*Journal of Speech and Hearing Research, Volume 38,* 889–901, August 1995

# Discourse Comprehension of Synthetic Speech Across Three Augmentative and Alternative Communication (AAC) Output Methods

**D. Jeffery Higginbotham**
*Department of Communicative Disorders and Sciences*
*State University of New York at Buffalo*

**Christine A. Scally**
*Clark County School District*
*Las Vegas, NV*

**Debra C. Lundy**
*University of Illinois at Chicago*
*Assistive Technology Unit*
*Chicago, IL*

**Kim Kowarsky**
*Bloorview Children's Hospital*
*Willowdale, Ontario, Canada*

The purpose of this investigation was to determine the relative effects of three different Augmentative and Alternative Communication (AAC) speech output methods (word, sentence, mixed words and letters) on a listener's ability to summarize paragraph-length texts. Based on previous work on the discourse processing of synthetic speech, a monotonic trend in a listener's ability to accurately summarize texts produced by different output methods was hypothesized (word > sentence > mixed). Thirty able-bodied adults were individually tested over a 2-day period, listening to four texts produced by a DECtalk speech synthesizer at a speech presentation rate of 7.5 wpm. Using a discourse summarization taxonomy developed by Higginbotham, Lundy, and Scally (1993), the experimental hypothesis was confirmed. Word-method listeners produced significantly more accurate renditions of the original texts than sentence-method listeners, who, in turn, did better than mixed-method listeners. Statistically significant differences also were found for the topic familiarity of the texts. The role of the above variables on AAC device comprehension and technology design is discussed.

**KEY WORDS: augmentative and alternative communication (AAC), speech synthesis, voice output communication aids (VOCA), discourse comprehension**

In the everyday transaction of spoken communication, discourse comprehension is typically regarded as an "effortless process" in which listeners encode and parse the discourse into constituent elements, develop the gist of the text, and integrate this information with other knowledge sources (Byrd, 1985; Kintsch, 1988). Due to the redundant characteristics of the linguistic information and the listener's ability to use a variety of linguistic and contextual resources to produce a reasonable interpretation of the speaker's communication, misunderstandings between natural speakers occur infrequently (Harris, 1987). When such mistakes do happen, interactants possess a wealth of repair mechanisms such as self-correction and contingent query to resolve these problems efficiently (Harris, 1987; Hirst, McRoy, Heeman, Edmonds, & Horton; 1993; Sacks, Schegloff, & Jefferson, 1974).

Unlike the relative "effortlessness" associated with understanding natural speech, comprehension of synthesized speech message output from Augmentative and Alternative Communication (AAC) can be more problematic, resulting in significant misunderstandings and frequent attempts at message repair (Higginbotham, Drazek, Kowarsky, Scally, & Segal, 1994; Raghavendra & Allen, 1993). Two general characteristics that largely contribute to these problem are the quality of the synthetic

0022-4685/95/3804-0889

speech signal and the message production methods associated with the communication technology.

## Quality of the Synthetic Speech Signal

In contrast to natural speech, synthetic speech contains less acoustic redundancy, less suprasegmental information, and at times, misleading phonetic cues (Duffy & Pisoni, 1992). Because of these acoustic deficiencies, listeners take more time to perceptually process the synthetic speech signal, and have more difficulty correctly identifying words produced by speech synthesis versus natural speech (Logan, Greene, & Pisoni, 1989; Mirenda & Beukelman, 1987, 1990; Ralston, Pisoni, Lively, Greene, & Mullennix, 1991). Listeners also display more difficulty transcribing sentences, identifying propositional content, and making correct inferences for sentence and discourse materials produced by speech synthesis than for productions of natural speech (Higginbotham et al., 1994; Luce, Feustel, and Pisoni, 1983; Raghavendra & Allen, 1993; Ralston, Pisoni, Lively, Greene, & Mullennix, 1991).

In a review of recent research on the comprehension of synthetic speech, Duffy et al. (1992) concluded that increased demands for perceptual level processing of the synthetic speech signal limit the availability of resources for higher level comprehension processing (e.g., syntactic processing, inferencing, instantiation maintenance of complex mental representations). Because of this competition for limited processing resources, comprehension errors may result, particularly for lower quality synthetic speech, for noisy listening environments, and for complex text materials (Higginbotham et al., 1994; Pisoni, Manous, & Dedina, 1987; Raghavendra & Allen, 1993; Ralston et al., 1991; Talbot, 1987). Because of these processing difficulties, listeners pay more attention to the surface structure and rely more on topical knowledge and the semantic context for synthesized speech compared to natural speech in sentence and discourse contexts.

## Synthetic Speech Output and AAC Devices

Synthetic speech output from an AAC device is also affected by the rate at which a user can produce text, as well as the design characteristics of the communication technology. The communication rates for individuals using augmentative communication are extremely slow (Foulds, 1980; Higginbotham, 1992; Mathy-Laikko, West, & Jones; 1993; Vanderheiden, 1988). Estimates range between one-half and 20 words a minute, depending on the input rate of the communicator and the selection technique (scanning, direct selection, encoding) and the vocabulary stored in the communication. In a review of the existing empirical studies in this area, Mathy Laikko et al. (1993) found the average output rates of device users to range between 2.3 and 8.2 words per minute.

Higginbotham et al. (1994) examined listeners' abilities to summarize discourse length texts presented by a DECtalk and Echo+ speech synthesizers at slow (7.5 wpm) and normal (140 wpm) speech production rates. Texts were also differentiated according to length, propositional complexity, and topic familiarity. The written summaries were rated by judges as either full, partial, changed or fragmented renditions of the original texts[1] to provide a global measure of the concordance between the listener's summary and the original text (i.e., “Did the listener understand the text?”) and to take into account the practical effects of mishearings and the influence of background knowledge. As expected, individuals listening to high quality DECtalk speech produced more accurate summaries than did individuals listening to lower quality Echo+ speech output. Slow speech production rate (word method) listeners did better than individuals listening to synthetic speech produced at a normal speech production rate (sentence method). A three-way interaction between voice, speech production rate, and text complexity (i.e., number of unique propositions in the text) showed that for both simple and complex texts, Echo listeners' performance declined for texts presented at a normal rate. However, for the DECtalk group, normal rate summaries declined in quality only for propositionally complex texts. Using a limited processing capacity model developed by Pisoni et al. (1987), the authors hypothesized that device-related factors such as lower quality speech, faster speech production rates (i.e., sentence method), and increased text complexity, will raise the speech processing demands and threaten the listener's ability to accurately comprehend the spoken text.

