A Sequential Sentence Paradigm Using Revised PRESTO Sentence Lists

DOI: 10.3766/jaaa.15074

Andrea R. Plotkowski*
Joshua M. Alexander*

Abstract

Background: Listening in challenging situations requires explicit cognitive resources to decode and process speech. Traditional speech recognition tests are limited in documenting this cognitive effort, which may differ greatly between individuals or listening conditions despite similar scores. A sequential sentence paradigm was designed to be more sensitive to individual differences in demands on verbal processing during speech recognition.

Purpose: The purpose of this study was to establish the feasibility, validity, and equivalency of test materials in the sequential sentence paradigm as well as to evaluate the effects of masker type, signal-to-noise ratio (SNR), and working memory (WM) capacity on performance in the task.

Research Design: Listeners heard a pair of sentences and repeated aloud the second sentence (immediate recall) and then wrote down the first sentence (delayed recall). Sentence lists were from the Perceptually Robust English Sentence Test Open-set (PRESTO) test. In experiment I, listeners completed a traditional speech recognition task. In experiment II, listeners completed the sequential sentence task at one SNR. In experiment III, the masker type (steady noise versus multitalker babble) and SNR were varied to demonstrate the effects of WM as the speech material increased in difficulty.

Study Sample: Young, normal-hearing adults (total n = 53) from the Purdue University community completed one of the three experiments.

Data Collection and Analysis: Keyword scoring of the PRESTO lists was completed for both the immediate- and delayed-recall sentences. The Verbal Letter Monitoring task, a test of WM, was used to separate listeners into a low-WM or high-WM group.

Results: Experiment I indicated that mean recognition on the single-sentence task was highly variable between the original PRESTO lists. Modest rearrangement of the sentences yielded 18 statistically equivalent lists (mean recognition = 65.0%, range = 64.4–65.7%), which were used in the sequential sentence task in experiment II. In the new test paradigm, recognition of the immediate-recall sentences was not statistically different from the single-sentence task, indicating that there were no cognitive load effects from the delayed-recall sentences. Finally, experiment III indicated that multitalker babble was equally detrimental compared to steady-state noise for immediate recall of sentences for both low- and high-WM groups. On the other hand, delayed recall of sentences in multitalker babble was disproportionately more difficult for the low-WM group compared with the high-WM group.

Conclusions: The sequential sentence paradigm is a feasible test format with mostly equivalent lists. Future studies using this paradigm may need to consider individual differences in WM to see the full range of effects across different conditions. Possible applications include testing the efficacy of various signal-processing techniques in clinical populations.

Key Words: adverse listening conditions, speech recognition, working memory

Abbreviations: ANOVA = analysis of variance; ISI = interstimulus interval; PI = performance intensity; PRESTO = Perceptually Robust English Sentence Test Open-set; PRESTO-R = PRESTO revised; SD = standard deviation; SNR = signal-to-noise ratio; SRT = speech reception threshold; WM = working memory

*Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN

Corresponding author: Joshua M. Alexander, Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN 47907; E-mail: alexan14@purdue.edu

INTRODUCTION

The purpose of this study is to extend work by Gilbert et al (2013) and Schurman et al (2014) on the development of a more sensitive test for assessing internal and external factors that may influence individuals' reported difficulties when attempting to comprehend a running stream of speech. Factors internal to listeners include hearing impairment, short-term and working memory (WM), linguistic skills such as vocabulary size, aspects of executive function, and the ability to use knowledge about speaker characteristics (indexical processing) to bootstrap speech recognition (Tamati et al, 2013). Factors external to listeners include signal-to-noise ratio (SNR), type of masker (e.g., steady noise versus modulated noise versus multitalker babble), environmental acoustics such as reverberation, signal processing (e.g., digital noise reduction), and so on.

A limitation of traditional speech recognition tests, in which a listener simply repeats back a word or sentence immediately after hearing it, is that they do not require the depth of processing needed to comprehend, retain, and respond to a spoken message that is necessary for successful communication in everyday listening situations (Fontan et al, 2015). Evidence suggests that traditional speech recognition tests have a modest relationship, at best, with self-report measures of communication difficulties, activity limitations and lifestyle changes. For example, Cox and Alexander (1992) found that <40% of the variance in self-report measures of new hearing aid users was predicted by recognition of sentences mixed with a six-talker babble. Furthermore, Walden and Walden (2004) found that performance on the Quick Speech-in-Noise test (Killion et al, 2004) with and without hearing aids could not predict individuals' scores on self-report measures after the participants' age was parsed out of the correlation.

In addition, a simple percent correct on a repetition task does not take into account the amount of concentration and attention necessary to achieve that level of performance. Due to these limitations, these scores tend to be insensitive to perceptual changes or preferences resulting from signal processing despite listeners reporting noticeable differences in their listening effort and/or their ability to understand speech (Preves, 1990; Keidser, 1996; Ricketts and Hornsby, 2005; Bentler et al, 2008; Sarampalis et al, 2009). On the other hand, tests that require a greater depth of processing whereby listeners must comprehend, make meaningful judgments, or retain information, may be inherently more sensitive to subtle changes in listening demands because of the increased cognitive load (see Picou and Ricketts [2014] for further review). Overall, there is a need for the development of speech tests that are more sensitive to individual differences in speech recognition in adverse, listening conditions and that are more predictive of real-world outcomes.

The Ease of Language Understanding model (Rönnberg et al, 2008; 2013) provides a conceptual framework for how comprehension of a spoken message can become effortful and require greater use of cognitive resources. If the incoming speech is a clear match with representations stored in the lexicon in long-term memory, then the spoken message can be processed implicitly and easily understood. However, if a mismatch occurs due to a degraded signal representation, then additional cognitive resources, such as inference to fill-in gaps and the inhibition of irrelevant information, are explicitly recruited to decode the message. Because WM is required to use contextual cues to fill in missing information and to encode spoken messages into long-term memory, individuals with relatively higher WM capacity should have a better ability to overcome lexical mismatches and to decode speech under poor listening conditions than those with relatively lower WM capacity. Therefore, individual variability in WM capacity appears to play a critical role in the amount of difficulty a listener faces when speech is degraded by background noise, reverberation, and/or hearing impairment and how much these difficulties may be alleviated by signal processing (Akeroyd, 2008; Rönnberg et al, 2008; Francis and Nusbaum, 2009; Rudner et al, 2012; Besser et al, 2013; Zekveld, George, et al, 2013; Sullivan et al, 2015). Understanding these mechanisms and documenting the challenges that individuals face while listening may help inform treatment interventions for individuals with hearing loss. For example, certain hearing technology and features may be better customized by choosing settings that lessen the cognitive demands required to listen in everyday environments and hence influence individual preferences and long-term outcomes (Desjardins and Doherty, 2013; Souza and Sirow, 2014; Arehart et al, 2015; Souza and Arehart, 2015; Souza et al, 2015).

Recognizing the limitations of basic speech recognition tests, Gilbert et al (2013) created the Perceptually Robust English Sentence Test Open-set (PRESTO), a high-variability test comprised of multiple talkers from eight different dialects speaking sentences with varying syntactic structures and with keywords that vary in familiarity and frequency. Previous research has indicated that speech recognition is significantly reduced when the talker's voice changes between test items (Mullennix et al, 1989; Sommers and Barcroft, 2006). Therefore, the intent of PRESTO was to increase listening demands by decreasing the predictability of the speech material in an attempt to capture individual differences in the aforementioned cognitive factors on performance (Magnuson and Nusbaum, 2007). Consistent with this goal, a follow-up study by Tamati et al (2013) found that individual performance on the PRESTO lists

by young, normal-hearing listeners was significantly associated with multiple cognitive measures, especially short-term and WM capacity.

In addition to increasing test sensitivity by manipulating the content of the speech materials, the task of the listener can be manipulated to evaluate the effects of different experimental conditions on cognitive demand. One example is delayed recall, in which listeners are instructed to hold their responses until a set number of trials have been presented. The intent of using this method is to explicitly increase cognitive demands on listeners, especially WM, which is needed to retain and recall the responses to earlier trials. Hence, this method is typically used as a memory span test to quantify the effective use of listeners' WM. When sentences are presented in the auditory modality under different experimental conditions, listeners may be asked to immediately repeat the entire sentence so as to include a classic measure of speech recognition and/or asked to simply make a semantic judgment about it, such as whether it made sense or whether the last word was predictable from the context (Pichora-Fuller et al, 1995). The memory component is then measured by how many of the sentence-final words can be recalled after a set number of sentences.

In a variation of the delayed auditory recall task, Schurman et al (2014) had listeners repeat back not just the sentence-final word, but the entire sentence of the trial 1-back from the current trial (hence, the term "1-back listening task"). After which, a new sentence was presented and the previously current sentence became the 1-back sentence to be repeated, and so on. The researchers' rationale for using this method was to increase WM and cognitive demands in a way that might more closely resemble the demands a listener faces in everyday conversational turn taking. This is because, unlike other manipulations used to increase test difficulty, both tasks were delivered in the auditory modality and both required verbal processing of complete sentences. Schurman et al (2014) used this method to assess the effects of listener age under several different speech-in-noise conditions, and found that older adults performed more poorly on delayed-recall (1-back) sentences compared with younger adults even though SNRs were individually adjusted to levels that produced equivalent performance on immediate-recall (0-back) sentences. Furthermore, they determined that differences between individuals on recognition of the delayed-recall sentences in the different noise conditions were significantly correlated with their performance on the listening span test (Daneman and Carpenter, 1980), a measure of WM. In contrast to the relationship with the results for delayed-recall sentences, performance on the WM test was generally not predictive of individual differences in recognition of the immediate-recall sentences. Together, these findings suggest that adding a second sentence to the traditional speech-in-noise test and requiring listeners to retrieve it from memory may be a more effective task for measuring differences associated with the listening demands of following an ongoing stream of speech.

