A GENERAL PLAN FOR CONDUCTING HUMAN FACTORS STUDIES OF COMPETITIVE SPEECH DICTATION ACCURACY AND THROUGHPUT

TR 29.2246
Raleigh, NC

James R. Lewis
Human Factors Group
West Palm Beach, FL

Abstract

This report describes a general plan for conducting human factors studies of accuracy and throughput for competitive speech dictation systems. The report covers the topics of (a) distinguishing between studies that are appropriate for measuring accuracy/throughput and those that are appropriate for discovering usability problems, (b) defining different measures of accuracy and throughput, (c) reviewing briefly previous tasks used at IBM to assess speech dictation systems and selecting those to continue to use in future studies, (d) developing efficient experimental designs that are appropriate for the study of up to four dictation conditions in a single study and are also appropriate for between-studies comparisons of measurements, and (e) describing a protocol for collecting accuracy and throughput measures (both performance and satisfaction). Use of these designs should allow developers of speech dictation systems to collect competitive dictation accuracy and throughput in an efficient and flexible manner.

ITIRC Keywords

Speech Dictation
Measurement of Speech Recognition Accuracy

Contents

Introduction

Developers of speech dictation products have a strong interest in measuring the usability of their systems and those of their competitors. Generally speaking, the usability of any system is a function of the extent to which it helps its users accomplish their goals. A highly usable system results in high user performance (both throughput and accuracy) and high user satisfaction. Currently, the development of usable products is an empirical process, and necessarily involves a fair amount of test and evaluation.

This report describes a general plan for conducting human factors studies of accuracy and throughput for competitive speech dictation systems. The report covers the topics of (a) distinguishing between studies that are appropriate for measuring accuracy/throughput and those that are appropriate for discovering usability problems, (b) defining different measures of accuracy and throughput, (c) reviewing briefly previous tasks used at IBM* to assess speech dictation systems, (d) developing efficient experimental designs that are appropriate for the study of up to four dictation conditions in a single study and are also appropriate for between-studies comparisons of measurements, and (e) describing a protocol for collecting accuracy and throughput measures (both performance and satisfaction). Use of these designs should allow developers of speech dictation systems to collect competitive dictation accuracy and throughput in an efficient and flexible manner.

Measurement Studies vs. Usability Problem Discovery Studies

Speech dictation systems have two major components that require investigation to assess system usability. One component is the recognition engine, and the other is the software structure around the recognition engine that supports the performance of system tasks unrelated (or only marginally related) to the act of dictating to the recognition engine. An appropriate assessment of the recognition engine involves the measurement of speech accuracy and throughput using techniques borrowed from traditional experimental statistics. Appropriate assessment of the non-dictation tasks that users perform with the system requires the use of more recently-developed techniques for discovering usability problems (Lewis, 1994).

The approach typically taken to assess non-dictation tasks is to create a set of realistic scenarios, then ask people believed to be appropriately representative of prospective system users to attempt to accomplish the goals described in the scenarios. Observers watch the test participants attempt to complete the tasks and carefully record (1) any problems the users encounter while performing tasks and (2) a set of measurements (including, but not limited to, standardized satisfaction questionnaires, scenario completion times, and rates of successful scenario completion). The measurements taken in these types of studies can serve as benchmarks for comparison with other similar systems, but the primary goal of such studies is the identification and resolution of usability problems. Efficient sample sizes for usability problem discovery studies are typically smaller than those required for precise measurement of system characteristics,

although the precise sample size depends on the average likelihood of problem occurrence. New systems usually have a higher likelihood of problem occurrence, thus requiring a very small sample size for the discovery of the majority of usability problems in the system. More mature systems have a lower likelihood of problem occurrence, and require a larger sample size to enable the discovery of additional usability problems. (For a detailed description of these considerations, see Lewis, 1994.)

Measurement of dictation system characteristics such as accuracy and throughput requires an entirely different approach, especially if the goal of the experiments is to produce data that allows comparison of outcomes between studies and comparison of multiple dictation conditions within a study. The sample size requirements for these types of studies depend on the minimum number of participants required to complete a digram-balanced Greco-Latin design and the expected variance of the measures.

The best estimate of the expected variance of a measure comes from historical data. When the expected variance is small, a relatively small sample size will still produce a small standard error of the mean, which in turn results in a reasonably precise estimate of the mean. (See any standard textbook on experimental statistics for more details about factors that affect the precision of measurement.)

Digram-balanced Greco-Latin designs require only a fraction of the number of test participants required to conduct a complete design, but counterbalance several aspects of the experimental design that might otherwise result in misleading outcomes. These designs are a special class of experimental designs, and efficiently counterbalance the immediate (first-order) sequential effects involved in the presentation of multiple experimental conditions and stimulus sets for within-subjects designs (designs in which all test participants experience all experimental conditions), ensure that all experimental conditions and stimulus sets occur an equal number of times in each position of the sequence, and also balance the pairing of experimental conditions and stimulus sets. (See Lewis, 1993, for more details about these experimental designs.)

The primary purpose of this report is to present a set of digram-balanced Greco-Latin designs appropriate for studying up to four dictation conditions in a single study. Before providing the details of these designs, however, it is important to define different measures of accuracy and throughput and to describe different tasks from previous IBM assessments of dictation systems (along with some important statistical information from global analyses of the data from these previous studies).

