Streamlining a General Test Plan for Competitive Evaluation of Dictation Accuracy and Throughput

TR 29.3158

James R. Lewis
Speech Product Design and Usability
West Palm Beach, FL

Abstract

I examined data from three studies of dictation accuracy and throughput executed using two test texts per experimental condition to determine the effect of reducing the number of test texts per condition to one on (1) estimates of the magnitudes of accuracy and throughput and (2) the precision of these estimates. Switching to one test text per condition has a negligible effect on both estimates of magnitudes and the precision of those estimates. Use of a single test text would substantially reduce the amount of time required to conduct these studies without significant degradation in the quality of the measurements.

ITIRC Keywords

Dictation accuracy
Dictation throughput
Experimental design

About the Author

James R. Lewis has worked as a human factors engineer for IBM since 1981, providing human factors support for both hardware and software development. He received his MA in experimental psychology (Engineering Psychology) in 1982 from New Mexico State University. He received his Ph.D. in experimental psychology (Psycholinguistics) from Florida Atlantic University in 1996. He has published over 100 technical papers, and is currently at the 6th level in the IBM Patent Program.

Contents

Introduction

Since early 1996 we have used a general test plan to evaluate competitive dictation accuracy and throughput (Lewis, 1997). Although that version of the test plan featured considerable streamlining relative to the previous procedure, it still required two test texts per experimental condition for each participant. The first test text for each pair contained embedded commands so participants did not have to memorize the recognizer's command sets to successfully dictate. The second text was plain, with no embedded commands. The final measurements for each condition were the average measurements for the pair of test texts. A large proportion of the time required to conduct these studies is due to the time spent dictating the two test texts. Given the availability of data gathered over the last 1.5 years, it is reasonable to examine the data to determine if it is possible to streamline this test procedure to an even greater extent by reducing the number of test texts used per experimental condition. These types of studies generate a large number of measurements, but the key measurements are primary accuracy (the percentage with which the recognizer correctly identifies a user's utterance) and throughput (the words per minute, or WPM, achieved by the participant, including both the time required to speak the text and the time required to correct misrecognitions).

The purpose of this report is to examine data from three studies executed using the general test plan to determine the effect of reducing the number of test texts per condition from two to one on (1) estimates of the magnitudes of primary accuracy and throughput and (2) the precision of these estimates. If this reduction would have adversely affected the magnitudes or precision of measurements from these studies, then we should continue to use the current general test plan. Otherwise, we should reduce the number of test texts in future dictation accuracy and throughput studies.

Method

I developed tables and charts to compare magnitudes and precision of measurement for both primary accuracy and throughput from three studies of IBM VoiceType Dictation (VTD). The first study (8 participants) compared VTD 1.32 with full enrollment, an early version of VTD 3.0 with no enrollment, 50-sentence enrollment, and full enrollment. The second study (12 participants) was a follow-up of the first, using a later (but still pre-release) version of VTD 3.0, providing measurements for no enrollment, 50-sentence enrollment, and full enrollment. The third study (8 participants) was a comparison of a newer version of VTD and IBM's new continuous speech dictation product. In this study, participants used VTD both without enrollment and with 50-sentence enrollment, and the continuous speech product without enrollment and with full enrollment.

Results

Study 1

Primary Accuracy. Table 1 and Figure 1 show the estimated means of primary accuracy for Study 1, with one line showing the estimated means based on averaging both texts and one using only the first test text. The grand means (averaged across experimental conditions) for the two sets of estimates differ by less than 1%. Table 2 and Figure 2 illustrate the effect of reducing the number of test texts on the size of the bounds for a 90% confidence interval on the means. The grand means for this analysis differ by only 0.1%.

Table 1. Primary Accuracy for Study 1: Mean Difference Between Both Texts' Average and Command Text Only

VTD 1.32 Full Enrollment	VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
-1.1	-0.4	-0.7	-0.3	-0.6