Using the PRESTO test materials from Gilbert et al (2013) and a delayed-recall task similar to Schurman et al (2014), the current study investigated a novel test paradigm—a variation of the sentence recall task—in which listeners responded to two sequentially presented sentences by opposite gender talkers. The purpose of using the delayed-recall sentence paradigm was the same as Schurman et al (2014), which was to encourage listeners to engage some of the explicit cognitive processes necessary for following an ongoing stream of speech. However, the task demands were different from Schurman et al (2014) because the listeners in the current study first heard a pair of sentences and then wrote down the first sentence in the pair (delayed recall) only after repeating back the second sentence (immediate recall). After listeners provided their written response to the delayed-recall sentence, they were presented with a new pair of sentences, thus they only had to operate on two sentences at a time. In contrast, the method used by Schurman et al (2014) required listeners to operate on three sentences at a time because they heard sentence N, responded to sentence N − 1 that they were holding in memory, heard sentence N + 1, and only then responded to sentence N. In addition to the less demanding pace of the trials, an advantage of the present modification was that a traditional test of speech recognition (immediate, spoken recall) was built into the design and separate blocks of trials for the delayed-recall sentences did not need to be conducted. This helped to minimize unwanted practice effects and/or fatigue effects when comparing performance across the two tasks.

Results of three experiments are reported. The first experiment tested the hypothesis that the PRESTO sentence lists were equivalent in a traditional speech-in-noise test involving immediate recall. This was necessary because when experimental comparisons are made between different sets of speech materials, it is critically important that intelligibility across sets, sentence lists in the current study, be as equivalent as possible. This is particularly true when task demands increase cognitive load, which can further exacerbate differences between lists. Taking the results from the first experiment, sentences were swapped between the original PRESTO lists to create revised, equivalent lists (PRESTO-R), which were then combined into pairs for the sequential sentence paradigm. The second experiment tested the equivalency of speech recognition performance between the new PRESTO-R lists within the immediate- and delayed-recall test conditions and the hypothesis that recognition for the immediate-recall sentences would

be unaffected by inclusion of the delayed-recall sentences. The third experiment manipulated the difficulty of the speech materials in the sequential sentence paradigm by changing the masker type and SNR and compared these effects to a measure of verbal WM.

EXPERIMENT I: PRESTO LIST EQUIVALENCY

Method

Sixteen normal-hearing listeners (5 males, 11 females), aged 18–27 (median = 21 yr), were tested on all the sentences from the PRESTO test (Gilbert et al, 2013). Listeners passed a hearing screening at 15 dB HL at the octave frequencies from 0.25 to 8 kHz in a sound-treated booth. Sentences for PRESTO were taken from the Texas Instruments/Massachusetts Institute of Technology (Garofolo et al, 1993) speech corpus and were compiled into 20 lists (with list 1 designated as practice) of 18 sentences from different talkers, including nine male and nine female talkers with a variety of dialects (see Gilbert et al, 2013, for more detail about the test construction). PRESTO list 1 was always presented first to help familiarize listeners with the materials and task. The remaining 342 sentences were then presented in a computer-generated random order, which varied for each listener. Here and throughout, stimuli were presented at their native sampling rate (16 kHz) monaurally using circumaural BeyerDynamic DT150 headphones (Heilbronn, Germany) at 65 dB SPL in a sound-treated booth. All listeners were tested using their right ears, except two in this experiment and one in experiment III who were tested using their left ears because they had an established mild high-frequency hearing loss in their right ears. Sentences were presented at −3 dB SNR using a steady, speech-shaped noise. This SNR was chosen based on the pilot data which indicated that it would produce scores that were not too close to the ceiling of the performance-intensity (PI) function for the immediate-recall sentences or too close to the floor for the delayed-recall sentences. Listeners were instructed to repeat back each sentence aloud immediately, which was scored by an experimenter who also controlled the rate of trial presentations. Keyword scoring was used following rules outlined in the PRESTO manual. Each list contained 76 keywords, except the practice list (list 1) which had 75 keywords. Mean percent correct was calculated separately for sentences comprising each of the original lists.

Results

Figure 1 shows the results for the PRESTO lists, which are rank ordered along the abscissa from lowest to highest percent correct. Here and throughout, error bars correspond to the standard error of the mean. Mean recognition for the practice list (dark gray bar) was the lowest at 48.42%, while the test lists (light gray bars) had a mean recognition of 65.08% with a range of 55.63–73.60% (standard deviation [SD] = 4.97%). Here, and throughout, all statistical analyses were conducted with percent correct transformed to rationalized arcsine units (Studebaker, 1985) to normalize

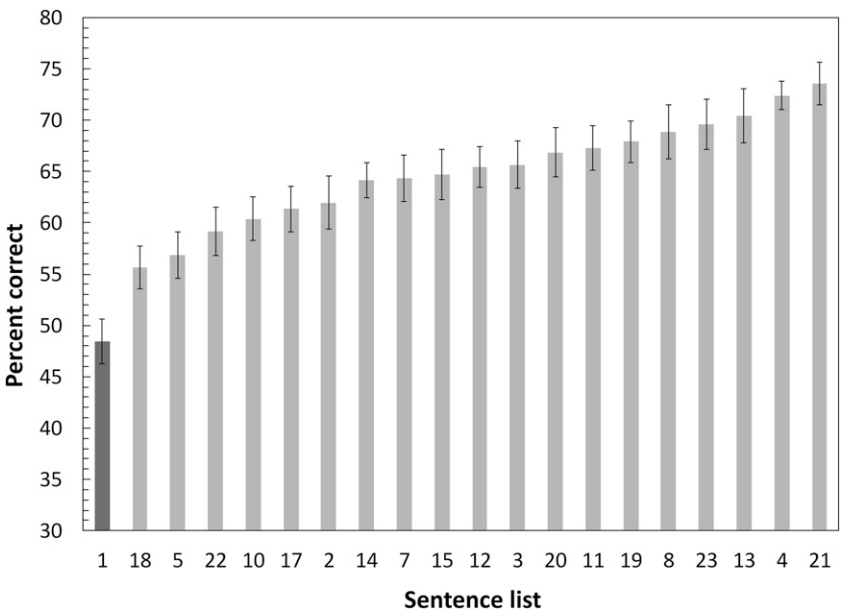

Figure 1. Mean percent correct recognition for each of the original PRESTO lists using a traditional speech recognition task. Lists are arranged from lowest to highest. List 1 is the practice list and is depicted with darker shading. Error bars represent the standard error of the mean.

the variance across the range and all post hoc testing of paired comparisons was completed using paired *t* tests and the Benjamini-Hochberg false discovery rate correction for multiple comparisons (Benjamini and Hochberg, 1995). A within-subjects analysis of variance (ANOVA), excluding the practice list, revealed a significant main effect [$F_{(18,270)} = 16.9$, $p < 0.001$], indicating that speech recognition was not equivalent across the 19 test lists. The results of post hoc tests from all pairwise list comparisons are shown in Table 1, with statistically equivalent list pairs in boldface. Of the 171 list pairs, only ~38% were equivalent.

Discussion

The mean recognition of 65.08% across the test lists is located near the middle of the underlying PI function, which is desirable to avoid floor and ceiling performance effects in future applications. This range of performance was similar, but slightly higher than that obtained by a similar population at −3 dB SNR by Gilbert et al (2013), which was 55.00% (SD = 7.02). However, the masker used by the two studies was different. The present study used steady noise rather than multitalker babble, which may have decreased the demands on WM due to the removal of the informational masking (Schneider et al, 2007; Zekveld, Rudner, et al, 2011; Koelewijn et al, 2014) and could explain the difference in recognition (cf. experiment III).