Different Measures for Dictation Accuracy and Throughput Studies

For competitive evaluations of speech dictation systems, it is important to precisely define different measures of dictation accuracy and throughput. Following are descriptions of the planned measures for these studies.

Accuracy measures

Primary accuracy. This measure is a percentage. It is the number of words and commands correctly recognized and acted upon during a dictation session (times 100) divided by the total number of words and commands issued. This measure is important because it reflects the percentage of time that the system correctly recognizes input and places no additional demand on the user to correct misrecognized input.

Secondary accuracy. This measure is also a percentage. It is the number of words and commands correctly recognized and acted upon during a dictation session, plus the number of words and commands that might not have been acted upon but appeared in a selection window (times 100), divided by the total number of words and commands issued. This measure is important because it is typically easier to select an alternative from a selection window than to reissue or type the incorrectly-recognized input.

Out-of-vocabulary adjustments. Given system-independent test texts (texts developed from real-world materials, not constrained to contain only words in the vocabulary of the dictation system under test), it is likely that some proportion of the words in the test texts will be out-of-vocabulary (OOV). If it is important to estimate the accuracy of the speech engine for words in vocabulary, it is possible to adjust the primary and secondary accuracy.

First, determine the percentage OOV (OOVPER) for the system under study. After collecting data from a number of participants, you can identify the words that the system misrecognized for all participants. If the system lets you view its vocabulary, you can check the misrecognized words against the internal vocabulary to determine which are OOV. Otherwise, you can create a new user ID and read the misrecognized words into the system. During correction, it is often possible to determine if the misrecognized word was in the vocabulary (for example, it might appear in the correction list, or the system might provide feedback about the introduction of a new word into the vocabulary). The OOVPER is the number of words determined to be OOV multiplied by 100 and divided by the total number of words in the test texts.

Next, calculate the maximum accuracy possible given the OOVPER by subtracting OOVPER from 100.0.

MAXACC = 100.0 - OOVPER

From the accuracy measurements you get from your study, you get the observed accuracy (OBSPER).

Finally, calculate the adjusted accuracy (ADJACC):

ADJACC = OBSPER + (OBSPER/MAXPER)(OOVPER)

The primary assumption underlying this adjustment is that if the OOV words were in vocabulary, they would have the same recognition accuracy as the words that were actually in the vocabulary.

For example, suppose your observed accuracy (OBSPER) from a recognition study was 92.5%, and you determined that 2.5% of the words in the test texts were OOV. MAXACC would be 97.5% (100 - 2.5), and the adjusted accuracy (ADJACC) would be 94.87% (92.5 + 92.5/97.5*2.5).

Throughput measures

Words per minute (WPM). This measure is a rate. It is the number of correctly recognized words issued per minute in a dictation session (where the time for a dictation session includes the time for word entry and correction of misrecognitions). This measure is important because it is the truest measure of throughput from a user's point of view. It does not take into account the number of correctly recognized commands because commands exist in a dictation system to control text characteristics and to allow recovery from recognition error. Thus, a system that requires a user to issue more commands should result in fewer WPM than a system with greater intelligence for correctly recognizing words and identifying situations that allow automatic, correct control of text characteristics (such as capitalization and identification of pluralization versus possession). To the extent that this is true, calculating WPM in this way will allow assessment of the benefit to the user of enhanced system intelligence.

Words per correction (WPC). This is the inverse of the error rate. It demonstrates the nonlinear effect of improving accuracy at higher levels of accuracy. For example, if the accuracy is 90%, the error rate is .10 and WPC is 10. At that accuracy level, users must correct one in every ten words dictated. If the accuracy is 95%, the error rate is .05 and the WPC is 20. Users must correct one in every twenty words dictated. If the accuracy is 99%, then the error rate is .01 and WPC is 100. Users only correct one in every hundred words dictated.

Others per minute (OPM). This measure is a rate. It is the number of non-word actions (others) issued per minute in a dictation session. This measure is important because it reflects the amount of additional user activity beyond the speaking of the words required to enter the words into the system. It replaces the Words and Commands per Minute (WCPM) measure formerly taken in IBM dictation studies. The formula for calculating OPM is:

OPM = ((1 * Class 1 Errors) + (2 * Class 2 Errors) + (2 * Extras)) / Time

Class 1 errors are the number of words and commands not correctly recognized during a dictation session, but that appear in an alternative selection window. A user typically only needs to perform one additional action (such as say a phrase like "Choose 2" or click on an item in the selection list) to correct this type of error.

Class 2 errors are the number of words and commands not correctly recognized during a dictation session and that do not appear in an alternative selection window. A user typically needs to perform two additional actions (such as to type the correct word/command and press the Enter key) to correct this type of error.

Extras are any activities other than saying words and commands or correcting Class 1 and 2 errors that a user performs to obtain perfect copy in a dictation session. For example, if a user coughs or otherwise gets a response to a nonlinguistic sound, then the user must perform actions to delete the word that appears. If there are capitalization or spacing problems in the final copy, then the user must perform actions to correct the format and spacing.