## Speech Output Methods in AAC

The speech output of an AAC device is also constrained by the method by which the text material is output or spoken. Three output methods are utilized by most AAC technologies.

*Word method* output consists of sequences of spoken words, interspersed by periods of silence during which the communicator formulates the next word. Using the word method, the sentence “Robert Fripp is my favorite guitarist” would be constructed and spoken as follows:

*Example of Word Method*

++++++_Robert +++++_Fripp +is +my ++favorite +++++++++_guitarist[2]

*Sentence method* output consists of spoken sentences produced after the communicator has constructed the message. Typically, each sentence is separated by a period of silence in which the message is being constructed by the *communicator. Using the sentence method, the example sentence would be constructed and spoken the following way:*

*Example of Sentence Method*

+++++_+++++_+++++++++++++= Robert Fripp is my favorite guitarist

[1] See the Methods section for more complete description of the discourse rating method.

[2] The preceding characters signify the keystrokes made by the AAC user to produce the message (“+”-letter/code, “_”-space/speak preceding word, “=”-speak preceding message). Each keystroke also indicates a fixed amount of time (e.g., 1 sec.) associated with letter or word selection. The relatively few keystrokes used with some words (e.g., “favorite”) demonstrates word encoding.

Finally, *mixed method* output consists of a combination of spoken words and spelled-out words interspersed with periods of silence. The formulation and production of a sentence using a mixed method would appear as follows:

*Example of Mixed Method*
+R+o+b+e+r+t _+F+r+ip+p _+is +my ++favorite +g+u+i+t+a+r+i+s+t

*Current output method use by AAC users.* The current output methods developed for AAC users have not been guided by empirical research findings. These methods have been developed using common sense views of spoken communication (e.g., people speak in sentences) or by the technical limitations of the speech synthesis technology (e.g., the particular device will not output sentences with silent intervals between individual words). For example, the production of spelled-out words characterizing the mixed output method is generally due to the inability of a particular AAC device to pronounce words not stored in its lexicon (e.g., Prentke Romich TouchTalker). This output method could be particularly problematic for the listener who must process and memorize the ongoing spelling, while attempting to remember the preceding words in the text and make a meaningful interpretation of the ongoing discourse. Following a language-processing model discussed by Duffy et al. (1992) and Higginbotham et al. (1994), the additional short-term memory requirements associated with the mixed output method could tax the listener's comprehension-processing abilities and threaten comprehension accuracy. In contrast, the sentence output method should be easier to comprehend because the listener does not have to remember letter sequences. However, Higginbotham et al. (1994) showed that the relatively fast speech production rate associated with sentence method output may contribute to misperception and incomplete processing of syntactic, semantic, and discourse relationships. Word method output should be the easiest of the three methods to comprehend since each spoken word is followed by a silent interval, providing additional time for processing and rehearsal.

The purpose of this study is to extend the work of Higginbotham et al. (1994) by evaluating the effects of the three most commonly used AAC voice output methods on the quality of discourse summary abilities of adult listeners. Based on the findings of Higginbotham et al. (1994) and the rationale discussed above, the word-by-word output method was hypothesized to be easier to comprehend than the sentence output method, which, in turn, should be easier to comprehend than the mixed output method. The accuracy with which subjects comprehended synthetic speech passages was assessed using discourse summarization measures developed by Higginbotham et al. (1994). The complexity and familiarity of text materials were manipulated to determine how these factors, in combination with the ouput methods, affected discourse comprehension. Finally, examples of misunderstood and fragmented texts were provided to specify the types of listener comprehension problems encountered and their relationship to the experimental conditions.

## Methodology

### Subjects

Thirty able-bodied university students from the State University of New York at Buffalo served as subjects for this study. Their participation was solicited through posters on university bulletin boards and an ad in the student newspaper. Subjects ranged in age from 18 to 47 (M = 25, SD = 7.57). Subjects reported no history of reading, language, or learning difficulties. All subjects passed a basic screening for cognitive and language competence using the Aphasia Language Profile[3] (Keenan & Brasswell, 1975) and passed a pure tone screening (25db SPL @.25,.5,1,2,4 kHz) (ANSI, 1969). All subjects spoke English as their first language and spoke American Standard English dialect. Subjects reported no or only minimal prior exposure listening to synthetic speech (e.g., television, software demonstrations). Subjects received $5 per hour for their participation in this experiment.

### Materials and Instrumentation

The four paragraphs used for this study were originally developed by Kintsch, Kozminsky, Streby, McKoon, and Keenan (1975)[4] (see Appendix A). These passages were adapted from a children's history book and physical science articles from *Scientific American*. The texts were controlled across three dimensions: text length, text complexity, and topic familiarity. As shown in Table 1 the number of words per text ranged between 66 and 71 words with sentence length. The texts were equated for the number of propositions per passage and proposition structure. For the purposes of this study, a proposition consisted of a predicate (e.g., verb) plus one or more arguments (e.g., referring expressions). If a proposition contained an argument stated in a prior proposition, it was considered dependent to the original proposition. Appendix B shows the propositional structure and dependency relations of a short text used in Kintsch, et al.s' (1975) investigation. Propositions with Level 1 and 2 dependency relations were classified as main propositions, whereas those containing argument dependencies greater than Level 2 were considered subordinate propositions. The four stimulus texts were roughly equivalent across propositions and argument types (see Table 1). For this study text complexity was related to the proportion of unique arguments within a text passage with complex texts possessing twice the number of unique arguments as simple texts. Kintsch et al. (1975) found comprehension difficulties directly related to the proportion of unique arguments in a text.

The other dimension, topic familiarity, related to the psychological difficulty of the text material. History passages focused on well-known stories framed in a narrative discourse structure that was thought to facilitate discourse

[3] Note that all subjects achieved a perfect score on the ALPS.

[4] These paragraphs were also the long stimulus passages used by Higginbotham et al. (1994).

**TABLE 1. Quantitative characteristics of the stimulus texts.**

| | Text length (propositions) | | | | Text complexity (arguments) | | | | Content familiarity | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Reading difficulty[c] | | | |
| Name | # Words | Total | (Main/Sub)[a] | Level | Total | Unique | Percent | Type | Familiarity rating[b] | Flesch | SMOG | Sentence length | Spelled-out words Number/Percent |
| Joseph | 66 | 23 | (9/14) | Simple | 39 | 7 | 18 | History | 4.8 | 71 | 10 | 18.7 | 21/32 |
| Assyria | 67 | 24 | (12/12) | Complex | 42 | 17 | 40 | History | 5.2 | | | | 25/37 |
| Astroid | 70 | 25 | (11/14) | Simple | 38 | 6 | 16 | Science | 4.7 | 60 | 10 | 15.7 | 25/36 |
| Comet | 71 | 25 | (9/16) | Complex | 45 | 15 | 33 | Science | 4.7 | | | | 28/39 |

[a]The propositional analysis shows the distribution of main propositions (Level1 and Level2) and subordinate propositions (Level 3 and lower).