Gilbert et al (2013) administered multiple lists across a range of SNRs to each listener during two sessions and determined that the overall performance at each SNR was strongly correlated between test and retest. However, because only three to four of the lists were presented at each SNR for each listener, list equivalency could not be established. Therefore, a follow-up study by Faulkner et al (2015) evaluated the equivalency of subsets of PRESTO lists using normal-hearing listeners. The 19 test lists were divided into two subsets of ten (with list 17 common to both) which were presented in multitalker babble at 0 dB SNR to two groups of listeners. List equivalency was determined by an evaluation of overall percent correct as well as by the correlation of individual performance across lists. The authors reported that lists 5 and 18 had the lowest mean recognition scores for each subset while lists 13 and 21 had the highest mean recognition scores. These results are consistent with the findings from the present study because lists 5 and 18 also had the lowest mean recognition scores and lists 21 and 13 had the first and third highest mean recognition scores, respectively. The results for the lists with the lowest and highest recognition scores from the present study also compared favorably to another study by Faulkner et al (2015). These authors tested a different group of normal-hearing listeners who identified a subset of the PRESTO lists that had been processed without noise using an eight-channel vocoder to simulate some of the spectral degradation associated with cochlear implant signal processing. Therefore, there is high agreement on the end points of the performance continuum between Faulkner et al (2015) and the present study despite differences in masker type, SNR, signal processing, and the fact that listeners in the present study listened monaurally and

Table 1. Type I Error Probability (*p* Values) Following Correction for Multiple Comparisons Using the Benjamini-Hochberg (Benjamini and Hochberg, 1995) False Discovery Rate for Each List Pair of the PRESTO Speech Intelligibly Test

	18	5	22	10	17	22	14	7	15	12	3	20	11	19	8	23	13	4
5	**0.609**																	
22	**0.105**	**0.235**																
10	**0.080**	**0.072**	**0.515**															
17	0.022	0.001	**0.273**	**0.564**														
22	0.011	0.008	**0.090**	**0.431**	**0.628**													
14	0.000	0.004	0.019	**0.068**	**0.180**	**0.353**												
7	0.004	0.001	0.004	0.034	**0.063**	**0.120**	**0.818**											
15	0.000	0.000	0.004	0.041	0.033	**0.078**	**0.656**	**0.833**										
12	0.000	0.000	0.008	0.010	0.039	**0.151**	**0.493**	**0.656**	**0.651**									
3	0.001	0.000	0.012	0.007	0.001	**0.079**	**0.477**	**0.548**	**0.567**	**0.877**								
20	0.001	0.000	0.010	0.006	0.001	0.036	**0.199**	**0.258**	**0.301**	**0.527**	**0.313**							
11	0.000	0.000	0.000	0.002	0.000	0.006	**0.077**	0.043	**0.096**	**0.397**	**0.303**	**0.811**						
19	0.000	0.000	0.000	0.002	0.000	0.005	**0.068**	**0.052**	**0.105**	**0.303**	**0.179**	**0.564**	**0.729**					
8	0.000	0.000	0.000	0.001	0.002	0.004	0.034	0.021	0.012	**0.102**	**0.158**	**0.477**	**0.373**	**0.609**				
23	0.000	0.000	0.000	0.002	0.001	0.001	0.031	0.039	0.008	0.033	0.034	**0.273**	**0.357**	**0.440**	**0.720**			
13	0.000	0.000	0.001	0.000	0.001	0.004	0.008	0.024	0.003	0.004	0.009	**0.104**	**0.223**	**0.275**	**0.516**	**0.644**		
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.002	0.002	0.004	0.039	0.010	0.026	**0.151**	**0.275**	**0.484**	
21	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.004	0.004	0.022	0.023	**0.096**	**0.428**

Note: Lists are arranged in order according to mean intelligibility from lowest to highest. Entries for equivalent list pairs ($p > 0.05$) are in bold.

spoke their responses, while listeners in the Faulkner et al (2015) study listened binaurally and typed their responses.

Using their criteria, Faulkner et al (2015) determined that ten of the lists with multitalker babble and nine of the lists with vocoding were equivalent, with six lists common to both. In the present study, in which each listener identified the entire PRESTO corpus, different pairs of lists were found to be equivalent to each other, but because the range of mean performance on the lists was ~18%, without modification it would be difficult to find a mutually inclusive group of equivalent lists larger than that found by Faulkner et al (2015). If these lists are to be used to make within-subject or between-subject comparisons across a variety of conditions, list equivalency needs to be improved because list variations may mask more subtle differences attributed to various experimental manipulations and/or between-subjects factors.

CREATION OF PAIRED PRESTO-R LISTS

Due the large variation in mean recognition between the original PRESTO lists, the lists were revised to improve equivalency. Based on the mean recognition scores for each sentence, low- and high-recognition sentences were swapped between lists while controlling for talker gender and the number of keywords to produce 18 test lists of 18 sentences with a mean predicted recognition of 64.99% (SD = 0.36%, range = 64.43–65.71%) and two practice lists with predicted recognition of 48.42% and 66.74%, respectively. Using the individual sentence data from experiment I, a within-subjects ANOVA on the 18 new test lists indicated that predicted speech recognition was equivalent across lists [$F_{(17,255)} = 0.1$, $p = 1.0$].

The newly created lists maintain most of the variability in syntax and speaker characteristics from the original lists; however, the integrity of the original PRESTO lists, which were balanced in terms of dialect and keyword frequency and familiarity, cannot be guaranteed. Therefore, to avoid confusion with the original PRESTO lists, the revised lists are identified as PRESTO-R and labeled using letters instead of numbers, with lists A and B reserved for practice.

Following the creation of the PRESTO-R sentence lists, 10 paired lists with 18 sentence pairs in each were created. Sentences from the first list in each pair were used for immediate-recall (lists A, C, E, G, I, K, M, O, Q, S) and sentences from the second list were used for delayed-recall (lists B, D, F, H, J, L, N, P, R, T). First, the sentences in each list were divided by talker gender and then sorted by total number of keywords. Next, each sentence was paired with a sentence from the other list with the opposite gender talker and the reverse ranking of keywords to control for the total number of keywords in each pair. The majority of sentence pairs (>82%) had eight or nine total keywords (42.2% and 40.0%, respectively) and the remaining pairs had seven or ten total keywords (10.6% and 7.2%, respectively). The sentence pairs were also checked for semantic similarities to prevent two similar topics from being presented sequentially. The revised PRESTO lists (PRESTO-R) are provided in Supplemental Appendix S1, supplemental to the online version of this article.

As a practical matter, to reduce confusion about task demands, the response to the immediate-recall sentences was spoken and the response to the delayed-recall sentences was written. The spoken response was scored immediately by an experimenter who was in the booth facing the listener while the written response was scored at a later time. While having to remember two sentences from different talkers would have increased listeners' cognitive load and tapped into the same pool of cognitive resources (Wong et al, 2004), the effect of adding a second sentence on speech recognition was predicted to be different for the two response formats. As reviewed by Gosselin and Gagné (2010), when listeners are given a secondary task to complete in addition to a primary listening task, they can prioritize completing the primary task over the secondary task based on the experimental instructions such that they will allocate whatever cognitive resources they have to the primary task before allocating them to the secondary task (see also, Picou and Ricketts, 2014). Along these lines, because of (a) the minimal demand on WM, (b) the naturalness of the response format, and (c) the social demand of having an authoritative figure actively judging the spoken response, it was expected that listeners would implicitly prioritize remembering the immediate-recall sentence over the delayed-recall sentence. Therefore, it was hypothesized that relative to recognition of the single sentences in experiment I, recognition of the immediate-recall sentences in experiment II would be minimally affected while recognition of the delayed-recall sentences would be degraded. More specifically, recognition of the delayed-recall sentences was predicted to be dependent on the amount of cognitive resources required to extract and repeat back the content of the immediate-recall sentences, thereby, creating a metric that could be used to test individual differences in WM and changes in cognitive demands associated with different experimental manipulations.

EXPERIMENT II: PAIRED PRESTO-R LIST EQUIVALENCY IN SEQUENTIAL SENTENCE PARADIGM

Method

To test the equivalency of the PRESTO-R lists and the hypothesis that recognition for the immediate-recall

sentences would be unaffected by inclusion of the delayed-recall sentences, a different group of 16 normal-hearing listeners (4 males, 12 females), aged 18–29 (median = 20 yr) participated in the sequential sentence paradigm. All listeners completed a hearing screening as previously described. All of the sentence pairs were combined and presented in two random orders across listeners, one being the reverse order of the other, excluding the practice list pair which was always presented first so that listeners would be familiarized with the test materials and task. The sentences for each pair were presented sequentially with a 500-msec pause at a 65-dB SPL presentation level with an SNR of −3 dB using a steady speech-shaped noise.

Listeners were instructed to pay careful attention to both sentences and to immediately repeat aloud the second sentence before writing down the first sentence on the numbered answer sheet that was provided. Listeners were encouraged to provide a response for each sentence, even if the response was a guess with only a few words. If a listener began repeating the first sentence instead of the second sentence during the first practice list, they were reinstructed. After this reinstruction, all listeners correctly completed the task. The experimenter immediately scored the spoken responses and ensured that the listener did not begin writing until after the spoken sentence had been completed. The written responses were scored at a later time. Two raters scored each written response, with a third rater making a final score determination if any discrepancies occurred. The mean interrater reliability across listeners was $r_{(179)} = 0.994$ (SD = 0.0044).

Results

The gray bars in Figure 2 display the mean recognition for the immediate-recall sentence lists in experiment II. As predicted during the creation of the lists, recognition was equivalent between each pair of test lists (lists C, E, G, I, K, M, O, Q, S) as indicated by a within-subjects ANOVA [$F_{(8,120)} = 1.8$, $p > 0.05$]. Mean recognition across the immediate-recall lists in the sequential sentence paradigm was 67.39% (SD = 1.76%), which was statistically the same as the mean recognition predicted from the single-sentence paradigm (experiment I) [$F_{(1,30)} = 0.1$, $p > 0.05$], indicating that adding the delayed-recall sentences did not increase the WM load for the immediate-recall sentences.