Overhead (OH). This is a new measure intended to assess the demand placed on a user by a dictation system during a dictation session. The formula for OH is:

OH = (OPM * 100) / (WPM)

If, on average, a user must complete two additional other actions to produce the correct dictation of one word, then OH will be 200.0. If the average proportion of others to words is 1:1, then OH will be 100.0. A dictation interface with perfect recognition and perfect language understanding would have an OH of 0.0 because all a user would have to do is to say the words with the correct inflection to indicate appropriate punctuation -- dictation with no overhead.

Three questionnaires

To measure participants' satisfaction with a given dictation system, participants will complete three different types of questionnaires: the ASQ, AQ and PSSUQ. Examples of these questionnaires appear in Appendix A.

After-scenario questionnaire (ASQ). This measure is a rating taken with a 3-item standardized questionnaire (Lewis, 1995). Administration of this questionnaire immediately follows completion of each individual dictation task. (See Lewis, 1995 for details about the development and administration of this questionnaire, and for information about its desireable psychometric properties.)

Accuracy questionnaires (AQ). This measure is also a rating taken with a 4-item questionnaire based on the ASQ, but specialized for the assessment of dictation accuracy. Administration of this questionnaire follows the completion of the ASQ for the final dictation task in a dictation condition.

Post-system usability questionnaire (PSSUQ). This measure is also a rating taken with a 19-item standardized questionnaire plus an additional 4 items (for a total of 23 items). Administration of this questionnaire follows the completion of the AQ for the final dictation condition for a dictation system. (See Lewis, 1995 for details about the development and administration of this questionnaire, and for information about its desireable psychometric properties.)

Definition and Evaluation of Different Tasks Used to Assess Dictation

Over the last three years, dictation studies at IBM have used the following four key tasks, generally presented to test participants in this order:

Task A: Reading text with embedded commands. To perform this task, test participants read from a document that had, mixed in with the text, the required commands (indicated in boldface) for producing punctuation and controlling other text characteristics. For example:

> I just finished reading your book **COMMA** , **BEGIN-CAPITALIZE** The Challenge **END-CAPITALIZE** of the **CAPITALIZE-NEXT** Unknown **COMMA** , and had a few questions about some of your book's assertions **PERIOD** .

Task B: Reading text with embedded asterisks. Rather than embedding the commands in the text, the text in this task had boldface asterisks to indicate the need to utter a command at that point. For example:

> I just finished reading your book ***** , ***** The Challenge ***** of the ***** Unknown ***** , and had a few questions about some of your book's assertions ***** .

Task C: Reading plain text. For this task, the user dictated from text that had no indication of when it was necessary to utter commands. For example:

> I just finished reading your book, The Challenge of the Unknown, and had a few questions about some of your book's assertions.

Task D: Composition. This task required the test participant to compose a short letter on a given topic, such as a job application. The purpose of this task was to have participants dictate in a non-reading situation that would allow assessment of the ease of recovery from misstated words, nonlinguistic utterances, and changes in a user's train of thought.

Intertask correlations

To assess the value of continuing to use all four tasks, I created correlation matrices for both primary accuracy and WPM using the average values from nine dictation system evaluations conducted over the last three years. The intertask correlations from these studies were:

Primary Accuracy		**A**	**B**	**C**
	B	1.00		
	C	1.00	0.99	
	D	0.89	0.93	0.91

WPM		**A**	**B**	**C**
	B	0.97		
	C	0.99	0.95	
	D	0.84	0.80	0.80

Correlations that exceed 0.80 are quite strong, and those that exceed 0.95 are near perfect. Correlations this high indicate that all the tasks appear to be measuring the same thing, and that examination of the data for one task to identify the best system would lead to the same conclusions as examination of the data for a different task. Therefore, to improve test efficiency, future studies of dictation systems will use only two of the four previous tasks: Task A and Task C.

Variance and standard deviations for task measures

Examination of the raw data for a recently completed evaluation of speech dictation systems showed that the average variance for a reading task's primary accuracy (average taken over Tasks B and C for two different recognition systems) was 6.35% (standard deviation = 2.52%). The equivalent average variance for a reading task's throughput was 12.75 WPM (standard deviation = 3.57). This information will be useful for balancing sample size requirements and measurement precision requirements for future studies using the digram-balanced Greco-Latin designs described in the next section of this paper.

Digram-Balanced Greco-Latin Designs for Speech Dictation Studies

This section contains experimental designs that control the influence of several effects that could create misleading data. These designs have the following desirable characteristics:

- Male and female speakers contribute equally to each system measurement.

- Four different text sets contribute equally to each system measurement. This is important because the use of multiple text sets reduces the likelihood that apparent system differences might be due to an unknown interaction between systems and a single text set. Including four text sets allows the development of experimental designs that can assess up to four dictation conditions in a single study (where different dictation conditions might be different dictation systems or meaningful manipulations of the same dictation system).

- For designs that investigate more than one dictation condition, each text set appears an equal number of times in each position of the test sequence.

- For designs that investigate more than one dictation condition, each text set immediately precedes and immediately follows every other text set an equal number of times.

- For designs that investigate more than one dictation condition, each dictation condition appears an equal number of times in each position in the test sequence.

- For designs that investigate more than one dictation condition, each dictation condition immediately precedes and immediately follows every other dictation condition an equal number of times.

- For designs that investigate more than one dictation condition, each text set appears with each dictation condition an equal number of times.

The designs

Design for one dictation condition. This design requires a minimum sample size of eight participants, with one dictation session per participant. Any additional data collection should replicate this design, collecting data from an independent group of eight participants.