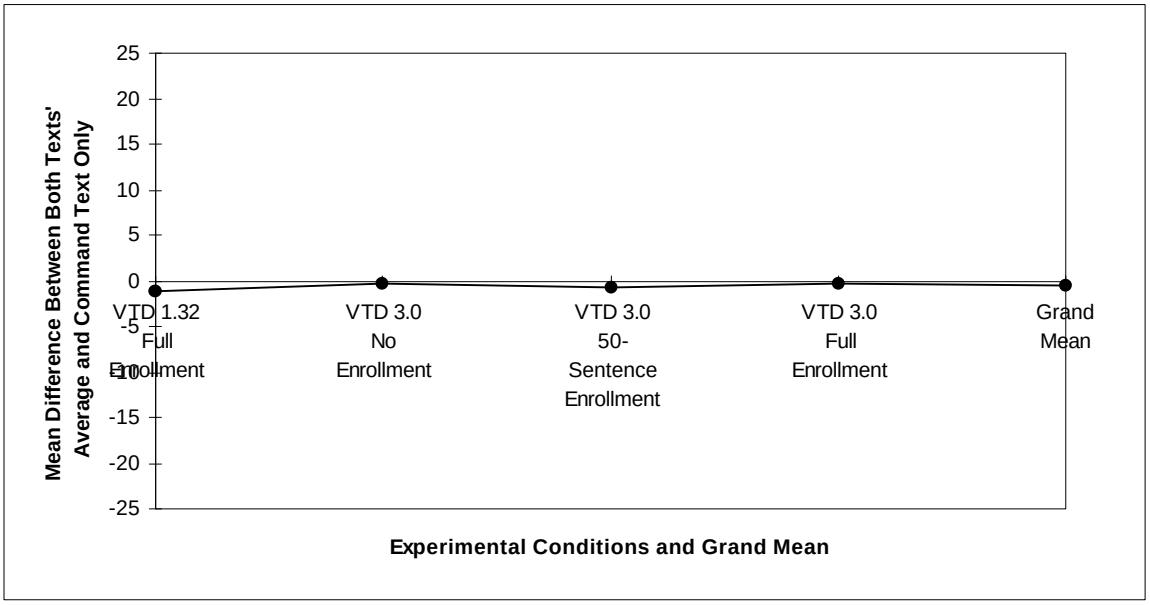

Figure 1. Primary Accuracy for Study 1: Profile of Mean Differences

Table 2. Primary Accuracy for Study 1: 90% Confidence Interval Bounds

	VTD 1.32 Full Enrollment	VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
Both texts	2.7	2.6	1.2	1.2	1.9
Commands only	2.7	3.1	1.1	1.2	2.0
Difference	0.0	-0.5	0.1	0.0	-0.1

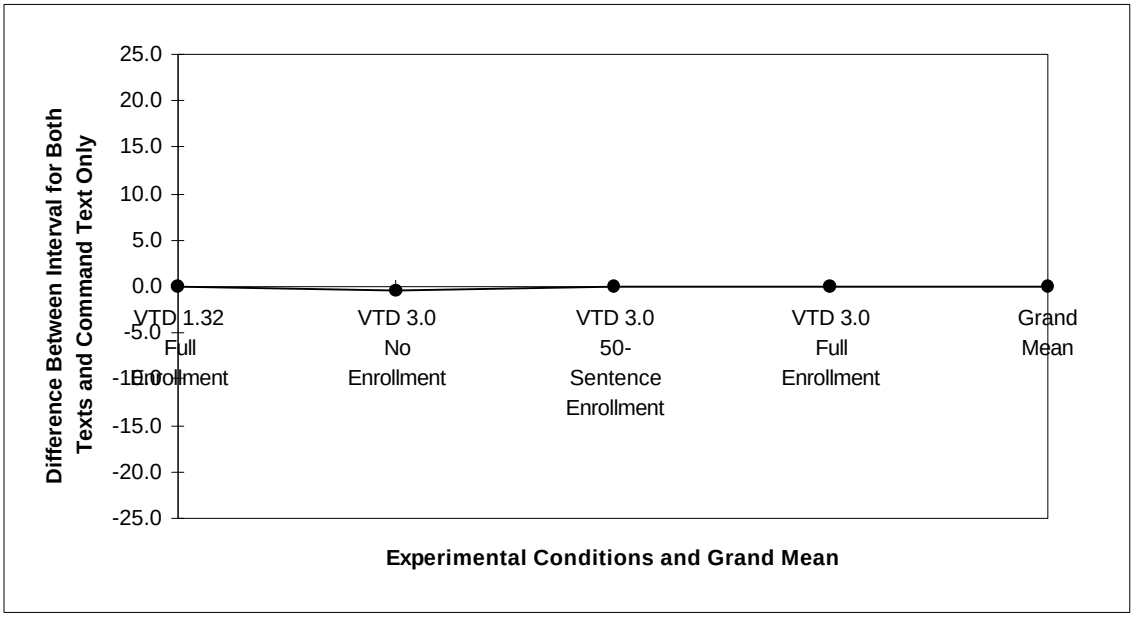

Figure 2. Primary Accuracy for Study 1: Profile of 90% Confidence Interval Bound Differences

Throughput. Table 3 and Figure 3 show the estimated means of throughput for Study 1, and Table 4 and Figure 4 illustrate the effect of reducing the number of test texts on the size of the bounds for a 90% confidence interval on the throughput means. For estimated means, the difference between the grand means was only 0.1 WPM, and for 90% confidence interval bounds, the difference between grand means was only 0.3 WPM.