The white bars in Figure 2 display the mean recognition for the delayed-recall sentence lists in experiment II. Mean recognition across the test lists (lists D, F, H, J, L, N, P, R, T) was 44.97% (SD = 2.49%), which as predicted was much lower than for the immediate-recall sentences. A within-subjects ANOVA indicated that, unlike the immediate-recall sentences, performance for the delayed-recall sentences was not equivalent between lists [$F_{(8,120)} = 3.35$, $p < 0.01$]. Post hoc tests indicated that the source of significant main effect was list D, which was significantly higher than lists N and T ($p < 0.05$).

Discussion

By comparing performance both in a traditional single-sentence paradigm and in the sequential sentence paradigm, it was demonstrated that performance on the

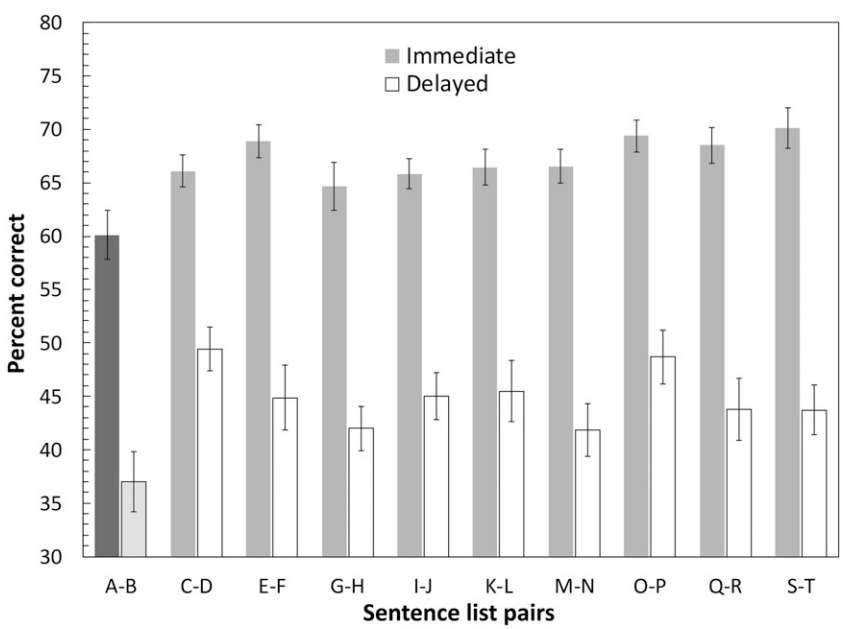

Figure 2. Mean percent correct recognition for the PRESTO-R lists with the immediate-recall (spoken) response format (gray bars) and the delayed-recall (written) response format (white bars). Lists are grouped according to the pairs that were presented to the listeners. The first list pair was used for practice and is depicted with darker shading. Error bars represent the standard error of the mean.

immediate-recall sentences was not affected by the addition of the delayed-recall sentences. As discussed in section "Creation of Paired PRESTO-R Lists", this likely occurred because of various factors that caused listeners to implicitly assign the immediate-recall sentences as the primary task and the delayed-recall sentences as the secondary task and to allocate cognitive resources accordingly. It was also demonstrated that when tested in an immediate-recall format, the PRESTO-R lists were equivalent to each other when presented in the sequential sentence paradigm. On the other hand, there was more variation in performance between lists using the delayed-response format and not all of the lists achieved statistical equivalency, which may be attributed to some sentence pairings being more difficult or easier than others. The written modality associated with the delayed-recall sentences may have also added the difficulty of trying to spell the words, and some lists may have been more challenging in this aspect even though listeners were told spelling was not scored. It may be partially for this reason that recognition for the delayed-recall sentences was lower compared with the immediate-recall sentences; however, it is largely assumed that this component of the task was more sensitive to cognitive processing demands and reflected restraints on WM. This assumption was tested in the following experiment, which manipulated the difficulty of the speech materials by changing the masker type and SNR and compared the effects of these changes on speech recognition to a measure of verbal WM.

It was hypothesized that speech recognition would be lower in the presence of a multitalker babble masker relative to a steady noise masker due to the increased demands on WM associated with inhibiting conflicting speech information from the masker (Schneider et al, 2007; Zekveld, Rudner, et al, 2011; Koelewijn et al, 2014). The effect of masker (noise versus multitalker babble) was predicted to interact with the effect of recall (immediate versus delayed), such that the predicted decrease in recognition for the multitalker babble masker (i.e., informational masking) would be greater for the delayed-recall task because of the associated increase in demands on WM. Finally, it was hypothesized that the effects of masker and recall would interact with WM, where the predicted decrements in recognition associated with the multitalker babble masker and with the delayed-recall task would be greater for listeners with relatively low WM compared to those with relatively high WM.

EXPERIMENT III: EFFECTS OF WM, MASKER TYPE, AND SNR

Method

The effects of masker type and SNR were tested using a different group of 21 normal-hearing listeners (8 males, 13 females), aged 18–32 (median = 20 yr). All listeners completed a hearing screening as previously described. The same pairs of lists as in experiment II were presented at −5, 0, 5, and 10 dB SNR with two masker types: steady, speech-shaped noise and multitalker babble. The multitalker babble consisted of six talkers (three males, three females) with standard Midwestern dialects who were recorded while reading 5 min of prose from different Encyclopedia passages. Each passage was recorded in a mono channel at a 44.1-kHz sampling rate using a headworn Shure Beta 53 professional subminiature condenser omnidirectional microphone with foam windscreen and filtered protective cap (Shure, Niles, IL). The published frequency response of this microphone arrangement is flat within 3 dB up to 10 kHz. Master recordings were downsampled at 22.05 kHz for further processing. Silent intervals >250 msec were shortened. Recordings derived from each talker were then band-pass filtered and spectrally shaped to approximate the one-third octave band levels from the International Long-Term Average Speech Spectrum (Byrne et al, 1994). Using stored random seeds, random selections derived from each talker with durations 500 msec longer than the target sentence were mixed together to form a six-talker masker. The masker started 250 msec before and ended 250 msec after the target sentence. The unique masker generated for each sentence did not change across SNR or masker type. The steady, speech-shaped noise masker for each sentence was produced using the same random selections, the only difference being that the temporal envelope was flattened by randomizing the phase of a fast Fourier transform of the masker followed by inverse fast Fourier transform.

To give listeners exposure to the different conditions and to the task itself, they responded to one list pair for practice in which the masker type alternated between presentations and SNR systematically decreased, followed by another list pair in which the masker type and SNR varied randomly. The masker condition was the same for both the immediate-recall and delayed-recall sentences in a pair, with the masker type and SNR applied to each list pair being approximately counterbalanced across listeners using one of eight pregenerated presentation orders (five orders were presented to three listeners each and three orders were presented to two listeners each). In the pregenerated test orders, two list pairs were assigned to each of the eight masker conditions. Sentences from each list pair were combined and randomized so that listeners heard a constantly changing masker level and type.

Consistent with the previous methods, sentences were presented monaurally at 65 dB SPL. Listeners were provided the same instructions as in experiment

II and were reinstructed as necessary. Again, the experimenter immediately scored the spoken responses and ensured that the listener did not begin writing until after the spoken sentence had been completed. The written responses were scored at a later time. The mean interrater reliability across listeners was $r_{(179)} = 0.991$ (SD = 0.0073).

Since performance on the sequential sentence paradigm was predicted to be related to individual differences in cognitive processing, listeners also completed the Visual Letter Monitoring test, a measure of WM (Gatehouse et al, 2003). In this test, letters were presented in large font one at a time on a touch-screen monitor. Letter sequences were the same for all listeners. Listeners were instructed to monitor the letters and to respond when three consecutive letters spelled a word. Responses were made by pressing a button on the touch-screen labeled "Yes, this makes a word." Listeners were tested on three different letter sequences, each lasting 2–5 min: a practice run of 31 letters with a 3-sec interstimulus interval (ISI), followed by a "slow" test run consisting of a 151-letter sequence with a 2-sec ISI, followed by a "fast" test run consisting of a different 151-letter sequence with a 1-sec ISI. The performance measure consisted of d' averaged across the two test runs (Macmillan and Creelman, 2005): $d'_{\text{avg}} = \sqrt{d'_{\text{slow}}{}^2 + d'_{\text{fast}}{}^2}$. A k-means cluster analysis (Arthur and Vassilvitskii, 2007) of the Visual Letter Monitoring scores was used to separate listeners into high (n = 10: 3 males, 7 females; mean age = 21.8 yr, SD = 3.6) and low (n = 11: 5 males, 6 females; mean age = 20.6 yr, SD = 1.7) WM groups.

Results

Figures 3A and B show the PI functions for the immediate- and delayed-recall sentences, respectively. Black lines with circle symbols and gray lines with triangle symbols correspond to the high- and low-WM groups, respectively, and solid and dotted lines correspond to the noise and babble maskers, respectively. The most apparent effects were improvements in sentence recognition with increasing SNR, with immediate-recall versus delayed-recall sentences, and with noise versus multitalker babble maskers.