Participant	Gender	System	Text Set
1	M	1	A
2	M	1	B
3	F	1	C
4	F	1	D
5	F	1	A
6	F	1	B
7	M	1	C
8	M	1	D

Design for two dictation conditions. This design requires a minimum sample size of eight participants, with two dictation sessions per participant. Any additional data collection should replicate this design, collecting data from an independent group of eight participants. If collecting a second set of data, put text sets across participants as A-C and B-D. If collecting a third set of data, put text sets across participants as A-D and B-C. As shown below, the initial design distributes text sets across participants as A-B and C-D.

		First Condition		Second Condition	
Participant	**Gender**	**System**	**Text Set**	**System**	**Text Set**
1	M	1	A	2	B
2	M	2	B	1	A
3	F	2	A	1	B
4	F	1	B	2	A
5	M	1	C	2	D
6	M	2	D	1	C
7	F	2	C	1	D
8	F	1	D	2	C

Design for four dictation conditions. Because a design for four conditions and four text sets is very efficient relative to a design for three conditions and four text sets, the design for four conditions is described next. This design requires a minimum sample size of eight participants, with each participant going through four dictation sessions. If a sample size greater than 8 is desirable, collect the additional data in groups of eight participants. To enhance the generalizability of the results for this experimental design, randomly reassign text sets to letters and systems to numbers for each additional group of data.

		First Condition		Second Condition		Third Condition		Fourth Condition	
Participant	**Gender**	**System**	**Text**	**System**	**Text**	**System**	**Text**	**System**	**Text**
1	M	1	A	4	B	2	D	3	C
2	F	2	B	1	C	3	A	4	D
3	F	3	C	2	D	4	B	1	A
4	M	4	D	3	A	1	C	2	B
5	F	4	A	1	D	3	B	2	C
6	M	1	B	2	A	4	C	3	D
7	M	2	C	3	B	1	D	4	A
8	F	3	D	4	C	2	A	1	B

Design for three dictation conditions. Due to the arithmetic complexity of crossing three conditions with four text sets, the minimum sample size that achieves a complete digram-balanced Greco-Latin design is 24. Each participant goes through three dictation sessions. If a sample size of 24 is untenable, see the next section for an alternative design that requires 12 participants.

		First Condition		**Second Condition**		**Third Condition**	
Participant	**Gender**	**System**	**Text Set**	**System**	**Text Set**	**System**	**Text Set**
1	M	1	B	3	C	2	A
2	M	2	C	1	A	3	B
3	M	3	A	2	B	1	C
4	F	2	A	3	C	1	B
5	F	3	B	1	A	2	C
6	F	1	C	2	B	3	A
7	F	1	C	3	D	2	B
8	F	2	D	1	B	3	C
9	F	3	B	2	C	1	D
10	M	2	B	3	D	1	C
11	M	3	C	1	B	2	D
12	M	1	D	2	C	3	B
13	M	1	B	3	D	2	A
14	M	2	D	1	A	3	B
15	M	3	A	2	B	1	D
16	F	2	A	3	D	1	B
17	F	3	B	1	A	2	D
18	F	1	D	2	B	3	A
19	F	1	C	3	D	2	A
20	F	2	D	1	A	3	C
21	F	3	A	2	C	1	D
22	M	2	A	3	D	1	C
23	M	3	C	1	A	2	D
24	M	1	D	2	C	3	A

An alternative design for three dictation conditions. If resource constraints prevent consideration of a design requiring 24 participants, the following alternative design only gives up digram-balancing for stimuli, and maintains all other desirable counterbalancing characteristics of the full design while requiring only 12 participants. The absence of digram-balanced stimulus sets is not likely to affect the measurements taken in a study of dictation accuracy and throughput.

		First Condition		Second Condition		Third Condition	
Participant	**Gender**	**System**	**Text Set**	**System**	**Text Set**	**System**	**Text Set**
1	M	1	A	2	B	3	C
2	M	2	C	3	A	1	B
3	M	3	B	1	C	2	A
4	F	3	D	2	B	1	A
5	F	1	B	3	A	2	D
6	F	2	A	1	D	3	B
7	F	1	D	2	B	3	C
8	F	2	C	3	D	1	B
9	F	3	B	1	C	2	D
10	M	3	C	2	D	1	A
11	M	1	D	3	A	2	C
12	M	2	A	1	C	3	D

Adequacy of minimum sample sizes

The following table shows, for sample sizes of 8, 16, and 24, the expected standard errors for the mean and precision associated with a 95% confidence interval around the mean for measurements of primary accuracy and WPM, using the values for the variances of these measures estimated from historical data.

Measure	**Sample Size**	**Standard Deviation**	**Standard Error of the Mean**	**Precision for 95% Confidence Interval**
Accuracy	8	2.52	0.89	1.73
	16	2.52	0.63	1.23
	24	2.52	0.51	1.00
WPM	8	3.57	1.26	2.47
	16	3.57	0.89	1.75
	24	3.57	0.73	1.43

As seen above, most of these experimental designs require a minimum sample size of eight participants to provide complete Greco-Latin counterbalancing. Given the expected variances and eight participants, the true mean accuracy should lie within +/- 1.75% of the observed mean, and throughput measures should be accurate within about +/- 2.5 WPM.