[b]The familiarity score is derived from a frequency analysis of nouns found in the 250,000 word Wordnet© corpus (Miller, 1990). These data were based on an average score for each noun type in the passages that were computed using MacWordnet (version 1.4). The scores are based on an 8-point scale ranging from 1 (extremely rare) to 8 (extremely familiar). The scores 4 and 5 indicate common and very common words, respectively. WordNet is a copyright of Princeton University. All rights reserved.

[c]The Flesch procedure is a measure of abstractness of a text. The scores vary from 0 (very abstract) to 100 (very concrete). The SMOG score is a measure of text complexity that is derived by determining the proportion of multisyllabic words in a text. The score reflects the grade level at which the text can be read with at least 90% comprehension. The individual texts were collapsed into history and science groups for analysis purposes. Despite this combination, not enough sentences were available to meet the criteria for a formal SMOG. Therefore, this score should be regarded with some caution.

recall. On the other hand, the science texts contained new information that should increase comprehension processing effort due to a smaller established memory base and less supportive expository discourse structure. However, the history and science passages were equivalent in terms of noun frequency or familiarity and concreteness, and were comprehensible at a 10th grade reading level (Flesch, 1948; McLaughlin, 1969; Miller, 1990). Sentence length for the combined history texts (18.7 words) was slightly longer than for the science texts (15.7 words). It should be noted that Kintsch and his colleagues (1975) found science texts to be significantly more difficult to recall than history texts.

The portable DECtalk Speech Synthesizer™[5] generated the speech stimuli used in this investigation. DECtalk was chosen for this study based on its demonstrated high level of segmental intelligibility and comprehensibility and wide use in current AAC technologies. It was also thought that if comprehension differences were found with this synthesizer, more substantial differences are likely to occur with other synthesized voices. The default voice setting (Perfect Paul) was used for all output.

To precisely control the speech presentation rate and ensure uniformity across the output methods, all synthetic speech stimuli were digitized for presentation. The vocabulary items (alphabet, single words, sentences) were first typed into text files and processed by the DECtalk synthesizer. Output from the synthesizer was low-pass filtered at 10kHz and digitized using an 8-bit Farralon SoundRecorder™ A-D converter at 22 kHz sampling rate. The digitized stimuli were stored on the hard drives of three Macintosh Classic computers used for stimulus presentation.

A Hypercard-based software application (Discourse Delivery System[6]) was developed to control all phases of training and delivery of the experimental stimuli. A separate computer was used to deliver the experimental materials to each subject (i.e., 3 subjects could be run simultaneously). The speech stimuli were presented via a Realistic SA-150 amplifier and SS110 headphones. Subjects sat in front of their computers, each in a small partitioned "cubbie" within a quiet room. All speech was presented at approximately 30 dB above the ambient noise level of the room. However, subjects were allowed to adjust the amplifier volume upward from this level to maximize listening comfort.

Subjects listened to paragraph materials presented using one of three different output methods described below:

1. *Word method.* Paragraphs were constructed with single words interspersed with periods of silence.
2. *Sentence method.* Paragraphs are produced as sentence-level connected speech with silent periods interspersed between sentences.
3. *Mixed method.* The 200 most frequently used words were produced as single words and the rest were spoken as single letters. (Beukelman, Yorkston, Poblete, and

[5]The portable DECtalk II synthesizer was developed by the Institute on Applied Technology, Children's Hospital, 300 Longwood Avenue, Fegan Plaza, Boston, MA 02115, and is an adaptation of the commercial DECtalk III speech synthesis technology.

[6]All software and digitized speech samples can be obtained by writing to the author.

Naranjo, 1984). Spelled-out words accounted for 32% to 39% of the words in each paragraph (see Table 1). Textually complex passages included a slightly higher percentage of spelled out words as did science passages (see Appendix A).

To accurately represent the temporal characteristics of the AAC device user's communications, the duration of each silent interval was calculated based on the probable number of keystrokes required to encode the word with a duration value of 1 second per keystroke. Single letters, punctuation, and spaces were assigned a keystroke value of 1. High frequency words were assigned a keystroke value of 2 to simulate a two-keystroke encoding technique. Each of the low frequency words in the word method condition were preceded by a silent interval of a duration equivalent to the number of keystrokes required to produce the word. Similarly, the silent interval preceding sentences in the sentence method condition was based on the number of keystrokes required to encode the sentence. The speech production rate for all output methods averaged 7.5 wpm. For the sentence method condition the within-sentence speech production rate was 149 wpm, but 7.5 wpm when the silent intervals were considered. Two short beeps were placed at the beginning and the ending of each paragraph to notify the listener when the task began and ended.

## Procedure

Listeners were randomly assigned to one of the three output method conditions and individually tested during two 1-hour sessions. Sessions took place on consecutive days. On the first day, listeners were screened for hearing and language comprehension ability, then took part in a self-paced training procedure run by the computer program. During the first part of the tutorial participants learned how to perform the experimental tasks and how to produce written summaries of the stimulus passages. In particular they were told to listen to the passage, turn to the appropriate page in their response books after the passage was spoken, complete a 1-minute mathematics task, then to summarize the passage in writing. They were instructed to be as complete as possible, but that exact recall was not necessary. The first two passages consisted of two of Aesop's fables presented in the sentence mode. The first passage was accompanied by a written transcript to diminish the immediate burden of comprehending synthetic speech. Each subject then took a short break, in which a research assistant would look over the two written summaries to make sure that the listener understood the comprehension task. If summarization problems occurred, the listener was reminded to produce complete paragraphs. After the break, subjects practiced listening and summarizing two additional paragraphs similar to the experimental texts, except shorter in length (approximately 16 words). These texts were produced by the same output method (word, sentence, mixed) that the subject would hear in the experiment. Again, listeners were instructed to write down what they remembered about the text. One or 2 days after the practice session, subjects listened to four experimental texts that were randomly presented for summarization.

Each listening/summary trial went the following way: First, a passage was presented to the listener. After the passage was completed, listeners were instructed by the computer to turn a page in their response packets and to work a set of multiplication problems for a 1-minute period. This step was taken to diminish the effect of rote memorization of the text and to reduce the chance of a ceiling effect for the word and sentence conditions. When the math period was completed, the computer then instructed listeners to turn the page in their response booklet and write down a summary of the preceding passage. After completing the passage, the listener then depressed a button on the computer, initiating the next listening/recall sequence.