For the purpose of data reduction, the PI functions for each listener and condition were converted into a speech reception threshold (SRT) corresponding to the SNR that would yield ~50% correct. This was accomplished by linearly regressing the Z scores corresponding to the proportions correct against the four SNRs and using the intercept of the fitted functions (Z score = 0, percent correct = 50%) for the SRT. SRTs for each condition are shown in Figure 4 separately for the high- and low-WM groups (black and gray bars, respectively). Lower SRTs indicate better performance. A mixed-design omnibus ANOVA was conducted using masker (noise versus multitalker babble) and recall (immediate versus delayed) as the within-subjects factors and WM (high versus low) as the between-participants factor. Consistent with the observations from Figure 3, there were significant main effects for masker [$F_{(1,19)} = 152.8$, $p < 0.001$] and recall [$F_{(1,19)} = 91.1$, $p < 0.001$]. There was also a significant three-way interaction of these factors with WM [$F_{(1,19)} = 5.3$, $p < 0.05$]. To explore the interaction, the difference in SRT between the babble and noise maskers was

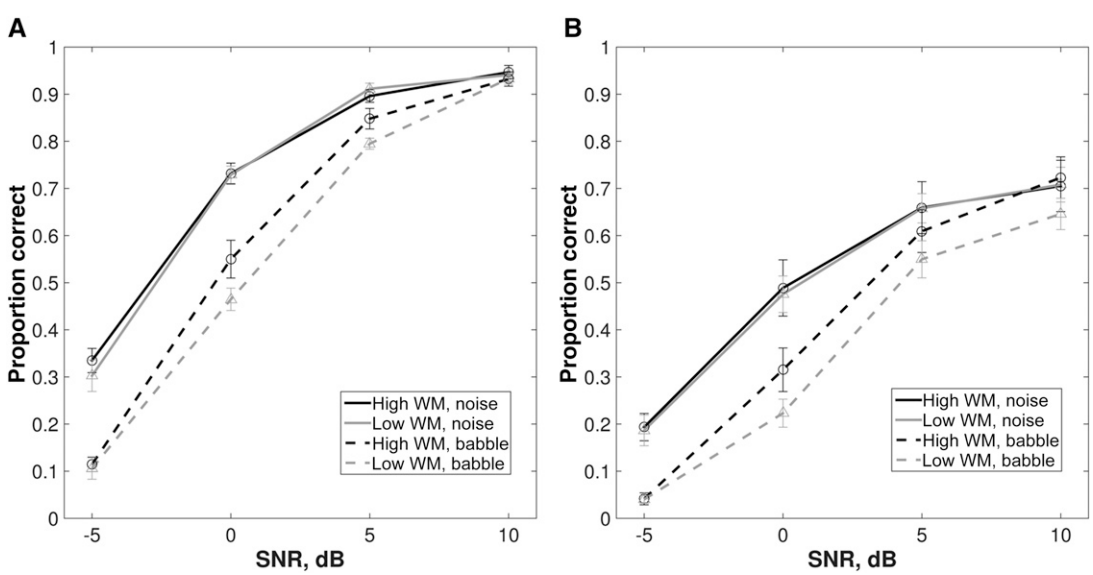

Figure 3. Results for experiment III in which masker type and SNR were varied. Mean results for the (A) immediate- and (B) delayed-recall sentences are shown. Black lines with circle symbols and gray lines with triangle symbols represent the high- and low-WM group noise, respectively. Solid and dotted lines represent the noise and multitalker babble maskers, respectively. Error bars represent the standard error of the mean.

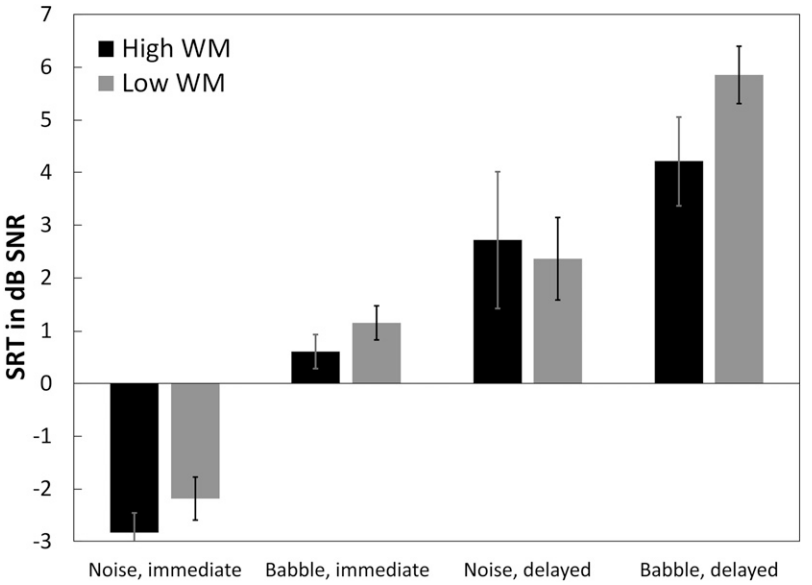

Figure 4. Plotted are the mean SRTs (lower equals better) for the two types of sentences (immediate recall and delayed recall) and the two types of maskers (noise and multitalker babble) which were derived by estimating the SNR corresponding to 50% correct for each listener and condition. The mean of high-WM group is plotted in black and the mean of the low-WM group is plotted in gray. Error bars represent the standard error of the mean.

computed as an index of "informational masking" (Agus et al, 2009). A two-sample t test showed that there was no statistical difference [$t_{(19)} < 1.0$, $p > 0.05$] in informational masking between the high- and low-WM groups for the immediate-recall sentences (Figure 3A), indicating that the negative effect of babble on speech recognition was equal for both groups. On the other hand, for the delayed-recall sentences (Figure 3B), the low-WM group had significantly greater informational masking [$t_{(19)} = 2.4$, $p < 0.05$] compared with the high-WM group. This result indicates that speech recognition deficits attributed to interference from the linguistic content of the babble masker were most apparent for listeners with lower WM under conditions that required successful storage and retrieval of sentences.

Discussion

The main effect for masker in the immediate-recall condition, in which recognition was significantly poorer for the sentences presented in six-talker babble compared with noise, is consistent with the previous findings (Carhart et al, 1975; Freyman et al, 2004; Simpson and Cooke, 2005; Garcia Lecumberri and Cooke, 2006; Van Engen and Bradlow, 2007; Cooke et al, 2008; Koelewijn et al, 2014) and is indicative of limitations in encoding the target speech while inhibiting the informational content of the masker. Furthermore, the immediate-recall task was not sensitive to individual differences in WM since the informational component associated with multitalker babble was equally detrimental to both WM groups. This result may be somewhat at odds with the assumption that compared with steady-state maskers, babble maskers yield lower sentence recognition because they require WM to inhibit competition of linguistic information from the competing speech (Schneider et al, 2007; Zekveld, Rudner, et al, 2011; Koelewijn et al, 2014). One explanation for this apparent discrepancy is that immediate-recall may not be sufficiently demanding of WM in young, normal-hearing listeners and thus it may be less sensitive to individual differences in WM ability. For example, Zekveld, Rudner, et al (2011) found that sentence recognition by young, normal-hearing listeners was correlated with a measure of WM only for conditions where intelligibility was intentionally made low by decreasing the SNR and by using misleading contextual cues about the sentence content. Based on the findings that indicate that middle-aged, normal-hearing adults perform more poorly on WM tasks compared with younger adults (Jaeggi et al, 2009; Neidleman et al, 2015), it is hypothesized that larger effects of WM on sentence recognition in babble versus noise will be observed for middle-aged and older listeners.

In the present experiment, mean sentence recognition in the immediate-recall condition for babble was lower toward the middle of the PI functions for the low-WM group compared with the high-WM group, which suggests that there may have been a role for WM that was too small to be detected statistically. Results from 40 young, normal-hearing listeners who were tested on PRESTO by Tamati et al (2013) corroborate the observed difference between WM groups in the present experiment because they found that listeners in the upper quartile on the PRESTO test had statistically

higher indices of WM compared with listeners in the lower quartile.

Recognition was significantly poorer for both groups of listeners and both types of maskers in the delayed-recall condition compared with the immediate-recall condition, which is indicative of limitations in WM and retrieval, and not in encoding, because the lists for each recall condition were previously equated for difficulty. Consistent with this assumption, some listeners reported that they heard the first sentence clearly but then forgot what to write down shortly after they repeated back the second sentence. The role of WM and retrieval in this task is further supported by the significantly greater amount of informational masking (difference in SRT for the babble and noise maskers) for the low-WM group compared with the high-WM group in the delayed-recall condition. That is, the significant difference in this condition between WM groups indicates that WM plays an important role in the ability to respond accurately to the delayed-recall sentences in the multitalker babble. Thus, while the babble may have been more effortful and cognitively demanding compared with noise maskers in both recall conditions, the low-WM group may have had fewer cognitive resources and WM skills available to devote to speech recognition when they had to retrieve the sentences from memory in the delayed-recall condition.

GENERAL DISCUSSION

This study provided list equivalency and normative data for the PRESTO test materials (Gilbert et al, 2013) in a single-sentence recognition task and in a sequential sentence recognition task. PRESTO sentence materials were used because they are inherently more difficult than traditional sentence recognition materials due to their high variability and low predictability. Furthermore, similar to previous delayed-recall sentence tests (Schurman et al, 2014), the sequential sentence paradigm was used as an additional means of increasing the cognitive demands placed on listeners compared with traditional speech recognition tests. Therefore, the purpose of combining the delayed-recall paradigm with the PRESTO sentence materials was to generate a more sensitive measure of speech communication difficulties in real-world situations that require constant attention to an ongoing message.

