving task, they spontaneously adopted a variety of time-saving interaction strategies, including word abbreviations, telegraphic utterances, and guessing, even when the strategies resulted in the production of partial words and grammatically incomplete utterances. The use of similar interactive strategies has been reported for interactive communication performance across a variety of communication media (telephone, note writing, interactive typing, TTY, instant messaging) (Clark, 1996; Clark & Brennan, 1991). Although Higginbotham et al. (2007) did not use word prediction in their investigation, their work does call into question whether attention-demanding communication technologies like word prediction can be productively used in interactive contexts, and whether communicators can make productive use of enhanced prediction vocabularies when having to actively participate in a face-to-face interaction activity. More generally, Todman and Alm (2003) presented a model regarding AAC communication effectiveness which emphasized the need to demonstrate a link between measures related to utterance production and measures of higher level, individual goals (e.g., self esteem, ability to live independently).

Thus, while researchers in AAC have focused on methodologies and device characteristics that improve specific measures of communication performance, questions remain regarding the degree to which these interventions alter performance at a higher, task-based level. That is, within AAC research, there is an assumption that increasing communicative rate, for instance, is a necessary – and perhaps even sufficient – condition for successful task-based interaction. Instead, examining the ultimate impact of device characteristics such as word prediction requires a shift to a systems-level viewpoint.

Researchers and practitioners in human factors engineering routinely study the interactions of humans, technology, and environments in terms of systems-level task performance (Czaja, 1997); that is, rather than rely on the performance characteristics of a single system component, they consider how tasks are accomplished by systems of humans and equipment in particular situations and physical environments. Using this perspective, AAC use can be considered a system comprising the individual who uses AAC, one or more communication partners, the AAC device, and the particular environment and situation surrounding the interaction. Relevant performance measures would include not only metrics of device or user performance (e.g., communicative rate, keystroke savings), but also measures related to task outcomes (e.g., success of task completion, task completion time), and subjective assessments regarding the interaction (e.g., workload, satisfaction of both the individual who uses AAC and his or her communication partners) (Drury, 1995; Stanton, Hedge, Brookhuis, Salas, & Hendrick, 2004; Wilson & Corlett, 2005).

The current research study was designed to investigate potential improvements in systems as well as device-level performance related to different word prediction techniques implemented in an AAC device. Standard human factors performance measures, such as task errors and completion times, system usability, and subjective evaluation of workload, were collected during three different interactive tasks. If device-level measures are predictive of successful augmented communication, and ultimately task performance, then we predicted a correspondence between more typical device-level measures and these systems-level task measures.

METHODS

Participants

Twenty-four pairs of adults with no disabilities volunteered to participate in this study. Participants were between the ages of 17 and 60 ($M = 23.15$, $SD = 7.531$); 22 were male and 26 were female. Participants were fluent English speakers with no prior AAC device experience, no diagnosis or treatment for hearing loss or for language disorder (self reported) and normal vision. All participants also passed a pure-tone hearing screening (20dB SPL at .25, 0.5, 1, 2, and 4 kHz). One participant in each pair was randomly selected to use the AAC device (designated as the "AAC User"). The other participant in the pair was designated as the "Partner". Most pairs of participants knew each other prior to the experiment.

Materials and Instrumentation

Figure 1 shows the Fujitsu Stylistic ST™ tablet computer used in this study. The touch screen interface measured 26.4 cm diagonally.

Figure 1. Fujitsu tablet computer with Enkidu Impact™ software.

The communication software installed was the Portable Tablet IMPACT Emulator™ (Moulton & Lesher, 2002)[1], which was produced by Enkidu Research, Inc. and the Rehabilitation Engineering Research Center on Communication Enhancement[1]. The AAC software interface included a QWERTY keyboard layout on the touch screen, word prediction of six items in alphabetic ordering, DECtalk™ speech synthesis, auditory feedback on word and sentence at 160 wpm, and a small prediction database of size 1,975 words. A set of portable amplified speakers was connected to the Fujitsu tablet and the participants adjusted the loudness to a comfortable level. Using a protocol established by Higginbotham, Drazek, Kowarsky, and Scally (1994) and Higginbotham et al. (2007), the device was programmed with a 1 s key-press delay to simulate conditions of typical AAC device use.

Tasks

Participants completed three tasks during the experiment: a tangram task, a map task, and a narrative task. These tasks were chosen because they varied in terms of the communication symmetry and interaction requirements of participant roles. We reasoned that the different task demands inherent in each task type could affect AAC device use, and could be assessed by the variety of measures used in this study.

Tangram task

The first task used a tangram puzzle, in which seven puzzle pieces (five triangles, a square, and a parallelogram) are arranged to match a given picture. Each pair of participants was given a picture to be matched along with seven puzzle pieces, as follows: three pieces were given to the AAC User, three pieces were given to the Partner, and one piece (the square) was placed on the table as a starting piece. To facilitate keeping track of the pieces, different colored tiles were provided to the AAC User and the Partner. Participants were told not to move their own tangram pieces. Instead, they were told to take turns instructing the other participant on how to move them. In other words, the AAC User instructed the Partner how to move the AAC User's pieces, and the Partner instructed the AAC User how to move the Partner's pieces. The AAC User gave the first instruction. During task performance, it was noted that participants did not consistently follow task instructions (e.g., moved their own piece, did not follow turn-taking protocol). To avoid disrupting the task flow, the researcher did not intervene. The tangram task was regarded as being the most symmetrical and collaborative of the three experimental tasks, in terms of the task roles and because it required active participation by both participants. In addition, the referents (puzzle pieces) were visibly available to both participants at the same time.