## Discourse Rating Procedure

The Qualitative Rating of Discourse Summaries (QRDS) (Higginbotham, Lundy, & Scally, 1993) consisted of four exhaustive, mutually exclusive, and ordinally related categories. The QRDS was used to evaluate the summarization accuracy of a listener or reader (Higginbotham et al., 1993, 1994). The categories included Full Summary (i.e., "the listener completely understood the message"); Partial Summary (i.e., "the listener understood most of the message"); Changed Summary (i.e., "the listener misinterpreted the message); and Fragmented Summary (i.e., "the listener was unable to make sense of the message"). A full set of QRDS definitions and defining features are listed in Appendix C.

Two graduate students and the project director (first author) served as raters for the study. Two of the raters had classified transcripts for previous investigations and all had collaborated on the design of the QRDS protocol. To classify the texts, the raters first independently rated each transcribed text according to the established definitions and rating protocol[7]. Interrater agreement among the three judges for the independent rating phase was .74, .76, and .82. Employing a structured consensus procedure detailed in Appendix D, the raters then resolved the discrepant ratings by comparing the individual ratings with the published protocol, then discussing each rater's rationale for producing the score. Using this procedure, agreement was achieved for all but three (2.5 percent) of the texts. A majority vote was employed to resolve these rating discrepancies.

Nine months after the initial classification procedure, three judges (including two of the original raters) rescored 25% of the original summaries. Intersession agreement between structured consensus scores was .90 (.91), as measured by the Kappa coefficient (Suen & Ary, 1989)[8].

## Design and Analysis

One between-subjects variable, output method, was combined with two within-subjects variables, text complex-

[7] Information regarding the development of the QRDS can be found in Higginbotham et al. (1994). Training materials can be obtained by writing to the first author.

[8] Suen and Ary (1989, p. 113) indicate that kappa coefficients over .8 are an indication of good reliability. A simple agreement score is provided after the kappa coefficients (in parentheses).

ity and topic familiarity, to form a 3 × 2 × 2 mixed models design. Data were analyzed using a log-linear analysis approach for response variables with an ordinal level metric (SAS Institute, 1994). The main and interaction effects were tested using the Wald Chi Square test (SAS, 1994). The a priori hypothesis concerning a monotonic trend across output methods (word method > sentence method > mixed method) was tested using two, single-tailed, orthogonally paired comparison tests based on Goodman's approach to multiple comparisons for multifactor tests of independence (Marascuilo & Serlin, 1988). For this test, the two nonproblematic summary categories (full & partial) were compared to the two problematic summary categories (changed & fragmented). This comparison was performed because it provides a more specific test to the hypothesis developed above than does the analysis of variance procedure. All other significant comparisons revealed in the general analysis of variance were further analyzed using a log-linear analog of the post hoc Scheffé procedure (Marascuilo & Serlin, 1988).

## Results

Table 2 presents the summarization quality data for the three different output methods. Table 3 presents the results of the mixed model ANOVA.

### Output Methods

A significant monotonic trend was found confirming our prediction that the proportion of Full and Partial rated summaries would decline from the word method to the sentence method to the mixed method ($Z_{V = 3}$[9] = 2.02, $p < 0.05$). This finding was further supported by not finding a quadratic trend in the data ($Z_{V=3} < 1$). As shown in Table 2, 68% of the texts recalled by word method summaries were judged to be either full or partial renditions of the original texts, compared to 58% of the sentence method summaries and only 43% mixed method summaries. The decrease of the full and partial categories across output method groups was associated with a concomitant increase in the proportion of texts judged to be changed or misunderstood. Interestingly, only the mixed method contained summaries judged to be fragmented (5%). This group also possessed the largest percentage of changed summaries (53%) for any of the three Output Method groups. Post hoc contrasts of single and combined summarization categories across output method groups did not reveal any other statistically significant differences.

### Topic Familiarity and Text Complexity

A main effect for topic familiarity was noted ($X^2_1 = 30.04$, $p < .0001$). Post hoc testing showed that the more familiar history texts were rated as Full or Partial compared to the less familiar science texts, which were predominantly rated in the Changed category ($Z_{V = \infty} = 3.633$, $p < .05$). No other post hoc contrasts were statistically significant. No other main or interaction effects were noted for the variables studied.

**TABLE 2. Percent of listeners' summarization scores across word, sentence, and mixed output method conditions.**

| QRDS categories | Output method | | | |
|---|---|---|---|---|
| | Word | Sentence | Mixed | All |
| Full | 23 | 20 | 5 | 16 |
| Partial | 45 | 38 | 38 | 40 |
| Changed | 32 | 42 | 52 | 42 |
| Fragment | 0 | 0 | 5 | 2 |

### Analysis of Listener Miscomprehension

One way to assess the "real world" problems faced by the listeners in this study was to examine the kinds of misunderstandings evidenced in their written summaries. The following descriptive analysis was meant to familiarize the reader with some of the types of problematic summary responses produced by the subjects in this study.

When inspecting the individual changed and fragmented texts, several trends were noted. Table 4 shows that for each original text, the frequency of associated problematic summaries was related to output method, topic familiarity, and, to a lesser extent, text complexity. Across all conditions most of the problematic texts contained correct information, interspersed with factual errors. Also, many of the summaries were factually correct, but missing the majority of relevant text elements.

These findings are consistent with Higginbotham and Baird's (in press) analysis of synthesized discourse passages, in which the DECtalk summaries were characterized by lexical omissions and substitutions, and discourse level recasts.

For changed texts produced under the word method, misunderstandings were typified by factual errors related to incorrect inferencing. Such a problem is illustrated in the following passage about asteroids:

*Original*: Asteroids are miniature planets that orbit around the sun. Hundreds of asteroids have been identified, but it is difficult to keep track of them, since all asteroids are alike. An asteroid is identified only by the position of its orbit. Even after an orbit has been determined, it is often lost because the orbit changes due to the influence of the large planets which deflect the asteroid from its orbit.

*Summary*: Asteroids are planets WHICH ORBITS AROUND THE SUN. Asteroids can be identified by its change in orbits. Asteroids are similar. Asteroids are hard to be identified in hundreds. Asteroids changes its' orbits around planets.

[9] Z is the statistic used in Goodman's loglinear approach to multiple comparison testing (Marascuilo & Serlin, 1988). Degrees of freedom (v) is computed as (row-1)*(column-1) for the contingency table.