Table 3. Throughput for Study 1: Mean Difference Between Both Texts' Average and Command Text Only

VTD 1.32 Full Enrollment	VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
0.1	-0.2	-0.1	-0.2	-0.1

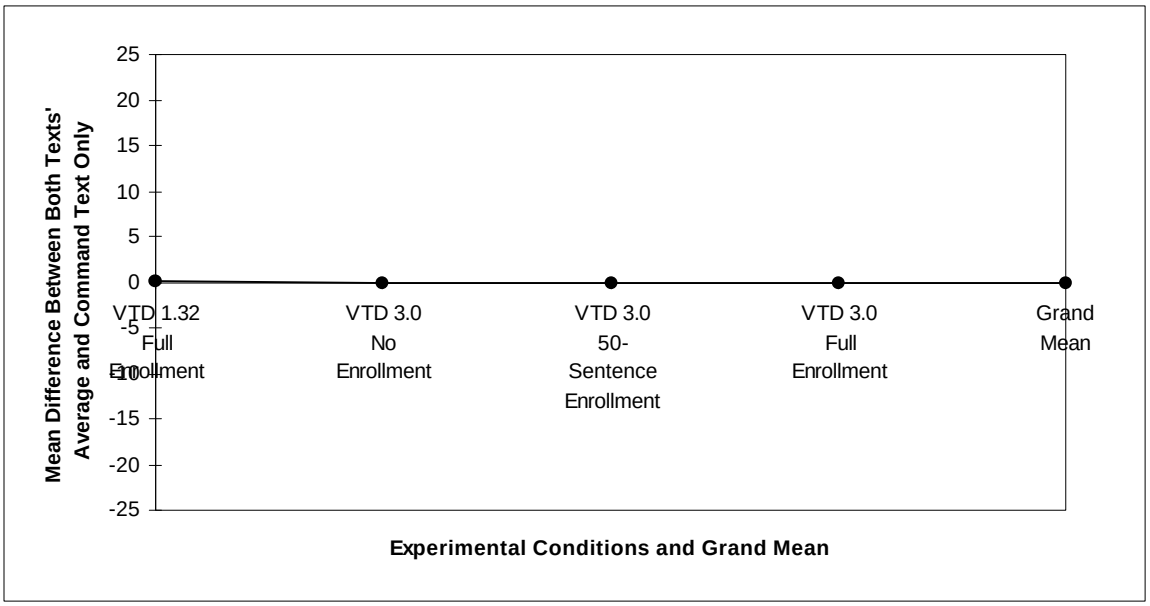

Figure 3. Throughput for Study 1: Profile of Mean Differences

Table 4. Throughput for Study 1: 90% Confidence Interval Bounds

90% CI Deltas for Mean Throughput	VTD 1.32 Full Enrollment	VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
Both texts	4.4	3.4	3.4	3.4	3.7
Commands only	4.9	4.6	3.4	3.1	4.0
Difference	-0.5	-1.2	0.0	0.3	-0.3

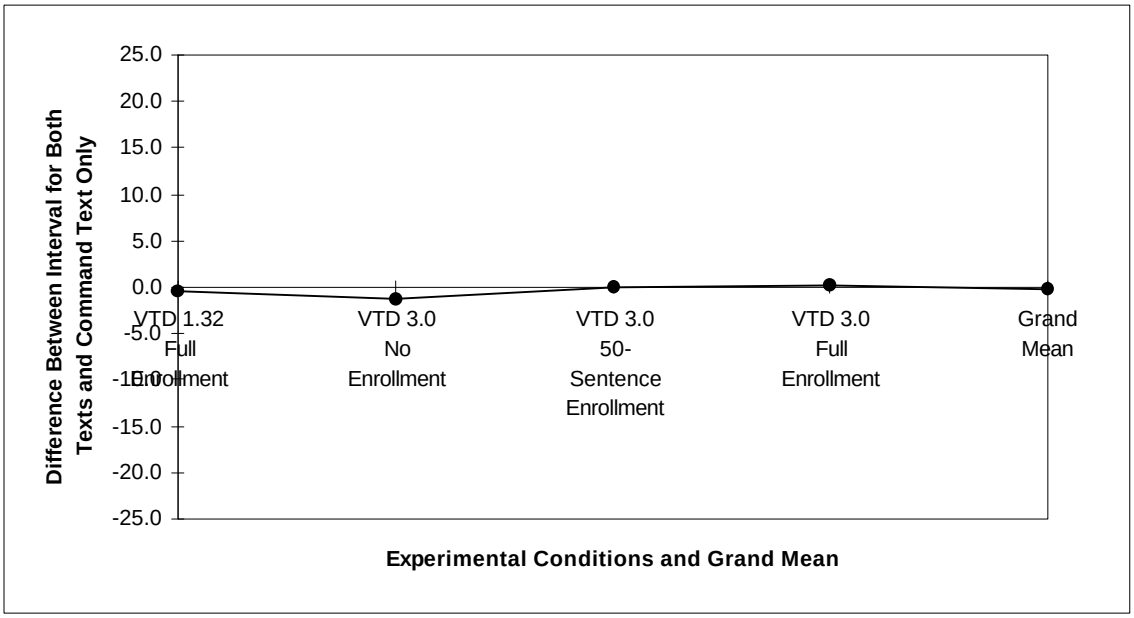

Figure 4. Throughput for Study 1: Profile of 90% Confidence Interval Bound Differences

Study 2

Primary Accuracy. Table 5 and Figure 5 show the estimated means of primary accuracy for Study 2, with one line showing the estimated means based on averaging both texts and one using only the first test text. The grand means (averaged across experimental conditions) for the two sets of estimates differ by less than 1%. Table 6 and Figure 6 illustrate the effect of reducing the number of test texts on the size of the bounds for a 90% confidence interval on the means. The grand means for this analysis differ by only 0.2%.