The sequential sentence paradigm was used as a research tool for investigating cognitive processes that may influence multiple facets of everyday communication such as comprehension, listening effort, ease of understanding, and listening fatigue. These facets can be evaluated to a limited extent by electrophysiological measures and by other objective laboratory-based speech measures (Hicks and Tharpe, 2002; Wong et al, 2004; Mackersie and Cones, 2011; Zekveld, Kramer, et al, 2011; Hornsby, 2013; Zekveld, Festen, et al, 2013; Picou and Ricketts, 2014; Fontan et al, 2015; Mackersie et al, 2015; Sullivan et al, 2015), but they are most often and best evaluated using self-report outcome measures of communication difficulties, activity limitations, lifestyle changes, and so on (Newman et al, 1991; Cox and Alexander, 1995; Dillon et al, 1997; Gatehouse, 1999; Cox et al, 2003; Gatehouse and Noble, 2004). Therefore, while the sequential sentence paradigm was motivated by the inability of traditional speech tests to capture these real-world consequences for communication, it too needs to be validated as a proxy of outcomes associated with the provision of hearing aids, cochlear implants, etc. Tamati et al (2013) provides encouraging evidence for the clinical utility of the sequential sentence paradigm because among their young, normal-hearing listeners they found that those who scored in the lower quartile on the PRESTO test also reported having difficulty understanding speakers at a public meeting or religious service at a significantly higher rate than those who scored in the upper quartile. While these two groups of listeners did not differ statistically on their responses to the other test items on the self-report measure, higher test sensitivity for PRESTO is expected when cognitive demand is increased as when testing a population of hearing-impaired individuals and/or when using the sequential sentence paradigm.

The experiments reported here demonstrated that the sequential sentence paradigm using PRESTO-R sentences in a mixed response format (spoken and written) is a feasible task for young, normal-hearing listeners. Research is underway to demonstrate the usefulness of the sequential sentence paradigm at a variety of SNRs for documenting communication difficulty in real-world environments by clinical populations (e.g., individuals with hearing loss and older adults) whose ability to comprehend ongoing speech depends greatly on verbal WM and other cognitive processes (Desjardins and Doherty, 2013; Souza and Sirow, 2014; Arehart et al, 2015; Souza and Arehart, 2015; Souza et al, 2015). Furthermore, feasibility of the paradigm was also indicated by the mean recognition scores for the delayed-recall sentences, which neither were at floor performance nor were they too low to make meaningful associations with individual differences when the most difficult babble masker was used. Because the delayed-recall task leads to lower levels of performance, it inherently leaves more room to demonstrate differences in cognitive effort. Therefore, this property of the sequential sentence task may help to elucidate differences between two signal manipulations or SNRs that otherwise produce near-plateau levels of performance on traditional tests of speech recognition (Preves, 1990; Keidser, 1996; Ricketts and Hornsby, 2005; Bentler et al, 2008; Sarampalis et al, 2009). In addition, immediate recall of sentences was not degraded when listeners were given an earlier sentence to remember. This finding is consistent

with evidence that shows that with explicit or implicit instruction, listeners are able to prioritize completing the primary listening task over a secondary task by allocating cognitive resources accordingly (Gosselin and Gagné, 2010; Picou and Ricketts, 2014).

It was assumed that due to the additional load on WM, recognition of the delayed-recall sentences would be more sensitive to various external factors that alter the difficulty of the speech material and to various internal factors related to the encoding and retrieval of this information. It is well documented (Carhart et al, 1975; Freyman et al, 2004; Simpson and Cooke, 2005; Garcia Lecumberri and Cooke, 2006; Van Engen and Bradlow, 2007; Cooke et al, 2008; Koelewijn et al, 2014) that compared with steady noise, multitalker babble creates additional "informational masking" because the informational content of the masker interferes with the lexical processing of the target speech. Other researchers (Schneider et al, 2007; Zekveld, Rudner, et al, 2011) have hypothesized that individual differences in speech-on-speech masking arise from differences in WM, which is used to inhibit competition of linguistic information from the competing speech. The significant effect of masker (babble worse than noise) replicates these previous findings. Furthermore, the role of WM in both the delayed-recall task and in the recognition of speech in multitalker babble is supported by a three-way interaction whereby the low-WM group demonstrated significantly more informational masking than the high-WM group in the delayed-recall condition, but not in the immediate-recall condition. Therefore, future studies using this paradigm may need to consider individual differences in WM to see the full range of effects across conditions.

Acknowledgments. The authors thank Kathleen F. Faulkner and David B. Pisoni for the provision of the PRESTO materials and for their engaging discussions about this work. The authors also thank two anonymous reviewers for their useful suggestions on an earlier version of this article.

REFERENCES

Agus TR, Akeroyd MA, Gatehouse S, Warden D. (2009) Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise. *J Acoust Soc Am* 126(4): 1926–1940.

Akeroyd MA. (2008) Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol* 47(2, Suppl):S53–S71.

Arehart K, Souza P, Kates J, Lunner T, Pedersen MS. (2015) Relationship among signal fidelity, hearing loss, and working memory for digital noise suppression. *Ear Hear* 36(5):505–516.

Arthur D, Vassilvitskii S. (2007) K-means++: the advantages of careful seeding. *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57:289–300.

Bentler R, Wu YH, Kettel J, Hurtig R. (2008) Digital noise reduction: outcomes from laboratory and field studies. *Int J Audiol* 47(8): 447–460.

Besser J, Koelewijn T, Zekveld AA, Kramer SE, Festen JM. (2013) How linguistic closure and verbal working memory relate to speech recognition in noise—a review. *Trends Amplif* 17(2):75–93.

Byrne D, Dillon H, Tran K, et al. (1994) An international comparison of long-term average speech spectra. *J Acoust Soc Am* 96: 2108–2120.

Carhart R, Johnson C, Goodman J. (1975) Perceptual masking of spondees by combinations of talkers. *J Acoust Soc Am* 58(1, Suppl):S35.

Cooke M, Garcia Lecumberri ML, Barker J. (2008) The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *J Acoust Soc Am* 123(1):414–427.

Cox RM, Alexander GC. (1992) Maturation of hearing aid benefit: objective and subjective measurements. *Ear Hear* 13(3):131–141.

Cox RM, Alexander GC. (1995) The abbreviated profile of hearing aid benefit. *Ear Hear* 16(2):176–186.

Cox RM, Alexander GC, Beyer CM. (2003) Norms for the international outcome inventory for hearing aids. *J Am Acad Audiol* 14(8): 403–413.

Daneman M, Carpenter PA. (1980) Individual differences in working memory and reading. *J Verbal Learn Verbal Behav* 19:450–466.

Desjardins JL, Doherty KA. (2013) Age-related changes in listening effort for various types of masker noises. *Ear Hear* 34(3): 261–272.

Dillon H, James A, Ginis J. (1997) Client Oriented Scale of Improvement (COSI) and its relationship to several other measures of benefit and satisfaction provided by hearing aids. *J Am Acad Audiol* 8(1):27–43.

Faulkner KF, Tamati TN, Gilbert JL, Pisoni DB. (2015) List equivalency of PRESTO for the evaluation of speech recognition. *J Am Acad Audiol* 26(6):582–594.

Fontan L, Tardieu J, Gaillard P, Woisard V, Ruiz R. (2015) Relationship between speech intelligibility and speech comprehension in babble noise. *J Speech Lang Hear Res* 58(3):977–986.

Francis AL, Nusbaum HC. (2009) Effects of intelligibility on working memory demand for speech perception. *Atten Percept Psychophys* 71(6):1360–1374.

Freyman RL, Balakrishnan U, Helfer KS. (2004) Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J Acoust Soc Am* 115(5 Pt 1): 2246–2256.

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL. (1993) *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia: Linguistic Data Consortium.

Garcia Lecumberri ML, Cooke M. (2006) Effect of masker type on native and non-native consonant perception in noise. *J Acoust Soc Am* 119(4):2445–2454.

Gatehouse S. (1999) Glasgow Hearing Aid Benefit Profile: derivation and validation of a client-centered outcome measure for hearing aid services. *J Am Acad Audiol* 10:80–103.

Gatehouse S, Naylor G, Elberling C. (2003) Benefits from hearing aids in relation to the interaction between the user and the environment. *Int J Audiol* 42(1, Suppl):S77–S85.

Gatehouse S, Noble W. (2004) The Speech, Spatial and Qualities of Hearing Scale (SSQ). *Int J Audiol* 43(2):85–99.

Gilbert JL, Tamati TN, Pisoni DB. (2013) Development, reliability, and validity of PRESTO: a new high-variability sentence recognition test. *J Am Acad Audiol* 24(1):26–36.

Gosselin PA, Gagné J-P. (2010) Use of a dual-task paradigm to measure listening effort. *Can J Speech Lang Pathol Audiol* 34(1):43–51.

Hicks CB, Tharpe AM. (2002) Listening effort and fatigue in school-age children with and without hearing loss. *J Speech Lang Hear Res* 45(3):573–584.