Map task

The map task was taken from the HCRC (Human Communication Research Centre) Map Task Corpus (HCRC, 1991). In this task, the AAC User was given a map on which a route was marked; the Partner was given a map without a marked route (see Appendix A). Both maps showed a number of landmarks (e.g., a tree, a telephone booth). The maps were slightly modified from the original map corpus in that landmark labels that reflected British naming conventions were changed to American usage. To increase task complexity, the maps (as designed) were not identical (e.g., landmarks were added or removed, or their position was altered). Prior to the start of the task, the participants were told that the maps may not be identical. Participants were seated opposite each other so that they could not see the other's map. The AAC User described the route to the Partner, who was instructed to draw the route on the unmarked map. The instructional nature of the map task resulted in less symmetrical role relationships than the tangram task (i.e., instruction giver, instruction follower) and required active role-specific participation by both participants. Referents were available but not shared by the participants: each participant had his or her own map, with a few discrepancies between maps.

Narrative task

The narrative task was taken from the Resource Allocation Paradigm of Pittsburgh (RAPP) (Doyle et al., 2000; Doyle & McNeil, 1998). The AAC User read a short paragraph about a baseball game silently for 5 min. Then the story was removed, the AAC User was given pictures illustrating the story (to provide memory cues), and was instructed to tell the story to the Partner using the device. A 10-question test of comprehension was then given to the Partner. This task was highly asymmetrical with respect to communication roles and responsibilities, in that the AAC User served as a story-teller, while the Partner was to listen, understand, and remember the story. Communication referents were not physically present but were conveyed linguistically by the AAC User.

Conditions: Context Vocabulary Priming

The AAC device was trained on task-specific vocabulary. Prior to the experiment, 10 pairs of fluent English-speakers (different from those who performed the experiment), with little or no AAC experience, performed two repetitions each of the tasks (narrative, map, and tangram). The vocabulary entered into the AAC device during the performance of the tasks was captured into a logfile. Logfiles from all pairs and repetitions for the same task were combined into a master logfile. These master logfiles were then used to prime the device separately, for each task, for one of the experimental conditions. Priming the device with the task-specific vocabulary increased the chances that the device would predict context-specific words during the task.

To empirically evaluate the potential effect of context vocabulary priming on keystroke savings, we processed each training transcript using Enkidu's word predictor, once with no priming, and once when the word prediction database had been primed with the nine other transcript pairs (Mosteller & Tukey, 1977). To accomplish this task, the Impact word predictor was set to Emulation mode (Higginbotham & Yik, 2004). Each transcript was input into the word prediction emulator, which produced a logfile reflecting the prediction process. The logfile was subsequently analyzed using the Augmentative Communication Quantitative Analysis (ACQUA) software program (Lesher, Moulton, Rinkus, & Higginbotham, 2003), which analyzed each transcript for keystroke savings. A statistically significant effect for context priming was found for all three tasks, Tangram: $t(9) = 5.67$, $p = 0.0003$, Map: $t(9) = 13.7$, $p < 0.0001$, Narrative: $t(9) = 15.9$, $p < 0.0001$: context priming improved performance as measured by keystroke savings for all three tasks. Effect sizes, measured in terms of keystroke savings, ranged from 25%

for the tangram task, to 15% for the map task, to 12.5% for the narrative task.

Independent Variables

There were two independent variables. The first, context condition, had two levels: context-on and context-off. In the context-on condition, the AAC device was primed for each task with the context-specific vocabularies captured in the training described above. In the context-off condition, the device was not primed. In both cases, the AAC device provided word prediction. The second independent variable was task: All participant pairs completed the tangram, map, and narrative tasks.

Experimental Design

This was a two-factor mixed design. Context condition was a between-subjects factor, while task was a within-subjects factor. Some measures that were task specific were analyzed only with respect to the context condition.

Procedure

Twelve pairs of participants were randomly assigned to the context-on condition, and 12 pairs were randomly assigned to the context-off condition. The AAC User within each pair received 5 min of instruction on using the AAC device, and was given 5 min of practice transcribing text and 5 min of practice answering questions prompted by the experimenter. The task order for each pair was balanced, so that four pairs from each condition performed one of these orders: narrative, map, tangram; map, tangram, narrative; tangram, narrative, map. The narrative task was limited to a maximum time of 20 min while the map and tangram tasks were each limited to 30 min (participants could stop if they finished prior to the maximum time). As part of the narrative task, a 10-question test of comprehension was given to the Partner. In addition, after each task was completed, each participant completed a workload assessment for that task using the NASA TLX (Hart & Staveland, 1988; Norman, Schneiderman, Harper & Slaughter, 1998).

Following the completion of all three tasks, participants were given a 5 min break. Subsequently, a user satisfaction measure, the Questionnaire for User Interaction Satisfaction (Version 7) (QUIS), was used to obtain a measurement of user satisfaction (Norman et al., 1998). Users are given a series of questions (that can be selected and tailored to meet the particular circumstance) and are asked to assign ratings (see Appendix B). In this study, the AAC User answered 47 questions, broken down into five sections: overall user reactions; screen; terminology and system information; learning; and system capabilities. The Partner answered 10 questions based on his or her experience communicating with the AAC User.