Table 5. Primary Accuracy for Study 2: Mean Difference Between Both Texts' Average and Command Text Only

VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
-0.8	-0.6	-0.4	-0.6

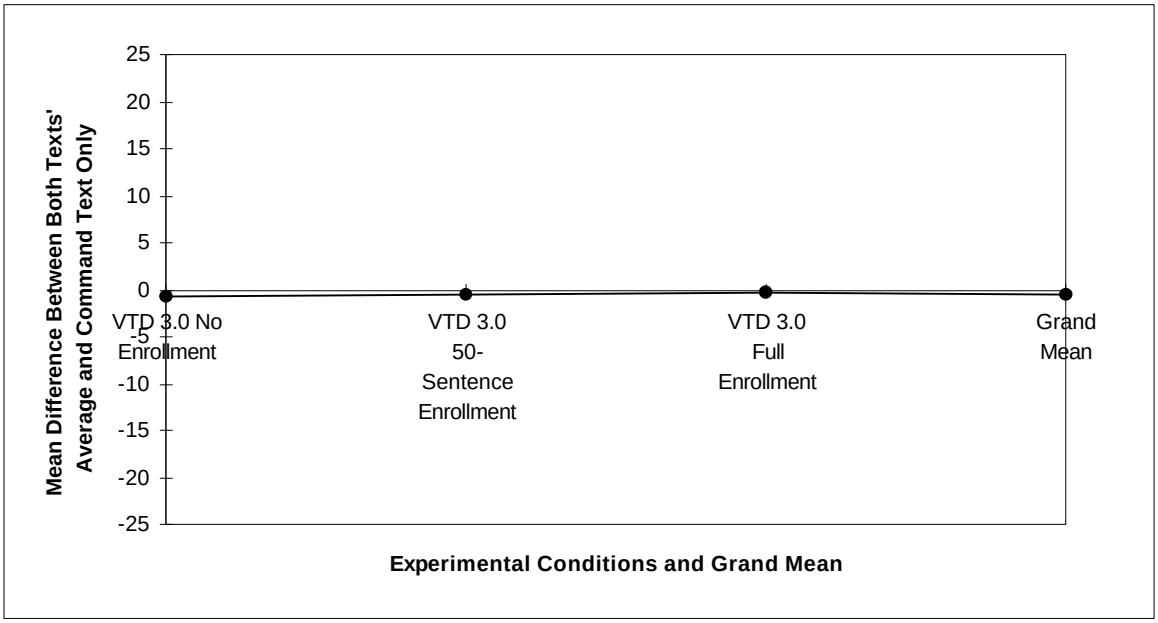

Figure 5. Primary Accuracy for Study 5: Profile of Mean Differences

Table 6. Primary Accuracy for Study 2: 90% Confidence Interval Bounds

	VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
Both texts	3.1	1.8	1.2	2.1
Commands only	2.8	1.7	1.2	1.9
Difference	0.3	0.1	0.0	0.2