Increasing the sample size to 24 participants would result in accuracy measurement within +/- 1.0% and throughput measurement within about 1.5 WPM. It is likely, for the purpose of these studies, that eight participants will provide adequate accuracy. Although I know of no substantial research on the topic, it does not seem likely that people can detect accuracy differences between systems as small as 1.75% or differences in throughput as small as 2.5 WPM. Thus, sample sizes of eight participants should produce results for which statistically significant differences between dictation conditions would also indicate practically significant differences from a user's point of view. Failure to achieve statistical significance would almost certainly correspond to differences between dictation conditions that are either nonexistent or too small for users to detect.

The text sets

To provide adequate and consistent testing of dictation systems with these text sets, each text set contains two letters, one to have embedded commands appropriate for the dictation system under study, and one to have plain text. (For all text sets, participants will dictate the letter with embedded commands first to help them learn the required commands. The text sets appear in Appendix B.) Each letter in each text set includes text with the following characteristics:

- Proper names
- Initial caps
- All caps
- Date and time
- Phone number
- Monetary amount
- New paragraphs
- New lines
- Plurals and possessives
- Questions and exclamations

Data Collection Protocol

This section contains a description of the data collection protocol to use when collecting dictation measurements. This description covers a single dictation session. If the study involves multiple dictation sessions, then the test monitor would repeat steps 3 to 13, as required. Additional sessions with the same system would drop steps 5-7.

1. Bring the participant into the lab.

2. Have the participant complete nondisclosure form and background questionnaire. (See Appendix C for the background questionnaire.)

3. If the participant needs to enroll to use the system, he or she would enroll.

4. After enrollment, the participant would complete the ASQ for that task.

5. The test monitor demonstrates for the test participant how to dictate the two practice sentences (shown in Appendix D), and discusses any system-specific appropriate dictation strategies.

6. The test monitor gives the participant a list of system-specific functions and commands needed to complete the dictation sessions, and goes over those commands with the test participant, answering any questions the participant might have about the functions or commands. (Because these are system-specific, a general test plan cannot specify them. They should appear as an appendix in any specific test plan.)

7. The participant dictates the two short sentences, with guidance from the test monitor, to receive practice dictating and correcting errors.

8. The participant dictates the first text in the assigned text set (the one with embedded commands).

9. The participant completes the ASQ for that task.

10. The participant dictates the second text in the assigned text set (the plain text).

11. The participant completes the ASQ for that task.

12. The participant completes the AQ for that dictation session.

13. After completing all assigned dictation conditions with a single dictation system, the participant completes the PSSUQ for that system.

During a dictation session, a test monitor will keep track of all extra commands uttered for an accurate assessment of OPM. The test monitor will keep track of both primary and secondary accuracy by marking on a score sheet those words that the system misrecognized, but provided in a selection box and those that the system misrecognized

and failed to provide in a selection box. The score sheet is identical to the sheet the test participant reads from, but is printed double space to leave room for scoring marks. The test monitor will ensure that participants correct all recognition errors using the most efficient correction strategy. For dictation systems that save all speech information in a session (such as the IBM recognizer), the monitor will prompt users to correct recognition errors at the conclusion of each paragraph if it appears that there are errors that the test participant did not notice. For dictation systems that lose speech information after dictation of a set number of words, the monitor will prompt users to correct any recognition errors as they occur if it appears that the test participant did not notice the error. If errors remain after dictation, the test monitor will instruct the participant to correct the remaining errors, using voice if possible and the keyboard otherwise. The time for the session used to calculate WPM and OPM will include both the dictation and any additional correction time. This procedure allows measurement of primary and secondary accuracy and ensures that measurements of throughput are based on input that is ultimately error-free.

Discussion

By planning in advance for studies that have multiple (up to four) dictation conditions, it is possible to collect data from studies that have fewer than four dictation conditions and still have data that will be comparable from one study to another. The trick is to design all studies so the four different text sets make equal contributions to the measurements taken. The designs are flexible in that dictation conditions can be many different types of things, such as different dictation systems (for example, System 1, System 2, System 3, and System 4), the same dictation system with different enrollment methods (such as enrolled versus an unenrolled system), or the same dictation system at different points in long-term training of the system (such as baseline, after 4 hours of input, after 8 hours of input, and after 12 hours of input). The structure of these experiments controls a number of uninteresting but potentially misleading effects with extremely compact and efficient designs. Examination of historical data indicates that using the minimum sample size required to complete these designs should provide data with adequate precision to discriminate among competitive speech dictation systems (and conditions). These designs should be valuable to product developers who want to obtain benchmark measurements of accuracy and throughput for competitive speech dictation systems.

References

Lewis, J. R. (1993). Pairs of Latin squares that produce digram-balanced Greco-Latin designs: A BASIC program. *Behavior Research Methods, Instruments, & Computers, 25*, 414-415.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors, 36*, 368-378.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*, 57-78.

Trademarks

Appendix A. The Satisfaction Questionnaires

The After-Scenario Questionnaire

Participant: _____

Scenario: _____

For each of the statements below, circle the rating of your choice.

1. Overall, I am satisfied with the ease of completing the tasks in this scenario.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

3. Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing the tasks.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

The Accuracy Questionnaire (AQ)

Participant: _____

Scenario: _____

For each of the statements below, circle the rating of your choice.