Finally, all participants completed three cognitive tests. The tests were (a) a narrative comprehension test, (b) a card rotation test and (c) a form board test (Ekstron, French, Harman, & Dermen, 1976). For the narrative comprehension test, participants were given a maximum of 10 min to read a passage, and then answered a 10-question comprehension test. The card rotation test and the form board test were designed to assess individual mental ability to perceive forms and conduct mental rotations. In the card rotation test, participants were asked to identify whether a shape was identical to an exemplar, but rotated either clockwise or counter-clockwise, for 20 problems. In the form board test, participants were asked to determine if four shapes could collectively form a single object, for 48 problems.

Dependent Measures

Dependent measures consisted of cognitive test performance measures, task process measures, task performance measures, and user satisfaction measures.

Cognitive Performance Measures

These measures were obtained from the three post-tests, narrative comprehension, card rotation test, and form board test, described above. Scores for these measures were computed as the percentage of correct responses.

Task Process Measures

These measures were total number of words produced, words per minute, and keystroke savings. Process measures were analyzed from log files collected automatically by the AAC device. The ACQUA was used to determine the total number of words used, words used per minute, and keystroke savings (Lesher et al., 2003).

The measure *total number of words* consisted of the number of words that the AAC User produced with the AAC device during communication with the Partner. The measure *words-per-minute* was the rate at which AAC Users produced words to communicate with their Partners. This was computed by dividing the total number of words produced by the total time taken to produce the words, with pauses of 10 s or more removed in order to eliminate non-communicative pauses from the rate calculations (Smith, Higginbotham, Lesher, Moulton, & Mathy, 2006).

Keystroke savings measured the number of keystrokes per word (including backspace, delete word, and speak display) as entered by the AAC User divided by the average number of characters per word in the output produced by the device.

Task Performance Measures

These were the narrative comprehension test score, the map task completion score, narrative comprehension test errors, percent of map task completion, map task error, tangram task completion, task completion time for all tasks, and workload assessment for all tasks. Note that only the Partner completed the narrative comprehension test. Others were measured for each pair of participants (AAC User and Partner).

For the measure tangram task *completion*, the task was counted as complete if the participants believed that they had completed the picture. For the measure tangram *correctness*, a task was counted as correct if, for a completed puzzle, the puzzle pieces were arranged correctly (e.g., it was possible for participants to believe they had completed the puzzle, but be incorrect).

Map task *completion* was the count of successful task completions and map task *completion percent* was computed as the ratio of the length of the path the partner drew to the length of the correct path. The map task *error* measure was the difference between the intended path and that drawn by the partner. As defined by the HCRC protocol (HCRC, 1991), it was computed by overlaying both paths on a 1 cm grid, and counting the number of 1 cm squares that were contained in the area between the two paths (squares that contained the paths were not counted). Essentially, this provided a rough estimate of the area between the paths.

The narrative comprehension test score was obtained by counting the total number of correct responses from the partner. Errors were classified as (a) a *comprehension error* if the information was provided by the AAC User to the Partner, but the Partner did not answer, (b) *no information provided* if the AAC User did not give the relevant information to the Partner, or (c) *incorrect information provided* if the AAC User gave the Partner the incorrect information.

For all tasks, task completion time was recorded as the total time participants needed to finish each task. If participants did not complete the task within the allotted time, the maximum time was used in further calculations. Task workload assessments were obtained from each participant for all tasks using the NASA TLX (Hart & Staveland, 1988).

User Satisfaction Measures

QUIS user satisfaction measures assessed the experiences of the AAC Users with the AAC device and the Partner performing the task. The QUIS score was obtained from individual responses for each question.

RESULTS

Two pairs of participants in the context-on condition were given an incorrect instruction regarding the maximum time allowed for the map task. Therefore, all two-way analyses for measures directly influenced by maximum task time, with task as one variable, were conducted without the data from these participant pairs. These measures with excluded data were task process measures (total words used), task performance measures (map task completion, map task completion percent, map task error, task completion time, workload assessment), and user satisfaction measures.

Table 1 provides descriptive statistics comparing the between-subjects condition, across context conditions, for task process, task performance, and user satisfaction measures. Note that overall, differences between means (provided as a percent of the overall mean) were small (except for keystroke savings), indicating that we saw little impact of the context priming intervention in this study.

Cognitive Performance Measures

There were no differences between context conditions for narrative comprehension, $F(1, 46) = 1.796$, $p = 0.190$, or card rotation, $F(1, 46) = 1.584$, $p = 0.215$. However, the form board test score was significantly lower for context-on participants $F(1, 46) = 17.954$, $p < 0.001$.

Task Process Measures

Task process measures were analyzed with a two-way (Context × Task) mixed ANOVA. There was a main effect of context condition on keystroke savings that approached statistical significance (Figure 2). Participants in the context-on group demonstrated a higher percentage of keystroke savings than participants in the context-off

TABLE 1 Means and SDs for between-subjects variables, comparing context conditions.

	Condition									
	Context-off			Context-on			Overall			
Measure	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	% dif. in means
Keystroke savings	23.56	12.59	36	32.73	13.60	36	28.14	13.81	72	33.0%
W/pm total	4.85	1.74	36	4.60	1.98	36	4.73	1.85	72	5.0%
No. of words	100.92	31.06	36	94.35	42.93	34	97.73	37.17	70	7.0%
Narrative comprehension	7.42	1.56	12	7.92	1.24	12	7.67	1.40	24	6.5%
% map completion	0.94	0.13	12	0.92	0.14	10	0.93	0.13	22	1.0%
Map task error	38.75	12.72	12	43.90	20.50	10	41.09	16.48	22	12.0%
Total task time	1345.53	444.83	36	1205.74	494.39	34	1277.63	471.40	70	11.0%
Workload (TLX score)	61.50	19.56	36	64.44	23.46	34	62.93	21.44	70	5.0%
QUIS score (partner)	5.48	1.05	12	6.03	0.95	12	5.76	1.02	24	9.5%
QUIS score (AAC)	6.19	1.15	12	6.43	0.52	12	6.31	0.89	24	3.8%

Note: Percent difference in means is computed as the difference between conditions divided by the overall mean.