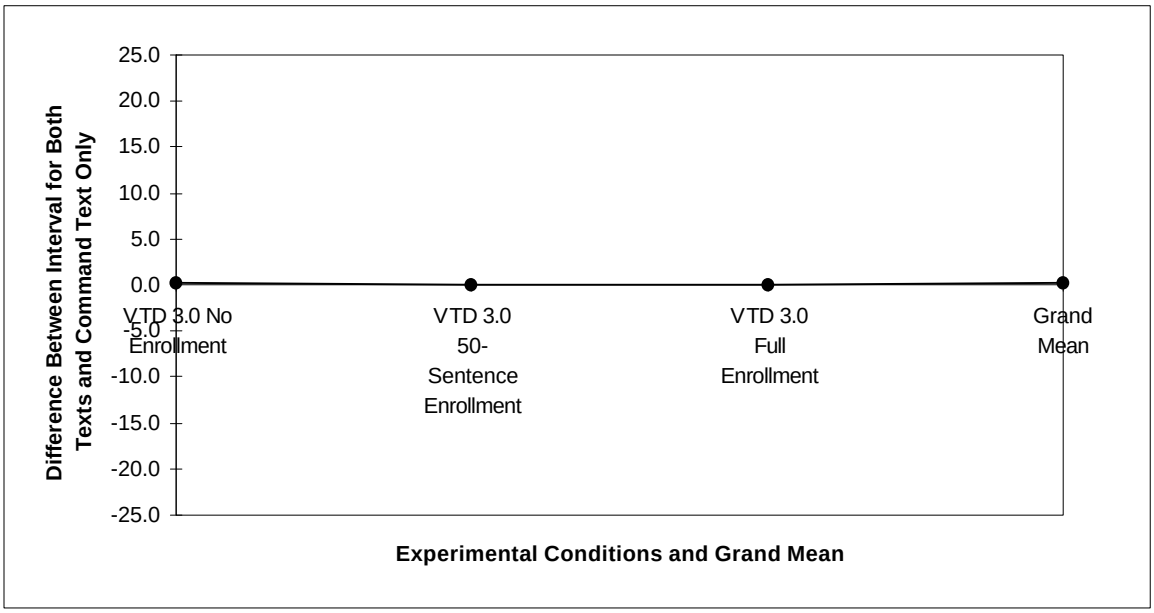

Figure 6. Primary Accuracy for Study 2: Profile of 90% Confidence Interval Bound Differences

Throughput. Table 7 and Figure 7 show the estimated means of throughput for Study 2, and Table 8 and Figure 8 illustrate the effect of reducing the number of test texts on the size of the bounds for a 90% confidence interval on the throughput means. For estimated means, the difference between the grand means was only 0.2 WPM, and for 90% confidence interval bounds, the difference between grand means was only 0.3 WPM.

Table 7. Throughput for Study 2: Mean Difference Between Both Texts' Average and Command Text Only

VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
-0.1	-0.4	-0.1	-0.2

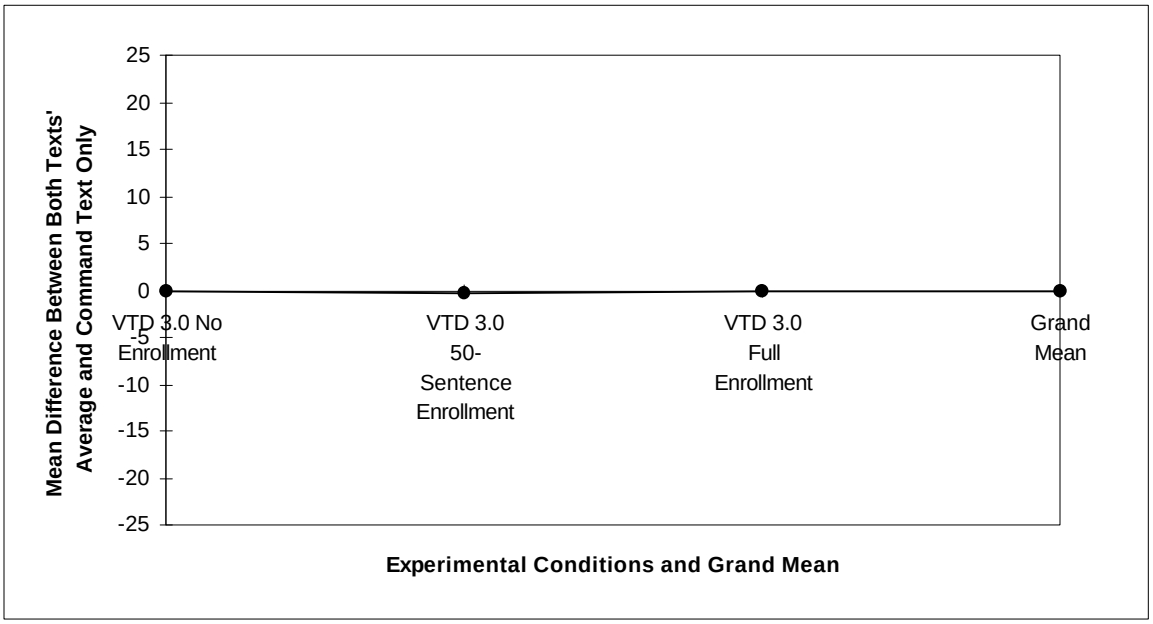

Figure 7. Throughput for Study 2: Profile of Mean Differences

Table 8. Throughput for Study 2: 90% Confidence Interval Bounds

	VTD 3.0 No Enrollment	VTD 3.0 50-Sentence Enrollment	VTD 3.0 Full Enrollment	Grand Mean
Both texts	3.1	4.3	3.0	3.5
Commands only	3.6	4.7	3.0	3.8
Difference	-0.5	-0.4	0.0	-0.3