**TABLE 3. Log linear analysis of variance table.**

| Source | DF | Wald ChiSquare | Prob>ChiSq |
|---|---|---|---|
| Output Method (OM) | 2 | 17.150672 | 0.0002* |
| Subjects | 27 | 45.116361 | 0.0158 {Error} |
| Text Complexity (TC) | 1 | 2.005464 | 0.1567 |
| OM*TC | 2 | 2.939743 | 0.2300 |
| Topic Familiarity (TF) | 1 | 30.042197 | 0.0000* |
| OM*TF | 2 | 1.154599 | 0.5614 |
| TC*TF | 1 | 2.803716 | 0.0940 |
| OM*TC*TF | 2 | 1.625628 | 0.4436 |

*Note*: Summary of Fit: Rsquare (U)—0.496; Observations (or Sum Wgts)—120.
* statistically significant results for $p < 0.05$.

Here, the listener makes several factual errors and omissions and fails to recount the major causal relationships in the passage (e.g., asteroids are hard to keep track of because they are alike; It is difficult to track their orbits because they are influenced by the gravitational forces of large planets). This particular summary was also characterized by a reduction in text elements compared to the original text, which was characteristic for most of the changed word method summaries.

Problematic sentence method texts contained fewer words per summary, more factual errors, and more omissions of entire clauses and sentences compared to the other two groups. In the following example, the listener indicates an inability to recall an individual lexical item by leaving a blank in the text:

*Summary*: ~~abro~~ are plants[10] that are in orbit. They all look alike. They are very clustered because there are so many.

Typical of the sentence method changed summaries, this text consists of only a few sentences while omitting many ideas contained in the original text. The ungrammatical style of the last sentence was noted throughout all text summaries.

Compared to either the word method or sentence method, mixed method texts contained more sentence fragments, omissions of obligatory lexical constituents, permutations of the original propositional order, and insertions of significant amounts of new information. The following mixed method summaries of the Comets passage depicts many of these phenomena:

*Original*: A comet is a celestial fountain spouting from a large snowball in space. We see the fountain as the head and tail of the comet. The tail extends for millions of miles but we never see the snowball which has a diameter of a few miles. A comet shines with reflected sunlight. Along its path it strews debris in space, which seen from the earth appears as the zodiacal light.

*Summary*: A comet is a celestial fountain. We see the fountain as a head & a tail. the tail (fountain?) is millions of miles long. the ? is only a few miles long.

*Summary*: An orbit is a celestial --------
which has a fountain, --------
a snowball. we can see the tail end of the orbit -- --- --------
4 million miles away
Its' diameter consists of
millions of orbits

*Summary*: Something about snowballs, not finding any with a diameter of miles long.

The first summary provides some correct factual material but is missing so many elements contained in the original

[10] Misspellings occurred frequently in the summarized texts across all output method conditions. Errors of this type did not lower the discourse rating for the summary.

**TABLE 4. Frequency and proportion[a] of discourse passages that were rated changed or fragmented.**

| Discourse passage | Complexity/Familiarity | Output method | | | |
|---|---|---|---|---|---|
| | | Word | Sentence | Mixed | Total (avg. prop.) |
| Asteroids | Simple/Science | 3 (.3) | 5 (.5) | 7 (.7) | 15 (.5) |
| Comets | Complex/Science | 6 (.6) | 7 (.7) | 8 (.8) | 21 (.7) |
| Joseph | Simple/History | 3 (.3) | 2 (.2) | 3 (.3) | 8 (.27) |
| Assyria | Complex/History | 1 (.1) | 3 (.3) | 5 (.5) | 9 (.3) |
| Total (avg. prop.) | | 13 (.33) | 17 (.43) | 23 (.58) | 53 (.44) |

[a] Proportion data are shown in parentheses.

passage that it was rated as being changed. The second summary is indicative of the magnitude of comprehension difficulties faced by some listeners in the mixed method condition. This summary is characterized by numerous sentence fragments, missing constituents, and the insertion of new and incorrect information. For example, the word "orbit" used as a descriptor in the comet summary was first heard in the "Asteroids" passage and was misapplied here. Also, the original phrase "for millions of miles" was incorrectly recast as "4 million miles." Taken together, these departures from the original passage attest to the listener's perceptual and higher level processing difficulties. The final summary was rated as a fragmented summary. The listener was unable to recall most of the elements of the passage, and the summary reflected a basic misunderstanding of the text.

## Discussion

The results of this study support the view that output method can have considerable impact on the comprehension performance of listeners. More specifically, we confirmed our hypothesis for a monotonic trend in the comprehension of synthetic speech discourse. Examination of the frequency distributions of the QRDS ratings showed systematic changes to occur primarily in the proportion of full and changed ratings across the three output method groups. Less substantial changes were observed for the partial and fragmented categories. Interestingly, only the mixed method condition contained fragmented summaries, bolstering the notion that increased processing difficulties are associated with this output method.

The statistically significant differences found in these data are consistent with our initial work in this area (Higginbotham et al., 1994) and support Duffy et al.'s (1992) resource model of comprehension processing for explaining the summarization difficulties experienced by sentence and mixed method listeners. The significant main effect for topic familiarity and nonsignificant main effect for text complexity were consistent with our earlier study. The failure of these factors to interact with the output method condition suggests that differences in proposition structure and topic may not influence summarization of texts produced by different output methods. On the other hand, visual inspection of the data (Table 4) indicates these factors may affect text comprehension, but not consistently enough or at a sufficient magnitude to produce statistically significant differences in this study.

Finding statistically significant differences for output method are particularly interesting given that the DECtalk voice is the highest quality commercial speech synthesizer available and is statistically indistinguishable from natural speech in some speech intelligibility studies (e.g., Mirenda & Beukelman, 1990). These findings provide evidence that, even with quality synthesized speech, output method may play a considerable role in the comprehension of discourse produced through augmentative means. The memory and processing requirements related to remembering individual letters and spelling out words appeared to interfere significantly with understanding the text material. The prevalence of constituent omissions, sentence fragments, and misunderstood passages reported above provide evidence for this problem. The sentence method appeared to cause listeners fewer specific problems compared to mixed method summaries. But the production of shorter and generally lower quality summaries compared to the word method suggests that listeners may not have been able to process the connected speech as completely as when the output was presented a word at a time. Given these results with high quality DECtalk speech, one may speculate that speech output from lower quality synthesizers would exacerbate discourse comprehension processing difficulties, particularly for sentence and mixed output methods.