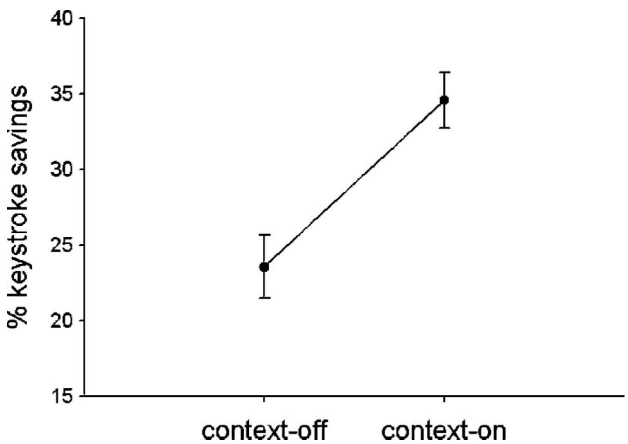


Figure 2. Mean percent keystroke savings in the context-on and context-off conditions. Error bars indicate standard error.

condition, $F(1, 20) = 4.310$, $p = 0.051$. There were no significant effects of context condition for words per minute nor for total number of words.

A main effect of task was found for total number of words $F(2, 40) = 16.081$, $p < 0.001$, and LSD post-hoc analysis showed that the number of words used for the map task was significantly higher than the other two tasks (Figure 3a). A main effect of task was also found for words per minute, $F(2, 40) = 19.719$, $p < 0.001$, shown in Figure 3b. LSD post-hoc analysis, with $p < 0.05$, showed that the narrative and map tasks had significantly higher word per minute scores compared with the tangram task. There was no main effect of task on keystroke savings, $F(2, 40) = 2.475$, $p = 0.097$. There were no significant interactions for any process measures between context condition and task.

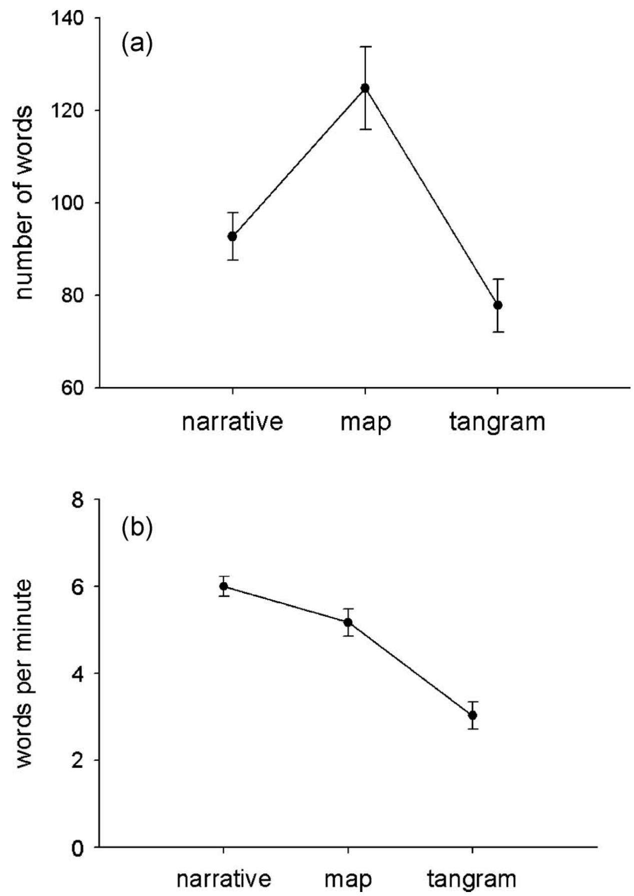


Figure 3. (a) Mean total number of words used across narrative, map and tangram tasks. Error bars indicate standard error. (b) Mean number of words per minute across narrative, map and tangram tasks. Error bars indicate standard error.

Task Performance Measures

Most task performance measures were analyzed separately for each task (since measures were task specific) using a one-way ANOVA, with context condition as the independent variable. There were no significant main effects of context condition for the narrative comprehension test score, $F(1, 22) = 0.753$, $p = 0.395$; map task completion percent, $F(1, 20) = 0.049$ $p = 0.827$; or map task error, $F(1, 20) = 0.520$, $p = 0.479$. Task completion for the map task and tangram task was similar across conditions. For the map task, 7 out of 10 pairs completed the task in the context-on condition and 9 out of 12 completed the task in the context-off condition. For the tangram task, 4 out of 12 pairs believed they completed the task in the context-on condition and 3 out of 12 believed that they completed the task in the context-off condition. Note that only three pairs in each context-condition correctly completed the tangram task.

Narrative comprehension errors are shown in Figure 4. For both context conditions, the incorrect information category accounted for the fewest errors. Also, it appeared that there were more errors in the context-off condition compared to the context-on condition because no information was provided to the Partner by the AAC User.

Task completion time was analyzed using a two-way mixed ANOVA. There was no main effect of context on task completion time, $F(2, 20) = 1.823$, $p = 0.192$. There was, however, a main effect of task as shown in Figure 5a, $F(2, 40) = 38.140$, $p < 0.001$. LSD post-hoc analysis for task completion time indicated that all three tasks were significantly different from each other. There was no task by context interaction effect.