1. This system's accuracy is acceptable.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

2. It is easy to correct errors in this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

3. The time it takes for the dictation program to recognize what I say is acceptable.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

4. In your estimation, what percentage of the time did the system correctly recognize your speech?

_____ %

5. Would you use this speech dictation system in your daily work?

Yes No

Comments:

The Post-Study System Usability Questionnaire: Dictation Scenarios

Participant: _____

System: _____

This questionnaire, which starts on the following page, gives you an opportunity to tell us your reactions to the system you used. Your responses will help us understand what aspects of the system you are particularly concerned about and the aspects that satisfy you.

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions.

Please read each statement and indicate how strongly you agree or disagree with the statement by circling a number on the scale. If a statement does not apply to you, circle N/A.

Please write comments to elaborate on your answers.

After you have completed this questionnaire, I'll go over your answers with you to make sure I understand all of your responses.

Thank you!

1. Overall, I am satisfied with how easy it is to use this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

2. It was simple to use this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

3. I could effectively complete the tasks and scenarios using this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

4. I was able to complete the tasks and scenarios quickly using this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

5. I was able to efficiently complete the tasks and scenarios using this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

6. I felt comfortable using this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

7. It was easy to learn to use this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

8. I believe I could become productive quickly using this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

9. The system gave error messages that clearly told me how to fix problems.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

10. Whenever I made a mistake using the system, I could recover easily and quickly.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

11. The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

12. It was easy to find the information I needed.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

13. The information provided for the system was easy to understand.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

14. The information was effective in helping me complete the tasks and scenarios.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

15. The organization of information on the system screens was clear.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

Note: *The interface includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, the screens (including their use of graphics and language).*

16. The interface of this system was pleasant.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

17. I liked using the interface of this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

18. This system has all the functions and capabilities I expect it to have.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

19. Overall, I am satisfied with this system.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

20. I would buy and use this system software.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

21. I would recommend this system software to others.

strongly agree						strongly disagree	
<==========						==========>	
1	2	3	4	5	6	7	N/A

Comments:

22. Please list the three things you liked most about this system software.

23. Please list the three things you liked least about this system software.

Appendix B. The Text Sets

This appendix contains the text sets for the dictation studies. The command text appears first in each set. This is the text that should contain embedded commands when presented to the test participants. Because the specific commands differ among dictation systems, no commands appear in this appendix. The second text in each text set is the plain text sample, which participants will see as is (without embedded commands). The number of words and approximate number of commands for each text sample are:

Text Set	Text Sample	Number of Words	Approximate Number of Commands
A	Command	261	60
	Plain	274	70
B	Command	353	75
	Plain	277	80
C	Command	288	50
	Plain	309	70
D	Command	283	60
	Plain	291	60

Text Set A

Command Text

May 3, 1994
9:00 a.m.

Dear Mr. Bartholomew:

The staff at the Skeptical Inquirer certainly enjoyed your article, Culture-Bound Syndromes as Fakery. We would like a series of follow-ups for a supplement that we are planning. Have you prepared any additional articles on this topic? We will offer $750 for each two page report. The portions that we feel most strongly require additional explanation follow:

"For the past one hundred years anthropologists and psychiatrists have debated the origin and nature of a curious behavior confined almost exclusively to the Southeast Asian neighboring cultures of Malaysia and Indonesia. Upon being startled, ordinarily timid, exceedingly polite women sometimes respond with vulgarities, obscenities, and outrageous gestures. In severe cases, the women experience automatic obedience, doing whatever they are told. Afterward they claim amnesia and are not held responsible for their actions. Episodes of this type last from a few minutes to several hours."

"Anthropologists have an unfortunate tendency to emphasize and glorify the exotic, especially in someone else's backyard, while psychiatrists are often overly eager to place a convenient disorder label on deviant behavior, no matter where it is found. When a community experiences a spate of flying saucer sightings, it is typically labeled as a form of epidemic hysteria, yet this behavior is not contagious and participants are not clinically hysterical."

If the staff's response to your article is any indication how our readers will respond, I think we will have a MAJOR SERIES OF SUCCESSES! Please call me at 375-5880 by May 15 and let me know your decision.

Sincerely,

Paul Kurtz

Plain Text

April 10, 1995
11:30 a.m.

Dear Mr. Gleick:

Congratulations! We have now sold 100,000 copies of your book, CHAOS: Making a New Science. Enclosed you will find your bonus check for $10,000 as we agreed when negotiating your contract. Personally, I want to let you know how much I have enjoyed your book, especially the following section:

"He decided to look more closely at the way two nearly identical runs of weather flowed apart. He copied one of the wavy lines of output onto a transparency and laid it over the other, to inspect the way it diverged. First, two humps matched detail for detail. Then one line began to lag a hairsbreadth behind. By the time the two runs reached the next hump, they were distinctly out of phase. By the third or fourth hump, all similarity had vanished."

"It was only a wobble from a clumsy computer. He could have assumed something was wrong with his particular machine or his particular model -- probably should have assumed. It was not as if he had mixed sodium and chlorine and got gold. Although his model was a gross parody of the earth's weather, he had a faith that they captured the essence of the real atmosphere. That first day, he decided that long-range weather forecasting must be doomed."

So, what do you plan to do with the bonus? I think you should take a well-deserved vacation, but the money's yours to do with as you please. And don't forget that if you reach a million copies sold by December 31, 1997, you'll receive a $50,000 bonus.