In a recent dissertation, Mathy-Laikko (1992) studied the effect of the three output methods on subjects' abilities to complete portions of the Revised Token Test that were administered via a DECtalk speech synthesizer (7 wpm). She found word and letter method (i.e., all words spelled out) to be superior to sentence method performance. Although the results of Mathy-Laikko's study appear to conflict with our results, her use of a highly specific direction-giving task, coupled with a small, closed vocabulary set could restrict the listener's lexical choices to relatively few words, thereby facilitating identification. In addition, the letter method consisted of spoken spellings for all words, in contrast to the mixed method, in which only lower frequency words were spelled out. Coupled with the small, closed vocabulary set and prior task experience, subjects in Mathy-Laikko's study could make accurate word guesses from the first or second letter of a word. The use of such strategies could significantly reduce the cognitive burden associated with remembering spoken spellings over protracted periods. However, with longer, less familiar, and more open-ended texts found in the present study, language processing difficulties related to letter sequence memorization would render a letter-based output method the most difficult output method to comprehend.

### Methodological Issues

The methodological modifications used to assess output method comprehension differences (e.g., silent interval correction, distractor task) appeared to make summarization more difficult and reduced the ceiling effect found with DECtalk speech in our earlier study (Higginbotham et al., 1994). For example, the proportion of full summaries was reduced from about 58% in the previous study to 23% in this investigation. The distractor task was not thought to threaten the external validity of the results but to increase the sensitivity of the experimental task. This investigation departed from normal communication situations in two important ways. First, in this study listeners were required to wait until the entire passage was completed before responding to the message. In a conversational context listeners would have the opportunity to interact with the speaker during message production (e.g., make comments, repeat message parts, request repair). Second, in the present investigation, listeners were given no feedback as to the validity of their interpretations during or after the stimulus passage was produced. In a real world situation interactants expect such feedback by the message producer. Thus, the introduction of a distractor task was not thought to limit the generalizability of these findings as much as the noninteractional quality of the experiment itself.

The results of this study may be considered fairly robust because of the fairly liberal criteria used for screening language competence. However, the possible variability within subject groups could have decreased the sensitivity of the experimental procedure in detecting less prominent comprehension differences. Assessing discourse processing skill prior to testing could provide a more homogeneous subject grouping, or data that could be used as a statistical covariate. Testing could include a different test of discourse comprehension, or a parallel version of the experimental materials administered the same way to each subject. Pretesting in this manner could provide a sensitive measure of comprehension for studying the impact of AAC designs and text-related factors on discourse processing.

### *Implications for AAC Design and Use*

The results presented here have implications for the design of AAC technologies as well as for clinical training. First, AAC manufacturers need to develop AAC technologies that allow spelled-out words to be spoken as single words after they are produced, rather than restrict output to the spoken production of individual letters. Second, our findings call into question the use of connected speech for transacting synthetic speech spoken communication. Rather, we suggest that the comprehension of preplanned texts (e.g., lectures, stories, reports) could be enhanced by inserting a short period of silence after each spoken word, phrase, or clause, especially if intonation and stress patterns are preserved. Research by Venkatagiri (1991) has shown that a 250-millisecond period of silence interspersed between words can significantly improve sentence transcription accuracy for low quality synthetic speech. Another means of improving comprehension could be to provide a visual display in addition to synthetic speech. Even a small display (e.g., 40 characters) oriented toward the communication partner, could provide enough information to successfully resolve mishearings and misunderstood utterances. These suggestions await verification through further experimental research and clinical trials.

For clinical training we suggest that clients should learn to employ all output methods available to them on their particular device. High quality speech synthesis allows both the word and sentence output methods to be used for everyday communication purposes. The word method could be employed if listener comprehension problems are anticipated or arise. Output methods involving spelling should be limited to situations requiring communication repair and not be used for typical communications.

## References

**American National Standards Institute** (1969). *Specifications for audiometers* (ANSI S3.6-1969). New York: ANSI.

**Beukelman, D., Yorkston, K., Poblete, M., & Naranjo, C.** (1984). Frequency of word occurrence in communication samples produced by adult communication aid users. *Journal of Speech and Hearing Disorders, 49,* 360–367.

**Byrd, M.** (1985). Age differences in the ability to recall and summarize textual information. *Experimental Aging Research 11,* 87–91.

**Duffy, S. A., & Pisoni, D. B.** (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech, 35,* 351–389.

**Flesch, R.** (1948). A new readability yardstick. *Journal of Applied Psychology, 32* (3), 221–233.

**Foulds, R.** (1980). Communication rates of nonspeech expression as a function of manual tasks and linguistic constraints. *Proceedings of the International Conference on Rehabilitation Engineering,* 83–87.

**Harris, J.** (1987). Speech comprehension and lexical failure. In R. Reilly (Ed.), *Communication failure in dialogue and discourse: Detection and repair processes* (pp. 81–98). New York: North-Holland.

**Higginbotham, D. J.** (1992). Evaluation of keystroke savings across five assistive communication technologies. *Augmentative and Alternative Communication, 8,* 258–272.

**Higginbotham, D. J., & Baird, L.** (in press). Discourse analysis of listeners' summaries of synthesized speech passages. *Augmentative and Alternative Communication.*

**Higginbotham, D. J., 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,* 258–272.

**Higginbotham, D. J., Lundy, D. C., & Scally, C.** (1993). *Qualitative ratings of discourse summaries: Definitions and procedure manual.* Unpublished manuscript, State University of New York at Buffalo.

**Hirst, G., McRoy, S., Heeman, P., Edmonds, P., & Horton, D.** (1993, November). *Repairing conversational misunderstandings and non-understandings.* Paper presented at the International Symposium on Spoken Dialogue, Waseda University, Tokyo, Japan.

**Jefferson, G.** (1974). Error correction as an interactional resource. *Language in Society* (3), 181–199.

**Keenan, J., & Brassell, E.** (1975). *Aphasia language performance scales.* Murfreesboro: Pinnacle Press.

**Kintsch, W.** (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182.

**Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M.** (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14,* 196–214.

**Logan, J. S., Greene, B. G., & Pisoni, D. B.** (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America, 86,* 566–581.

**Luce, P. A., Feustel, T. C., & Pisoni, D. B.** (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors, 25,* 17–232.

**Marascuilo, L., & Serlin, R.** (1988). *Statistical methods for the social and behavioral sciences* New York: W.H. Freeman.

**Mathy-Laikko, P.** (1992). *Comprehension of augmentative and alternative communication device output methods.* Unpublished doctoral dissertation, University of Wisconsin–Madison.

**Mathy-Laikko, P., West, C., & Jones, R.** (1993). Development and assessment of a rate acceleration keyboard for direct-selection augmentative and alternative communication users. *Technology and Disability, 2,* 57–67.