Hornsby BWY. (2013) The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear Hear* 34(5):523–534.

Jaeggi SM, Schmid C, Buschkuehl M, Perrig WJ. (2009) Differential age effects in load-dependent memory processing. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn* 16(1):80–102.

Keidser G. (1996) Selecting different amplification for different listening conditions. *J Am Acad Audiol* 7(2):92–104.

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. (2004) Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 116(4 Pt 1):2395–2405.

Koelewijn T, Zekveld AA, Festen JM, Kramer SE. (2014) The influence of informational masking on speech perception and pupil response in adults with hearing impairment. *J Acoust Soc Am* 135(3):1596–1606.

Mackersie CL, Cones H. (2011) Subjective and psychophysiological indexes of listening effort in a competing-talker task. *J Am Acad Audiol* 22(2):113–122.

Mackersie CL, MacPhee IX, Heldt EW. (2015) Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear Hear* 36(1):145–154.

Macmillan NA, Creelman CG. (2005) *Detection Theory: A User's Guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Magnuson JS, Nusbaum HC. (2007) Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J Exp Psychol Hum Percept Perform* 33(2): 391–409.

Mullennix JW, Pisoni DB, Martin CS. (1989) Some effects of talker variability on spoken word recognition. *J Acoust Soc Am* 85(1): 365–378.

Neidleman MT, Wambacq I, Besing J, Spitzer JB, Koehnke J. (2015) The effect of background babble on working memory in young and middle-aged adults. *J Am Acad Audiol* 26(3):220–228.

Newman CW, Weinstein BE, Jacobson GP, Hug GA. (1991) Test-retest reliability of the hearing handicap inventory for adults. *Ear Hear* 12(5):355–357.

Pichora-Fuller MK, Schneider BA, Daneman M. (1995) How young and old adults listen to and remember speech in noise. *J Acoust Soc Am* 97(1):593–608.

Picou EM, Ricketts TA. (2014) The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear Hear* 35(6):611–622.

Preves DA. (1990) Approaches to noise reduction in analog, digital, and hybrid hearing aids. *Semin Hear* 11(1):39–67.

Ricketts TA, Hornsby BWY. (2005) Sound quality measures for speech in noise through a commercial hearing aid implementing digital noise reduction. *J Am Acad Audiol* 16(5):270–277.

Rönnberg J, Lunner T, Zekveld A, Sörqvist P, Danielsson H, Lyxell B, Dahlström O, Signoret C, Stenfelt S, Pichora-Fuller MK, Rudner M. (2013) The ease of language understanding (ELU) model: theoretical, empirical, and clinical advances. *Front Neurosci* 7(31): 1–17.

Rönnberg J, Rudner M, Foo C, Lunner T. (2008) Cognition counts: a working memory system for ease of language understanding (ELU). *Int J Audiol* 47(2, Suppl):S99–S105.

Rudner M, Lunner T, Behrens T, Thorén ES, Rönnberg J. (2012) Working memory capacity may influence perceived effort during aided speech recognition in noise. *J Am Acad Audiol* 23(8): 577–589.

Sarampalis A, Kalluri S, Edwards B, Hafter E. (2009) Objective measures of listening effort: effects of background noise and noise reduction. *J Speech Lang Hear Res* 52(5):1230–1240.

Schneider BA, Li L, Daneman M. (2007) How competing speech interferes with speech comprehension in everyday listening situations. *J Am Acad Audiol* 18(7):559–572.

Schurman J, Brungart D, Gordon-Salant S. (2014) Effects of masker type, sentence context, and listener age on speech recognition performance in 1-back listening tasks. *J Acoust Soc Am* 136:(6):3337–3349.

Simpson SA, Cooke M. (2005) Consonant identification in N-talker babble is a nonmonotonic function of N. *J Acoust Soc Am* 118(5): 2775–2778.

Sommers MS, Barcroft J. (2006) Stimulus variability and the phonetic relevance hypothesis: effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J Acoust Soc Am* 119(4):2406–2416.

Souza P, Arehart K. (2015) Robust relationship between reading span and speech recognition in noise. *Int J Audiol* 54(10):705–713.

Souza PE, Arehart KH, Shen J, Anderson M, Kates JM. (2015) Working memory and intelligibility of hearing-aid processed speech. *Front Psychol* 6:526.

Souza PE, Sirow L. (2014) Relating working memory to compression parameters in clinically fit hearing AIDS. *Am J Audiol* 23(4): 394–401.

Studebaker GA. (1985) A "rationalized" arcsine transform. *J Speech Hear Res* 28(3):455–462.

Sullivan JR, Osman H, Schafer EC. (2015) The effect of noise on the relationship between auditory working memory and comprehension in school-age children. *J Speech Lang Hear Res* 58(3): 1043–1051.

Tamati TN, Gilbert JL, Pisoni DB. (2013) Some factors underlying individual differences in speech recognition on PRESTO: a first report. *J Am Acad Audiol* 24(7):616–634.

Van Engen KJ, Bradlow AR. (2007) Sentence recognition in native- and foreign-language multi-talker background noise. *J Acoust Soc Am* 121(1):519–526.

Walden TC, Walden BE. (2004) Predicting success with hearing aids in everyday living. *J Am Acad Audiol* 15(5):342–352.

Wong PCM, Nusbaum HC, Small SL. (2004) Neural bases of talker normalization. *J Cogn Neurosci* 16(7):1173–1184.

Zekveld AA, Festen JM, Kramer SE. (2013) Task difficulty differentially affects two measures of processing load: the pupil response during sentence processing and delayed cued recall of the sentences. *J Speech Lang Hear Res* 56(4):1156–1165.

Zekveld AA, George ELJ, Houtgast T, Kramer SE. (2013) Cognitive abilities relate to self-reported hearing disability. *J Speech Lang Hear Res* 56(5):1364–1372.

Zekveld AA, Kramer SE, Festen JM. (2011) Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear Hear* 32(4):498–510.

Zekveld AA, Rudner M, Johnsrude IS, Festen JM, van Beek JH, Rönnberg J. (2011) The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise. *Ear Hear* 32(6):e16–e25.

Supplemental Appendix S1

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 1: Lists A & B

Spoken Sentence	Gender	TIMIT No.	Dialect
1. To his puzzlement, there suddenly was no haze.	F	2007	1
2. Later, you shall know it better.	F	2019	6
3. But such cases were, in the past, unusual.	F	929	7
4. The rides were tame enough -- mostly we talked.	F	2212	6
5. If the farm is rented, the rent must be paid.	F	944	2
6. The angry boy answered but didn't look up.	F	273	2
7. Her wardrobe consists of only skirts and blouses.	F	403	1
8. The sculptor looked at him, bug-eyed and amazed, angry.	F	1244	1
9. The blue rug was suspiciously bright and new.	F	1775	5
10. That is only partly non-nonsense, he began.	M	2291	3
11. Carl lives in a lively home.	M	17	7
12. The causeway ended abruptly at the shore.	M	203	3
13. It looks like we did, when we made blastdown.	M	1826	2
14. Before Thursday's exam, review every formula.	M	14	6
15. Don't ask me to carry an oily rag like that.	M	2	2
16. Drop five forms in the box before you go out.	M	313	7
17. I assume moisture will damage this ship's hull.	M	360	2
18. Laugh, dance, and sing if fortune smiles upon you.	M	407	4

Written Sentence	Gender	TIMIT No.	Dialect
1. Range was a vital detail.	M	480	4
2. Not immediately, as the deputy demanded.	M	1928	2
3. The instinct to discipline has been lost.	M	2119	7
4. He had never felt particularly close to her.	M	2160	3
5. He recognized his jacket and trousers.	M	2153	2
6. Project development was proceeding too slowly.	M	100	5
7. They'll move around that rock all day, following the shade.	M	1073	6
8. The sermon emphasized the need for affirmative action.	M	215	2
9. The gorgeous butterfly ate a lot of nectar.	M	252	1
10. I just saw Jim near the new archeological museum.	F	378	5
11. Greg buys fresh milk each weekday morning.	F	410	3
12. Eating spinach nightly increases strength miraculously.	F	409	7
13. The saw is broken, so chop the wood instead.	F	265	6
14. Be careful not to plow over the flower beds.	F	271	3
15. My father ran him off here six years ago.	F	1894	1
16. Soon the office work claimed all her time.	F	1000	5
17. Most young rise early every morning.	F	26	1
18. Our successors will have an easier task.	F	1081	4

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 2: Lists C & D

Spoken Sentence	Gender	TIMIT No.	Dialect
1. But it was a hopeful sign, he told himself.	F	2146	3
2. You would drive us into shame, he declared.	F	1379	5
3. These men were without capital or experience.	F	1046	2
4. This doctrine was repugnant to my moral sense.	F	1208	2
5. Why, he's going to kill me, he thought wildly.	F	1932	8
6. Change involves the displacement of form.	F	1148	1
7. Heave on those ropes; the boat's come unstuck.	F	2016	1
8. I always say you've got a wonderful husband, Miss Margaret.	F	2307	7
9. Nice country to meet a lion in face to face.	F	1025	1
10. That touched off a total stampede.	M	1879	4
11. She must have put his clothes in the closet.	M	2154	2
12. Fine, day after tomorrow, she added.	M	2224	5
13. A tsunami is not a single wave but a series.	M	965	8
14. He did not mind the useless, kindly questions.	M	1478	4
15. Toothpaste tube should be squeezed from the bottom.	M	429	7
16. The frightened child was gently subdued by his big brother.	M	343	3
17. A lone star shone in the early evening sky.	M	443	6
18. Watch it, big shot, a hoarse voice yelled back.	M	2176	3