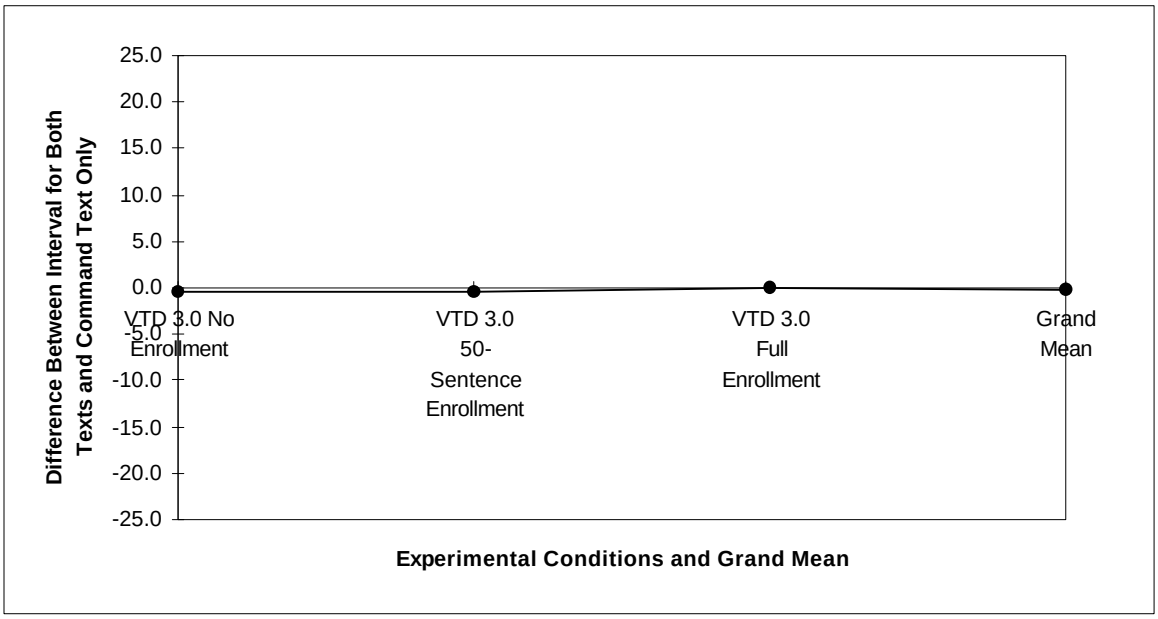

Figure 8. Throughput for Study 2: Profile of 90% Confidence Interval Bound Differences

Study 3

Primary Accuracy. Table 9 and Figure 9 show the estimated means of primary accuracy for Study 3, with one line showing the estimated means based on averaging both texts and one using only the first test text. The grand means (averaged across experimental conditions) for the two sets of estimates differ by only 1.2%. Table 10 and Figure 10 illustrate the effect of reducing the number of test texts on the size of the bounds for a 90% confidence interval on the means. The grand means for this analysis were identical.

Table 9. Primary Accuracy for Study 3: Mean Difference Between Both Texts' Average and Command Text Only

Discrete No Enrollment	Discrete 50-Sentence Enrollment	Continuous No Enrollment	Continuous Full Enrollment	Grand Mean
-0.5	-1.5	-1.5	-1.1	-1.2

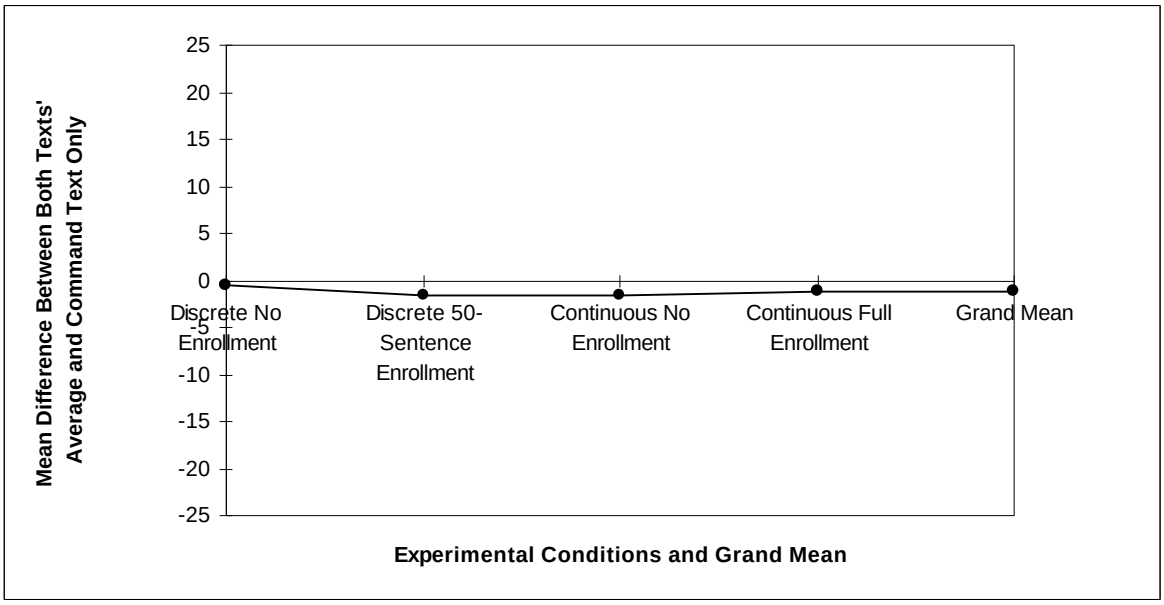

Figure 9. Primary Accuracy for Study 3: Profile of Mean Differences

Table 10. Primary Accuracy for Study 3: 90% Confidence Interval Bounds

	Discrete No Enrollment	Discrete 50-Sentence Enrollment	Continuous No Enrollment	Continuous Full Enrollment	Grand Mean
Both texts	5.6	4.0	3.2	2.4	3.8
Commands only	5.6	3.3	3.7	2.6	3.8
Difference	0.0	0.7	-0.5	-0.2	0.0