**McLaughlin G.** (1969). SMOG grading—a new readability formula. *Journal of Reading, 12,* 639–646.

**Miller, G. A.** (Ed.) (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography, 3,* (4).

**Mirenda, P., & Beukelman, D. R.** (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication, 3,* 120–128.

**Mirenda, P., & Beukelman, D. R.** (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternative Communication, 6,* 61–68.

**Pisoni, D. B., Manous, L. M., & Dedina, M. J.** (1987). Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language, 2,* 303–320.

**Raghavendra, P., & Allen, G. D.** (1993). Comprehension of synthetic speech with three text-to-speech systems using a sentence verification paradigm. *Augmentative and alternative communication, 9,* 126–133.

**Ralston, J. V., Pisoni, S. E., Lively, S. E., Greene, B. G., & Mullenix, J. W.** (1991). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors, 33,* 471–491.

**Sacks, H., Schegloff, E., & Jefferson, G.** (1974). A simplist systematics for the organization of turn-taking in conversation. *Language, 50,* 6996–735.

**SAS Institute** (1994). *JMP® (version 3.02).* Cary, NC: SAS Institute, Inc.

**Suen, H. K., & Ary, D.** (1989). *Analyzing quantative behavioral observation data.* Hillsdale: Lawrence Erlbaum Associates, Inc.

**Talbot, M.** (1987). Reaction time as a metric for the intelligibility of synthetic speech. In J.A. Waterworth (Ed.), *Speech and language-based interaction with machines: Towards the conversational computer.* Chichester, England: Ellis Norwood Ltd.

**Vanderheiden, G.** (1988). A unified quantitative modeling approach for selection-based augmentative communication systems. In L. Bernstein (Ed.), *The vocally impaired: Clinical practice and research* (pp. 40–83). Philadelphia: Grune & Stratton.

**Venkatagiri, H.** (1991). Effects of rate and pitch variations on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 7,* 284–289.

Received July 25, 1994
Accepted February 14, 1995

Contact author: D. Jeffery Higginbotham, PhD, Department of Communicative Disorders and Sciences, 109 Park Hall, State University of New York at Buffalo, Buffalo, NY 14260. E-mail: cdsjeff@ubvms.cc.buffalo.edu

## Appendix A

### *Stimulus Passages Used in This Study (From Kintsch et al., 1975)*

*Comet (Text Complexity = Complex/Topic Familiarity = Science)*

*A comet*[1] *is a celestial fountain spouting* from a large *snowball floating* through *space*. We see the *fountain* as the head and *tail* of the comet. The *tail extends* for *millions of miles* but we never see the *snowball* which has a *diameter* of a few *miles*. A *comet shines* with *reflected sunlight*. Along its *path* it *strews debris* in *space*, which seen from the earth *appears* as the *zodiacal* light.
(spelled-out words = 28)

*Astroids (Text Complexity = Simple/Topic Familiarity = Science)*

*Asteroids are miniature planets* that *orbit* around the *sun*. *Hundreds* of *asteroids* have been *identified*, but it is *difficult* to keep *track* of them, since all *asteroids* are *alike*. An *asteroid* is *identified* only by the position of its *orbit*. Even after an *orbit* has been *determined*, it is often *lost* because the *orbit* changes *due* to the *influence* of the large *planets* which *deflect* the *asteroid* from its *orbit.*
(spelled-out words = 25)

*Assyria (Text Complexity = Complex/Topic Familiarity = History)*

The *rich* country of *Assyria* was *ruled* by a *king* who lived at *Nineveh*. This *king* loved *luxury* and he *built* a *wonderful palace*. On the *roadway* that led to the *palace* he placed many *huge* and *fantastic statues*, *depicting bulls* with *wings* and *lions* with the heads of men, just as a *rich* man *nowadays* might *plant trees* along the *driveways* to his home.
(spelled-out words = 25)

*Joseph (Text Complexity = Simple/Topic Familiarity = History)*

Although *Joseph* was a *slave* in *Egypt* and it was *difficult* to *rise* from the class of *slaves* to a *higher* one, *Joseph* was so *bright* that he became a *ruler* in *Egypt*. *Joseph's wicked brothers*, who had once planned to *kill* him came to *Egypt* in *order* to *beg* for *bread*. There they found that *Joseph* had become a great *ruler.*
(spelled-out words = 21)

[1] Italics indicate the words spelled out in the stimulus passages. Spelled-out words were used in the calculation of silent intervals for each output method condition. These words were also spoken as individual letters in the mixed method condition.

## Appendix B

### *Example of a Propositional Analysis of a Short Text (From Kintsch et al., 1975)*

Each line indicates a proposition. The first item in the proposition is the predicate; subsequent words are the related arguments. Indentations indicate the level of dependency between propositions.

1 (LOVE, GREEK, ART)

2 (BEAUTIFUL, ART)

3 (CONQUER, ROMAN, GREEK)

4 (COPY, ROMAN, GREEK)

5 (WHEN, 3, 4)

6 (LEARN, ROMAN, 8)

7 (CONSEQUENCE, 3, 6)

8 (CREATE, ROMAN, 2)

Arguments: Greek, Art, Roman (3)
Text: The Greeks loved beautiful art. When the Romans conquered the Greeks, they copied them, and thus, learned to create beautiful art (21 words).

# Appendix C

## *Definitions for the Qualitative Rating of Discourse Summaries*

### FULL SUMMARY

*Gloss*

*The text was completely understood.*

A. *Definition*

The summary repeated the original text, including all textual elements (actors, relevant actions, events, locations, themes, and causal relations).

B. *Defining Features*

1. The summary is an exact repetition of the original text, or it includes all elements of the original text.

OR

2. The meaning of the summary is equivalent to that of the original text.

C. *Additional Considerations*

The summary is also considered to be Full in the following situations:

1. Change or omission of an element (or phrase) if:
   - the meaning of the altered element is consistent with that of the original text;
   - the altered element does not affect the meaning of other sentences or the central meaning of the summary compared to those of the original text.
2. Introduction of new words into the text, if the meaning of the text remains consistent with the meaning of the original paragraph (e.g., use of synonyms).
3. Omission of assumable elements and phrases (ellipsis).
4. Changes in punctuation, use of numbering (e.g., omission of commas, numbering of sentences), unless these changes appear to affect the meaning of the text.

### PARTIAL SUMMARY

*Gloss*

*My listener understood most of the text.*

A. *Definition*

One or more of the substantive elements of the original text were altered or omitted. However, the overall meaning of the summary is similar to that of the original text.