Task workload assessment was analyzed using a three-way mixed ANOVA, with two between-subjects factors: context condition and participant type (AAC User or Partner) and task as a within-subjects factor. There were no main effects of context condition, $F(1, 40) = 0.117$, $p = 0.734$ or participant type, $F(1, 40) = 0.378$, $p = 0.542$. Figure 5b shows a main effect of task on workload assessment, $F(2, 80) = 13.2$, $p < 0.001$. LSD post-hoc analysis indicated that the narrative task had a significantly lower NASA TLX score compared with the map and tangram tasks ($p < 0.05$). There were no significant two- or three-way interactions.

User Satisfaction Measures

QUIS scores were analyzed for AAC Users and Partners separately because questions were role specific. There were no significant effects of context condition on average QUIS scores for the AAC User, $F(1, 20) = 0.173$, $p = 0.168$; or the Partner, $F(1, 20) = 1.786$, $p = 0.196$.

Emulation Versus Human Performance

Using the emulation methodology described above, logfiles containing AAC device output for all the experiments were reprocessed using the Impact Emulator to determine the keystroke savings that could have been obtained for each participant's task performance under conditions of optimal word prediction. First, context-on/context-off results were compared and showed that the context-on condition had significantly

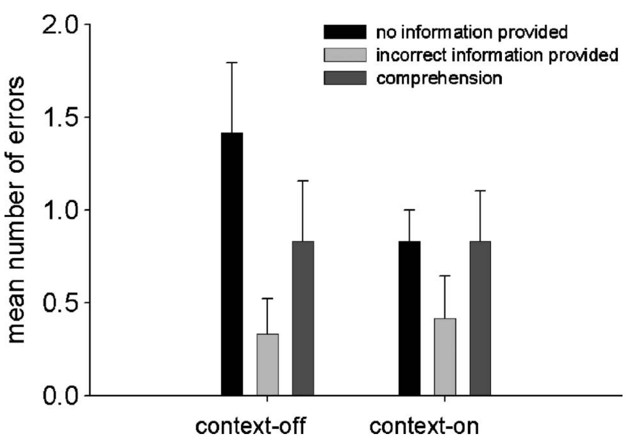


Figure 4. Mean number of errors for comprehension, no information provided, and incorrect information provided types of errors for the narrative comprehension test.

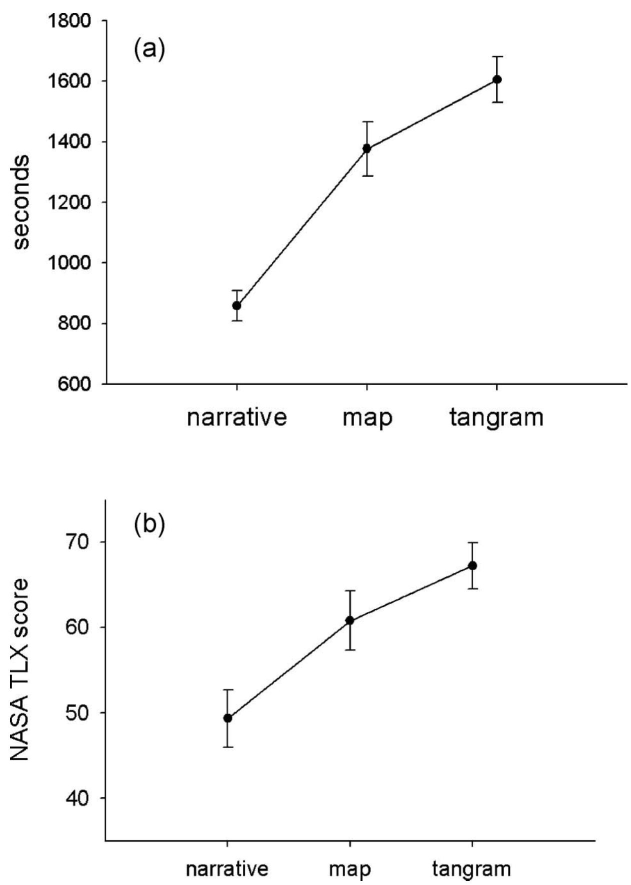


Figure 5. Mean task completion times across narrative, map and tangram tasks. Error bars indicate standard error. (b) Mean workload scores across narrative, map and tangram tasks. Error bars indicate standard error.

greater keystroke savings, $t(70) = 8.5113$, $p < 0.0001$. The keystroke savings percentage obtained during emulation was then compared to the experimental data in order to determine the extent to which actual performance differed from optimal word prediction. A significant difference between actual users and emulated results was found, $t(71) = 12.95$, $p < 0.0001$ and indicated that the actual users did not achieve optimal keystroke savings. The difference in keystroke savings between actual users and emulated performance averaged 15 percent (Users: $M = 0.29$, $SD = 0.13$; Emulation: $M = 0.44$, $SD = 0.07$).

DISCUSSION

Studies that examine the impact of context-based priming and other AAC device characteristics on task performance are critical because ultimately, the intent of AAC devices is to aid individuals in functional tasks and settings (home, school, community). Reducing keystrokes and improving communication rate are important indicators of the effectiveness with which word prediction algorithms or interface manipulations have been implemented. However, direct links between these measures (used commonly in AAC research) and task performance measures (used in human factors research) have not typically been investigated.