Respectively yours,

Nelson Merrick
Office phone: 443-1066

Text Set B

Command Text

June 21, 1997
4:00 p.m.

Dear Mrs. Ford:

I recently acquired some copies of an old book entitled Typing Behavior, by August Dvorak, published in 1936. I have in my possession over 500 copies of this book, which I would be glad to sell to you for $5000. If you want to pursue this matter, you MUST CONTACT ME NO LATER THAN THE END OF BUSINESS NEXT WEEK! If I don't hear from you by June 28 I will look for another buyer. By the way, I thought you might find the following observations interesting. This book's a classic.

"What is the oldest, greatest invention from the past? In considering, you may suddenly exclaim, I have it -- it really is language! During primitive centuries, while men were inventing certain sounds to help control one another's actions, they stumbled upon devices which would carry such human speech further than the unaided human voice. When people were still savages, a few crude pictures served that purpose. Some ridiculous drawing -- perhaps a crude map that designated a meeting place -- was the only love letter a primitive girl might expect from her boy friend."

"Outside the schools, typewriters have been widespread for over a generation. Yet as a student typist you now make your first formal contact with a writing machine! In all probability as a youngster you, too, once took a good position and practiced easy, sideways writing movements across a sheet of paper. Rhythm, as always, helps your speed. However, if you desire finer and faster writing, you can use a typewriter. Business men who require high quality or quantity writing buy machines to produce it. Why else did men invent -- amid other possibilities -- the typewriter? The words that you have so painfully organized are swiftly and precisely organized for you by this machine."

Don't you agree that this is interesting material? Although almost no one uses the Dvorak keyboard, the book might be of historical interest to your customers. If you need to contact me about this, please call me at 495-9033 during business hours.

Respectively yours,

Frank Gilbreth

Plain Text

August 30, 1995
10:00 a.m.

Dear Mr. Gardner:

Your recent article, Artificial Languages, was VERY ENTERTAINING! I have often wondered about the origins of Esperanto and other artificial languages. I especially enjoyed the following paragraphs:

"In the seventeenth century, among a variety of philosophers, the notion arose that perhaps a completely artificial language, based on logic, with simplified grammar and spelling, might serve to unify nations. This grandiose dream quickly gripped the minds of hundreds of linguistic cranks, who during the next three centuries proposed more than three hundred artificial tongues."

"Of these, by far the most successful has been Esperanto, the brainchild of a Warsaw eye doctor who used the name Dr. Esperanto. The word Esperanto means one who hopes. This expressed his desire that Esperanto would become the world's second language. More than 30,000 books have been translated into Esperanto. Reader's Digest published an Esperanto edition."

"There have been numerous attempts to improve Esperanto. Ido -- which means offspring -- was developed in 1907. Its developer regarded all Esperantists as depraved. Bertrand Russell recalled hearing this linguist complain that Ido had no word similar to Esperantist, which refers to a speaker of Esperanto. Russell suggested idiot, but the linguist was not pleased."

Do you plan to write any more articles about this subject? If you are interested in working together on a manuscript, please call me in my office between 2:00 p.m. and 4:00 p.m. any day after Tuesday, September 15 so we can plan our writing strategy. My office telephone is 982-4061. Also, could you recommend any additional sources? I am willing to spend up to $100 for preparation materials.

Sincerely yours,

Jerry Hobbs

Text Set C

Command Text

September 13, 1996
10:30 a.m.

Dear Mr. Nickell:

I read your latest article "Alien Autopsy Hoax" in the Skeptical Inquirer and found it to be MOST INFORMATIVE! Do you have a set fee for lecture engagements? I am in the process of hiring speakers for a series of seminars and hope $3000 is a great enough incentive to have you speak for an hour to a group of "reality buffs" from Roswell, New Mexico. The following paragraphs are what piqued my interest.

"None of us were of the opinion that we were watching a real alien autopsy, or an autopsy on a mutated human which has also been suggested. We all agreed that what we were seeing was a very good fake body, a large proportion of which had been based on a lifecast. Although the nature of the film obscured many of the things we had hoped to see, we felt that the general posture and weighting of the corpse was incorrect for a body in a prone position and had more in common with a cast that had been taken in an upright position."

"The Roswell myth should be permitted to die a deserved death. Whether or not we are alone in the universe will have to be decided on the basis of better evidence than that provided by the latest bit of Roswell fakery. Television executives have a responsibility not to confuse programs designed for entertainment with news documentaries."

If speaking at our seminar is a possibility please call me at 596-8760 during business hours. I can assure you an audience of about two hundred people, and you will find this group's enthusiasm will make this engagement worthwhile.

Looking forward to your reply,

Eugene Emory

Plain Text

April 7, 1995
3:15 p.m.

Dear Dr. Loftus:

I have read your article, Remembering Dangerously - The Repressed Memory Debate, and have some concerns about the article's claims. I am a professional therapist contracted to edit this paper, and I would like to spend SOME SIGNIFICANT TIME discussing the following sections!

"We live in a strange and precarious time that resembles at its heart the hysteria and superstitious fervor of the witch trials of the sixteenth and seventeenth centuries. Men and women are being accused, tried, and convicted with no proof or evidence of guilt other than the word of the accuser. Individuals are being imprisoned on the evidence provided by memories that come back in dreams and flashbacks -- memories that did not exist until a person wandered into therapy."