B. *Defining Features*

1. One or more elements of the summary are omitted or changed compared to the original text.

AND

2. The overall meaning of the summary is similar to that of the original text. The altered elements do not substantially depart from the overall jist of the summary.

C. *Additional Considerations*

The summary is also considered to be Partial in the following situations:

1. Text elements, phrases, sentences may be changed or omitted if the overall interpretation of the summary is comparable to that of the orginal text.
2. New elements may be introduced into the text if they do not significantly change overall meaning of the summary compared to that of the original text.

### CHANGED SUMMARY

*Gloss*

*My listener misinterpreted the text.*

A. *Definition*

Alterations made to the elements of the summary significantly change the overall meaning or jist of the original text. The summary is understandable unto itself.

B. *Defining Features*

1. A number of elements, phrases, and/or sentences of the summary are added, omitted, or changed compared to the original text.

AND

2. The altered elements significantly change the overall meaning of the summary compared to the original text.

AND

3. The summary makes sense in itself.

C. *Additional Considerations*

The summary is also considered to be Changed in the following situations:

1. The altered elements significantly affect the interpretation of the entire text or subsequent sentences in the text.
2. The summary may be reasonable and even elaborate, but its meaning is significantly different from that of the original text.

### FRAGMENTED SUMMARY

*Gloss*

*My listener was unable to make sense of the text.*

A. *Definition*

The summary is incoherent and/or only fragments of the original text were provided.

B. *Defining Features*

1. The majority of the elements are omitted or changed.

AND

2. It is difficult to make sense of the text.

OR

3. Subjects may report that they are unable to recall the paragraph.

C. *Additional Considerations*

The summary is also considered to be a Fragment in the following situations:

1. The subject may only comment on the difficulties related to comprehending the original paragraph.

## Appendix D

### *Procedures for Rating the Discourse Summaries*

*Preparation of Summary Materials and Work Area*

All summaries should be marked with an identification code and relevant group classification information. The identification code should be visible on the page and not reveal group identity.

The workspace should include a central area for placing summaries to be evaluated, four smaller areas for placing the rated summaries, as well as definitions and central examples. Guidelines for rating ambiguous or borderline summaries should be placed between relevant categories.

*Preparation for the Independent Rating Phase*

The independent rating procedure should be performed on all summaries raters.

Prior to beginning the independent rating phase, each rater should read the original texts in order to become familiar with the following elements of each text:

- characters and their actions
- objects and the setting
- relevant actions and events
- the theme
- causal relationships
- changes of state
- motivations of the characters

The raters should jointly generate a list of these elements for each text. The lists will be used during both the independent rating phase and the consensus phase as a tool to assist the raters in making their judgments.

*Independent Rating Phase*

All summaries written or spoken in response to a specific stimulus paragraph should be evaluated together. Before rating each group, the original text and list of elements that applies to that group should be reviewed.

Each summary should be compared only to the original paragraph. After rating all the summaries for the first time, the summaries should be reevaluated at least one time in order to ensure rating consistency. If after the second evaluation, the judge is still unsure about the rating of one or more summaries, those summaries that received a confidence rating of "unsure" (4) or "very unsure" (5) should be evaluated one more time before entering the consensus phase.

No time constraints are involved in the rating process. Reevaluation can occur immediately after the initial rating.
The rater should take care to evaluate each summary from the respondent's perspective. The rater should presume that the respondent understood the original paragraph unless there is an explicit indication that the text was not fully comprehended or was misunderstood (e.g., omitted, substituted, or added elements).

*Scoring Procedure*

The actual independent rating of each summary should be performed in the following manner:

1. Compare each summary to the original text and the operational definitions.
2. Determine whether the summary should be classified as a Fragmented Summary. If so, place into the Fragmented Summary tray and go to the next paragraph.
3. If the summary appears to be perfectly consistent with the original paragraph, then classify it as a Full summary. Examine the definitions and central examples for making a determination between the Full, Partial, and Changed categories, and place the summaries in the appropriate trays.
4. For those summaries that are difficult to score, consult the Guidelines for Rating Ambiguous and Borderline Summaries. Based on these guidelines, the definitions and central examples, place the summaries in the appropriate trays. All summaries should receive a summary rating at the end of each rating period.
5. Mark each classification on the space provided on the summary sheet. A rationale describing altered elements and justifications for the rating should be written down.
6. During reevaluation of each summary text, the rater should mark the summary rating and confidence rating on the area provided in the rating area, even if it is the same rating as provided previously. During the third independent evaluation, only mark summary and confidence ratings for those texts that were reviewed (those receiving a confidence rating of unsure or very unsure during the second evaluation).

*Preparation for Consensus Meeting*

Following the independent rating phase, a master database should be created that contains the following information for each summary:

- identifying information
- final summary rating for each independent rater
- final confidence rating for each independent rating
- space to record summary rating achieved through consensus

The information contained in the database should be analyzed to determine which summaries received discrepant ratings from the raters.

For those summaries that received the same rating by all 3 raters the data base should be amended to report these summary ratings as the final scores. The summary sheets (located at the bottom of each summary) should likewise be amended.

Discrepant ratings for a particular summary should be resolved by employing the consensus-making procedure listed below. After consensus has been achieved, the summary sheet and database should be amended to indicate consensus and the final score.

*Procedures for Achieving Consensus on Discrepant Ratings*

Raters should enter the consensus-making meeting with written ratings rationales for each summary to be discussed.

Similar to the independent rating task, summaries pertaining to specific original text should be evaluated together. The original text and list of elements generated should be reviewed by all raters prior to rating that group of summaries. As consensus is achieved, the consensus scores and rationale are recorded in the space provided in the summary sheet at the bottom of each reviewer's transcription. Ratings achieved by consensus should also be recorded as the final score in the space provided on the summary sheet.

For each summary, raters will take turns explaining their reasons for their particular ratings to the other rater. To assist with decision making, the following decision protocol should be employed, in the order listed:

1. If one rater immediately recognizes the other rater's scores as being more accurate, select that score.
2. Each rater then reviews both rationales against the published criteria (which include both the definitions, central examples, and guidelines) to determine if the rationale corresponds to the score given. If agreement is achieved, record the appropriate rating.
3. If disagreement still exists, the raters may present arguments for their rating, based on their interpretation of the published criteria. If agreement is achieved, record the appropriate rating.
4. All raters review their rationales and confidence ratings.
5. If one rater has a lower confidence rating for his or her score than the other rater, the rater with the lower rating may agree to accept the other rater's score based on the disparity in the confidence ratings. If agreement is achieved, record the appropriate rating.
6. If two of the raters are in agreement, use that classification as the final rating.
7. If there is no agreement between judges, use the middle rating.