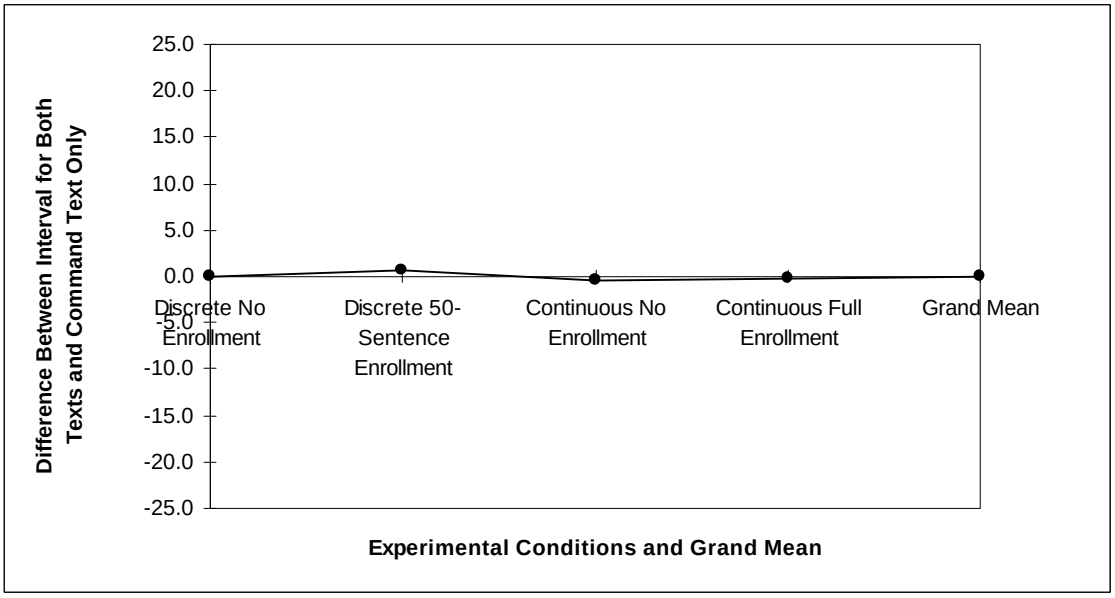

Figure 10. Primary Accuracy for Study 3: Profile of 90% Confidence Interval Bound Differences

Throughput. Table 11 and Figure 11 show the estimated means of throughput for Study 3, and Table 12 and Figure 12 illustrate the effect of reducing the number of test texts on the size of the bounds for a 90% confidence interval on the throughput means. For estimated means, the difference between the grand means was only 0.2 WPM, and for 90% confidence interval bounds, the difference between grand means was only 0.2 WPM.

Table 11. Throughput for Study 3: Mean Difference Between Both Texts' Average and Command Text Only

Discrete No Enrollment	Discrete 50-Sentence Enrollment	Continuous No Enrollment	Continuous Full Enrollment	Grand Mean
1.1	0.4	0.4	-1.1	0.2