"On her own initiative, Willa hired a private investigator to pose as a patient and seek therapy from her sister's therapist. The private investigator called herself Ruth. She twice visited the therapist and secretly tape-recorded both of the sessions. For long sections of the tape it was hard to tell who was the patient and who was the therapist."

"Why at this time in our society is there such an interest in repression and the uncovering of repressed memories? Why is it that almost everyone you talk to either knows someone with a repressed memory or knows someone who's being accused, or is just plain interested in the issue? Why do so many individuals believe these stories, even the more bizarre, outlandish, and outrageous ones?"

Is this really what you meant to say? Please call me in my office between 9:00 a.m. and 11:00 a.m. on Thursday, April 14 so we can go over some issues. My office telephone is 526-3692. For your information, I am charging the publisher $60 per hour for my services.

Respectfully yours,

Cheryl Norton

Text Set D

Command Text

November 17, 1998
8:15 a.m.

Dear Mrs. Hoffman:

I just finished reading your book, The Challenge of the Unknown, and had a few questions about some of your book's assertions. I am a practicing mathematician hired to review this manuscript, and am NOT SURE that I agree with the following text!

"You can imagine wanting to know an answer, especially in the old days, when you wrote your program on cards and then submitted them to the computer center. They'd run them overnight and get back to you the next day. And you'd have an account with, say, a hundred bucks in it. Every once in a while, the program would have an infinite loop and burn up gobs of money. You'd get nothing out of the program, since it was stuck in an infinite loop. Either your account would run out of money or somehow the machine would notice that it had been going for a very long time and shut itself off."

"Besides the proof of the impossibility of solving the halting problem, the year 1936 witnessed another assault on the illusory goal of absolute mathematical knowledge. Alonzo Church proved that the so-called decision problem was unsolvable: there can never be a general procedure for deciding whether a given statement expresses an arithmetic truth. In other words, no computer will ever exist that can spew out the truths of mathematics."

Have I misunderstood your arguments? Please call me at home between 8:00 p.m. and 10:00 p.m. on Wednesday, November 20 so we can discuss this. My home number is 360-9616. Also, do you know if the publisher still plans to sell the book for $37.50 per copy?

Thank you,

Cheryl Norton

Plain Text

February 8, 1996
1:20 p.m.

Dear Mr. Pinker:

It is our pleasure to let you know that we've decided to publish your new book, The Language Instinct. We expect the list price to be $25.75. We were very impressed when Noam Chomsky declared that it was "an extremely valuable book." There is one of the book's sections that WE MUST DISCUSS when we get together to sign the contract next Monday, February 12 at 10:00 a.m. - NO MATTER WHAT ELSE WE DO! Here's a quotation from that section:

"But it is wrong, all wrong. The idea that thought is the same thing as language is an example of what can be called a conventional absurdity: a statement that goes against all common sense but that everyone believes because they dimly recall having heard it somewhere and because it is so pregnant with implications. Think about it. We have all had the experience of uttering or writing a sentence, then stopping and realizing that it isn't exactly what we meant to say. To have that feeling, there has to be a what we meant to say that is different from what we said. Sometimes it is not easy to find any words that properly convey a thought. When we hear or read, we usually remember the gist, not the exact words, so there has to be such a thing as a gist that is not the same as a bunch of words."

Do you see why my boss wants us to work on this a little bit? I don't think it will be a problem, though, and I look forward to seeing you next Monday. If you have any questions about this, please call me at 437-0635.

Sincerely yours,

William Morrow

Appendix C. Background Questionnaire

Please circle the best response and fill in the blanks when appropriate.

1. I am:

    a. male b. female

2. My age group is:

    a. less than 20 years old b. 20-29 years old
    c. 30-39 years old d. 40-49 years old
    e. 50-59 years old f. over 59 years old

3. My educational level is:

    a. high school graduate b. vocational/technical graduate
    c. some college d. bachelor's degree
    e. advanced degree f. other (please describe):
    _____

4. I have previous experience using non-computerized dictation equipment.

    a. No b. Yes
    If yes, what equipment and how long?
    _____

5. I have previous experience using computerized dictation systems.

    a. No b. Yes
    If yes, what equipment and how long?
    _____

6. I have used computers for:

    a. more than 5 years b. 1-5 years
    c. less than 1 year d. I have never used a computer.

7. I have used computers at:

    a. home and work b. work only
    c. home only d. I have never used a computer.

    If applicable, describe your work computer:
    _____
    If applicable, describe your home computer:
    _____

8. Briefly, my work experience (type of work/years experience) has been:

_____
_____

9. When I use a computer, I typically use it for:

Circle all that apply

a. word processing
b. spreadsheets
c. graphics/paint/draw
d. video games
e. other (please describe):

_____

10. My typical typing speed is about _____ words per minute.

11. When I was a child (birth to 10 years old), I lived in the following locations:

List up to three locations/years there. If more than three, list those where you lived the longest.

_____ _____ _____

Appendix D. The Two Practice Sentences

Now is the time for all good persons, NOT just men, to come to the aid of the League of Concerned Citizens.

What did Jessie Valdez say when she called 976-3290 on June 27, 1992 at 3:30 p.m.?