Contributions of the Research

In the current experiment, the marginal differences in keystroke savings did not translate to higher level communication or performance measures such as communicative rate, task completion times, task success (as measured by a variety of task specific performance measures), or subjective measures of workload or usability. Thus, this research has empirically demonstrated that device measures that have been typically used to assess AAC device performance may not be predictive of aspects of AAC use in actual communication tasks. Todman and Alm (2003) also emphasized that measures of AAC-generated communication only have value to the extent to which they can be linked to achievement of communicator goals. In contrast to Todman and Alm's (2003) model, however, the goals of our research dealt with the completion of task-related activities, which are an intermediate step towards the accomplishment of personal goals such as self esteem and quality of life that are included in the model. It should be noted that this study was conducted using persons with no known perceptual, cognitive, or motor limitations. Further work in this area should be done in order to clarify implications for clinical practice.

The study also demonstrated clear effects of task on multiple measures of AAC use and task performance. This is a particularly important finding because research in AAC to date has not addressed task-specific performance. The relative utility of AAC devices in different task environments, and AAC device design that can enhance communicative performance in a variety of task environments is a fertile ground for future study. It is important to note that current psycholinguistic research has demonstrated task specific use of communication styles and modes, which suggests that task-specific interface designs may be important for ensuring optimal communication performance (Clark, 1996; Mantovani, 1996).

Taken as a whole, the results of this study indicate that typical device-level measures of AAC performance (e.g., keystroke savings) may not, in actual use, be predictive of task-level performance. As part of this study, two sets of simulation results (based on the vocabulary training transcripts, and the transcripts generated during the experimental tasks) clearly indicated that the context-on condition should provide a significant reduction in keystroke savings. For the training transcripts, the difference in keystroke savings (from the context-on versus context-off conditions) ranged from 10% to 15% across tasks. Analysis of emulated results from the experimental transcripts also showed that significant context effects were possible. However, keystroke savings during the actual task performance was only marginally different between the context-on and context-off conditions. We also showed that emulated performance was significantly better (by an average of 15%) in terms of keystroke savings over actual task performance. These results provide strong evidence that theoretical improvements in efficiency may be attenuated during task performance (i.e., individuals may not always select predicted words, but may type each letter instead), perhaps due to factors such as the interface itself, training, or the need to divide attentional resources across the device, the task, and other communication partners. Persons who use AAC may not notice that the intended word has been predicted and provided on the interface; or may be already engaged in typing the word and decide not to switch production methods (perhaps because the time and effort to switch methods is not perceived to be beneficial in terms of time or workload compared to selecting the predicted word), particularly in the context of a complex interactive task (Koester & Levine, 1996).

Limitations of the Study and Implications for Future Research

This study was performed by participants without disabilities rather than by individuals with disabilities who used AAC for daily communication. This limitation is common among AAC studies (Farrier et al., 1985; Higginbotham, 1989, 1995; Hochstein, McDaniel, Nettleton, & Neufeld, 2003) due to the difficulty in recruiting sufficient numbers of AAC users with similar characteristics for a laboratory-based (rather than field or clinical) study. A protocol established by Higginbotham and colleagues (Higginbotham, 1989; Higginbotham, et al., 2007) was used to simulate restrictions in use that were experienced by some typical AAC device users, by implementing a 1 s keystroke delay, in an attempt to reduce the impact of this limitation.

In addition, participants had limited device training. It may be that their use of the word prediction capabilities of the device would change with further practice. Recall, however, that the device provided word prediction in both context conditions, so that this limitation was unlikely to have differentially impacted participants in the context-on versus context-off conditions. Therefore, while the absolute values related to word prediction use may be altered (and likely, suppressed), comparisons across context conditions can still be made. Future studies should include individuals who use AAC to communicate in their daily lives.

Results may also have been affected by the tasks themselves and the laboratory context in which the study was conducted. It is possible that, for different tasks, there may have been a significant effect of context on both device and task outcome measures. Tasks were chosen, however, to provide a range of interactive experiences and AAC communication roles and thus to generalize to a variety of conditions. Within the experiment, potential task confounds were controlled by including task as a within-subjects variable.

To avoid potential experience effects, context was included as a between- rather than within-participant variable. Differences between groups may have differentially interacted with the context condition. To minimize this, participant pairs were randomly assigned to context conditions. Most of the cognitive tests showed that the characteristics of the groups did not differ, and the difference in the form board test did not appear to have an impact on any task performance measures for the tangram task. Somewhat more pairs in the context-on condition had known each other a longer time compared to the context-off condition, which may have affected the results.

The fact that this research showed a discontinuity between device level (e.g., keystroke savings) and task-level measures is a clear indication that more research is needed to determine the role that rate enhancement techniques and other AAC device characteristics may play in improving functional outcomes for individuals who use AAC. In particular, this research should be extended to include experienced users of AAC in experimental and field settings across a variety of real life tasks.

Finally, the task-related differences in communication rate and total words used may not have accounted for all of the language expressed by the AAC participants during the experimental tasks. Deictic and representational gestures and partner co-construction were noted throughout the experiment, particularly during the tangram and map tasks. Use of such communication modes may reflect communication demands of the particular tasks. This result was not particular to this study and affects all AAC studies that focus on device-specific measures. Future research should be conducted to address this issue.

Acknowledgements

This research was supported by an interdisciplinary grant provided by the University at Buffalo. We would like to acknowledge the contributions of Michael Buckley, Kris Schindler, Greg Lesher, Bryan Moulton, and Victor Paquet to the conceptualization of the study, and Kimberly Ghylin for assistance in data collection.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Note

1. The Impact Emulator is available for research use at http://ubceac.org/?q=node/52

References

Beukelman, D. R., & Mirenda, P. (2005). *Augmentative and alternative communication* (3rd ed.). Baltimore, MD: Paul H. Brookes.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick & J. Levine (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Czaja, S. (Ed.). (1997). *Systems design and evaluation*. New York: Wiley & Sons.