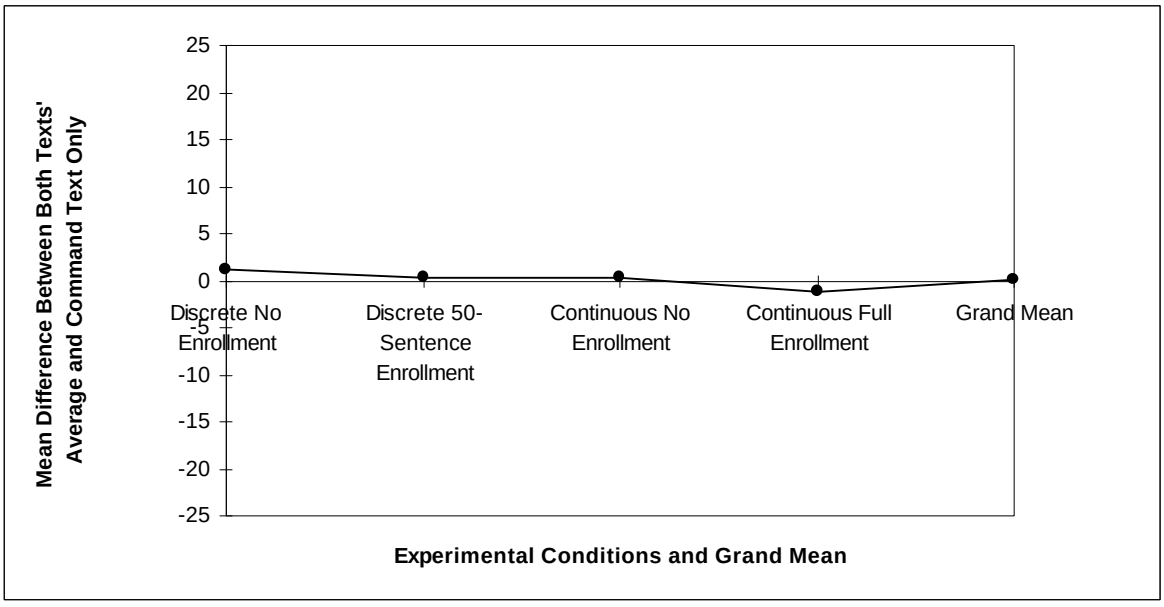

Figure 11. Throughput for Study 3: Profile of Mean Differences

Table 12. Throughput for Study 3: 90% Confidence Interval Bounds

	Discrete No Enrollment	Discrete 50-Sentence Enrollment	Continuous No Enrollment	Continuous Full Enrollment	Grand Mean
Both texts	7.2	8.8	5.1	6.6	6.9
Commands only	6.5	9.4	5.5	7.1	7.1
Difference	0.7	-0.6	-0.4	-0.5	-0.2

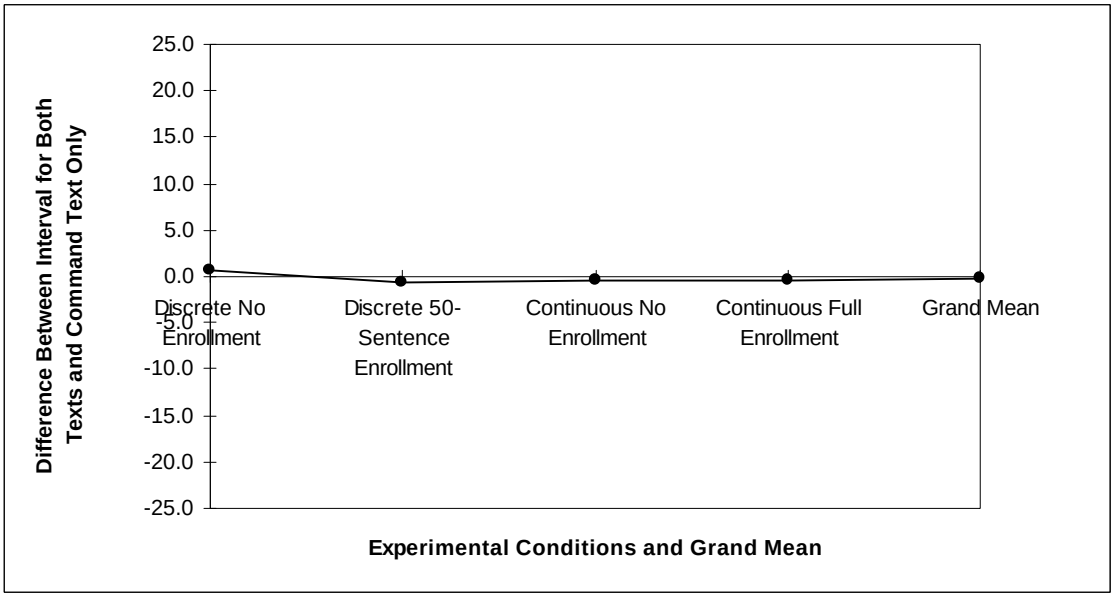

Figure 12. Throughput for Study 3: Profile of 90% Confidence Interval Bound Differences

Discussion

The results of these analyses, with highly consistent results across three independent studies, show that reducing the number of test texts from two per experimental condition to one per experimental condition has virtually no effect on either the magnitude of estimates of means of primary accuracy and throughput or the precision of those estimates. To take advantage of this finding and increase the efficiency of these types of evaluations, future studies of dictation accuracy and throughput should use only one test text (specifically, the text with embedded commands) per experimental condition. This will substantially reduce the amount of time required to conduct these studies without any significant degradation in the quality of the measurements.

With regard to the precision of measurement reported here, it is important to keep in mind that the studies we run are usually within subjects. Because within-subjects measurements are typically correlated and we generally focus on difference scores rather than raw scores, the precision of measurement associated with the difference scores will be greater (smaller confidence interval bounds) than the precision reported here for the standard accuracy scores (Steele and Torrie, 1960).

References

Lewis, J. R. (1997). *A general plan for conducting human factors studies of competitive speech dictation accuracy and throughput* (Tech. Report 29.2246). Raleigh, NC: International Business Machines Corp.

Steele, R. G. D., and Torrie, J. H. (1960). *Principles and procedures of statistics*. New York, NY: McGraw-Hill.