Dominowska, E., Roy, D., & Patel, R. (2002). *An adaptive context-sensitive communication aid*. Paper presented to the CSUN International Conference on Technology and Persons with Disabilities, Northridge, CA.

Doyle, P. J., & McNeil, M. R. (1998) *Resource allocation paradigms of Pittsburgh*. Program demonstration presented to the 1st VA Rehabilitation Research and Development Conference, Washington, DC.

Doyle, P. J., McNeil, M. R., Park, G., Goda, A., Rubenstein, E., Spencer, K., et al. (2000). Linguistic validation of four parallel forms of a story retell procedure. *Aphasiology*, *14*, 537–549.

Drury, C. G. (1995). Designing ergonomics studies and experiments. In: J. R. Wilson & E. N. Corlett (Eds.), *Evaluation of human work* (pp. 113–140). London: Taylor and Francis.

Ekstron, R. B., French, J. W., Harman, H. H., and Dermen, D. (1976). *Manual for Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Farrier, L. D., Yorkston, K. M., Marriner, N. A., & Beukelman, D. R. (1985). Conversational control in nonimpaired speakers using an augmentative communication system. *Augmentative and Alternative Communication*, *1*(2), 62–75.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp.139–183). Amsterdam: North-Holland.

Higginbotham, D.J. (1989). The interplay of communication device output mode and interaction style between nonspeaking persons and their speaking partners. *Journal of Speech and Hearing Disorders*, *54*, 320–333.

Higginbotham, D. (1995). Use of nondisabled subjects in AAC research: Confessions of a research infidel. *Augmentative and Alternative Communication*, *11*, 2–5.

Higginbotham, D., Drazek, A.L., Kowarsky, K., Scally, C., & Segal, E. (1994). Discourse comprehension of synthetic speech delivered at normal and slow presentation rates. *Augmentative and Alternative Communication*, *10*, 191–202.

Higginbotham, D. J., Kim, K., & Scally, C. (2007). The effect of the communication output method on augmented interaction. *Augmentative and Alternative communication*, *23*, 140–153.

Higginbotham, D. J, & Yik, F. (2004). The effect of priming on keystroke savings. CADL technical note. Retrieved 3 October 2007 from: http://ubceac.org/?q = node/69

Human Communication Research Centre (1991). The HCRC Map Task Corpus Annotations Version 1.0. University of Edinburgh & University of Glasgow.

Hochstein, D. D., McDaniel, M. A., Nettleton, S., & Neufeld, K. H. (2003). The fruitfulness of a nomothetic approach to investigating AAC: Comparing two speech encoding schemes across cerebral palsied and nondisabled children. *American Journal of Speech Language Pathology*, *12*, 110–120.

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

Koester, H. H., & Levine, S. P. (1997). Keystroke-level models for user performance with word prediction. *Augmentative and Alternative Communication*, *13*, 239–257.

Koester, H. H., & Levine, S. P. (1998). Model simulations of user performance with word prediction. *Augmentative and Alternative Communication*, *14*, 25–36.

Lesher, G. W., Moulton, B. J., Rinkus, G., & Higginbotham, D. J. (2003). *Software tools for emulation and analysis of augmented communication*. Paper presented at the CSUN International Conference on Technology and Persons with Disabilities, California State University, Northridge.

Lesher, G. W., & Rinkus, G. J. (2002 July). *Domain-specific word prediction for augmentative communications*. Paper presented to the RESNA 2002 Annual Conference, Reno, Nevada.

Luo, F., Higginbotham, D. J., & Lesher, G. W. (2007). *Enhancing word prediction efficiency through webcrawling*. Paper presented to the CSUN International Conference on Technology and Persons with Disabilities, Northridge, CA.

Mantovani, G. (1996). Social context in HCI: A new framework for mental models, cooperation, and communication. *Cognitive Science*, *20*(2), 237–269.

McNaughton, D., & Bryen, D. N. (2007). AAC technologies to enhance participation and access to meaningful societal roles for adolescents and adults with developmental disabilities who require AAC. *Augmentative and Alternative Communication*, *23*, 217–229.

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression*. Addison-Wesley, Reading, MA.

Moulton, B. J., & Lesher, G. W. (2002). *Tablet Portable IMPACT*. Spencerport, NY: Enkidu Research, Inc.

Norman, K., Schneiderman, B, Harper, B., & Slaughter, L. (1998). *Questionnaire for User Interaction Satisfaction* (version 7). College Park, Maryland, University of Maryland.

Smith, L. E., Higginbotham, D. J., Lesher, G. W., Moulton, B., & Mathy, P. (2006). The development of an automated method for analyzing communication rate in augmentative and alternative communication. *Assistive technology*, *18*, 107–121.

Stanton, N., Hedge, A., Brookhuis, K., Salas, E., & Hendrick, H. W. (Eds.). (2004). *Handbook of human factors and ergonomics methods*. Boca Raton, FL: CRC Press.

Todman, J., & Alm, N. (2003). Modelling conversational pragmatics in communication aids. *Journal of Pragmatics*, *35*, 523–538.

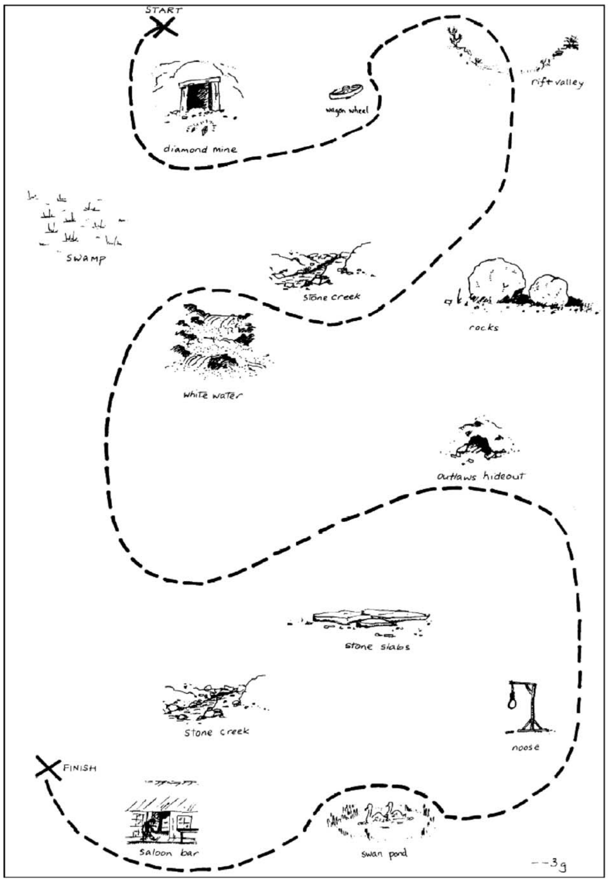
Wilson, J. R., & Corlett, E. N. (Eds.). (2005). *Evaluation of human work*. Boca Raton, FL: CRC Press.

Wisenburn, B. (2005). The use of natural language processing for augmentative-alternative communication utterance generation, (Doctoral dissertation, State University of New York at Buffalo, 2005), *Dissertation Abstracts International*, *65*, 6341.

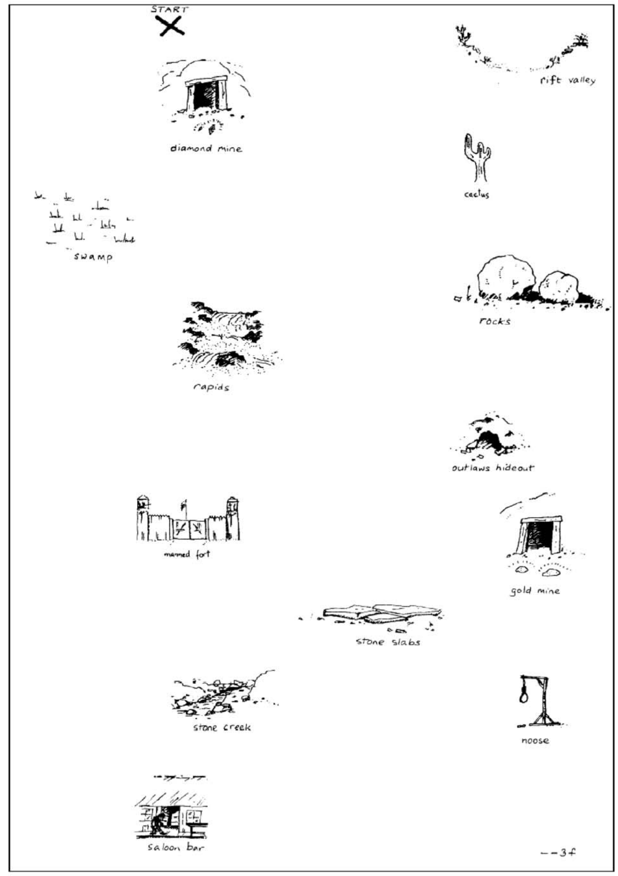
Wisenburn, B., & Higginbotham, D. J. (in press). An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances. *Augmentative and Alternative Communication*.

APPENDIX A

Map Task Stimuli



AAC User



Partner

APPENDIX B

Questionnaires

Sample Questions from QUIS

All questions were answered on a 9-point scale with anchors as shown, as well as a Not Applicable (NA) response. The QUIS had the 5 sections shown below. The actual number of questions in each section is indicated in parentheses. There were 47 total QUIS questions.

Section	Sample Items
Overall User Reactions (6)	*frustrating – satisfying* *rigid – flexible*
Screen (10)	Screen layouts were helpful: *never – always* Amount of information that can be displayed on screen: *inadequate – adequate* Progression of creating utterances: *confusing – clearly marked*
Terminology and System Information (5)	Performing an operation leads to a predictable result: *never – always*
Learning (11)	Learning to operate the system: *difficult – easy*
System Capabilities (15)	The computer-generated sounds are: *not understandable – understandable* Correcting your mistakes: *difficult – easy* Correcting typos: *complex – simple*

Partner Questionnaire

All questions were answered on a 9-point scale with anchors as shown, as well as a Not Applicable (NA) response.

How effective of a communicator was your partner? (*ineffective – effective*)
How many misunderstandings resulted from the way your partner communicated? (*a lot – none*)
How complete were your partner's messages? (*incomplete – complete*)
How much information did your partner offer? (*a little – a lot*)
How well did your partner keep your attention? (*badly – well*)
How was your partner's rate of communication? (*too slow – too fast*)
How smooth was your partner's speech? (*not smooth – smooth*)
How easy were your partner's messages to understand? (*difficult – easy*)
How interesting were the things that your partner communicated? (*not interesting – interesting*)
How well did your partner get to the point? (*did not get right to the point – got right to the point*)