Written Sentence	Gender	TIMIT No.	Dialect
1. Gregory and Tom chose to watch cartoons in the afternoon.	M	328	2
2. The advertising verse of Plymouth Variety Store never changes.	M	440	4
3. Cheap stockings run the first time they're worn.	M	304	5
4. They serve cracked wheat, oats or cornmeal.	M	897	8
5. Then fill the system and add a rust inhibitor.	M	872	7
6. Two things contribute to his popularity.	M	1162	4
7. Cottage cheese with chives is delicious.	M	305	4
8. Artificial intelligence is for real.	M	126	7
9. Don't do Charlie's dirty dishes.	M	34	3
10. The pilots' heads looked ridiculously small.	F	1964	6
11. Highway and freeway mean the same thing.	F	233	5
12. It took a tragedy to bring things to a head.	F	906	2
13. One wonders about its applicability to people.	F	1088	4
14. It was secured by an oversized padlock.	F	1942	5
15. Suddenly my reflexes are gone.	F	471	4
16. He had plunged into the dark woods beyond.	F	1544	2
17. Hispanic costumes are quite colorful.	F	86	2
18. Not surprisingly, this approach did not work.	F	1043	4

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 3: Lists E & F

Spoken Sentence	Gender	TIMIT No.	Dialect
1. While he was in custody his wife divorced him.	F	925	8
2. She can decrease the number of temptations.	F	577	4
3. He talked about unauthentic storylines too.	F	618	8
4. By eating yogurt, you may live longer.	F	253	2
5. The red pills are a vitamin-and-iron compound.	F	2220	8
6. Each stag surely finds a big fawn.	F	434	1
7. George seldom watches daytime movies.	F	358	8
8. Critical equipment needs proper maintenance.	F	37	7
9. The previous speaker presented ambiguous results.	F	151	7
10. She always jokes about too much garlic in his food.	M	377	7
11. A roll of wire lay near the wall.	M	16	4
12. Will you please describe the idiotic predicament.	M	79	7
13. See, he's already snapping it up, the dumb jerk.	M	1528	2
14. Regular attendance is seldom required.	M	64	5
15. Move the garbage nearer to the large window.	M	415	6
16. Please sing just the club theme.	M	436	4
17. Only incomplete, imperfect things move towards what they lack.	M	1154	4
18. Stimulating discussions keep students' attention.	M	40	7

Written Sentence	Gender	TIMIT No.	Dialect
1. A moth zig-zagged along the path through Otto's garden.	M	426	2
2. Clasp the screw in your left hand.	M	116	3
3. A connoisseur will enjoy this shellfish dish.	M	353	6
4. Call an ambulance for medical assistance.	M	103	8
5. Continental drift is a geological theory.	M	63	1
6. Family rationing probably will be necessary.	M	1285	3
7. I know I didn't meet her early enough.	M	344	7
8. Bury those uniforms so they won't be found.	M	1907	4
9. Any contributions will be greatly appreciated.	M	62	2
10. Higher toll rates also are helping boost revenues.	F	518	1
11. Husky young man, he said with mock distaste.	F	1474	1
12. Steve wore a bright red cashmere sweater.	F	275	4
13. Young people participate in athletic activities.	F	38	2
14. The new birth is immediate and instantaneous.	F	695	4
15. Only the best players enjoy popularity.	F	211	7
16. Correct execution of my instructions is crucial.	F	147	7
17. The fog prevented them from arriving on time.	F	323	8
18. He will allow a rare lie.	F	11	4

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 4: Lists G & H

Spoken Sentence	Gender	TIMIT No.	Dialect
1. The emperor had a mean temper.	F	127	4
2. What it does: stimulates rumen activity.	F	829	4
3. A pool is no place for a shut trap.	F	793	7
4. Summertime supper, outside, is a natural.	F	742	2
5. This is particularly true in site selection.	F	804	1
6. Our experiment's positive outcome was unexpected.	F	156	3
7. We'll serve rhubarb pie after Rachel's talk.	F	402	4
8. Michael colored the bedroom wall with crayons.	F	209	5
9. Eat your raisins outdoors on the porch steps.	F	69	6
10. A chosen few will become Generals.	M	306	6
11. The bungalow was pleasantly situated near the shore.	M	139	1
12. To his surprise his plan worked perfectly.	M	2063	4
13. The way is to rent a chauffeur-driven car.	M	882	7
14. Spherical gifts are difficult to wrap.	M	448	7
15. She came back the other day to reassure me.	M	542	4
16. He thought he saw a pale face at a window.	M	1663	6
17. Draw every outer line first, then fill in the interior.	M	267	1
18. The best way to learn is to solve extra problems.	M	110	7

Written Sentence	Gender	TIMIT No.	Dialect
1. It was a grotesque hen, five or six feet tall.	M	1563	3
2. My own experience has followed simpler lines.	M	1213	5
3. Comparable trends can be noted elsewhere.	M	677	4
4. Most of our aid will go to those nearing self-sufficiency.	M	1253	6
5. The data are presented in lists and tables.	M	1442	2
6. Whatever their faults, they are not hypocrites.	M	1053	1
7. Most assuredly ideas are invaluable.	M	1143	4
8. Nurses' training here doesn't cost anything.	M	1643	4
9. You don't belong in professional baseball.	M	2235	7
10. One could hear a very faint, ladylike sigh of relief.	F	2297	7
11. We apply auditory modeling to computer speech recognition.	F	251	2
12. Trim excess clay away from outer edges.	F	759	6
13. December and January are nice months to spend in Miami.	F	134	2
14. Put a dollar-and-cents limit on benefits.	F	856	3
15. They remained lifelong friends and companions.	F	163	5
16. With a whirling jump, it could get into gear.	F	1035	5
17. She didn't even give me a chance to refuse.	F	2191	5
18. It's impossible to deal with bureaucracy.	F	81	5

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 5: Lists I & J

Spoken Sentence	Gender	TIMIT No.	Dialect
1. His shoulder felt as if it were broken.	F	385	8
2. Withdraw only as much money as you need.	F	266	2
3. When all else fails, use force.	F	22	2
4. These exclusive documents must be locked up at all times.	F	292	2
5. Cut off every building at the seventh floor.	F	1159	7
6. Hello, boss, he said, and grinned.	F	2264	8
7. Lighted windows glowed jewel-bright through the downpour.	F	1938	5
8. A screwdriver is made from vodka and orange juice.	F	123	3
9. His portrayal of an edgy head-in-the-clouds artist is virtually flawless.	F	620	4
10. The Thinker is a famous sculpture.	M	90	5
11. It's not easy to create illuminating examples.	M	240	5
12. The opposing aircraft continued to come on.	M	1968	5
13. Fists pummeled him as he staggered forward.	M	1539	2
14. Some women get a real thrill out of housework.	M	917	6
15. That diagram makes sense only after much study.	M	372	2
16. The surplus shoes were sold at a discount price.	M	367	3
17. The full moon shone brightly that night.	M	325	1
18. Gwen planted green beans in her vegetable garden.	M	412	7

Written Sentence	Gender	TIMIT No.	Dialect
1. Tradition requires parental approval for under-age marriage.	M	137	5
2. Biological warfare is considered to be primarily a strategic weapon.	M	1319	7
3. The fifth jar contains big, juicy peaches.	M	431	8
4. The heavens refused to give up their weeping.	M	639	1
5. It's hard to tell an original from a forgery.	M	89	5
6. Otherwise, the outlook could be dark indeed.	M	1164	7
7. Please dig my potatoes up before frost.	M	318	7
8. Quietly he determined to foil her.	M	2254	3
9. I saw your horse outside.	M	2332	7
10. However, this inaugural feast did its sponsors no good whatever.	F	963	3
11. His name became synonymous with cold-blooded cruelty.	F	1137	7
12. The triumphant warrior exhibited naive heroism.	F	172	7
13. Here's where luck would normally step in.	F	1714	4
14. Academic aptitude guarantees your diploma.	F	56	5
15. The soft snow was deceitful underfoot.	F	1677	7
16. Pa don't care about the kid.	F	1615	4
17. The humor of the situation can be imagined.	F	627	8
18. I'm going to search this house.	F	2330	3

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 6: Lists K & L

Spoken Sentence	Gender	TIMIT No.	Dialect
1. But that explanation is only partly true.	F	601	3
2. They've never met, you know.	F	1612	8
3. Receiving no answer, they set the fire.	F	1406	1
4. Religion thus becomes integrated with life.	F	1214	6
5. The master's hand has lost none of its craft.	F	622	5
6. He ate four extra eggs for breakfast.	F	130	5
7. A second point requires more extended comment.	F	657	5
8. Consider couples visiting an art museum.	F	1346	6
9. His artistic accomplishments guaranteed him entry into any social gathering.	F	1202	4
10. He's a pinto and he photographs wonderfully.	M	2112	7
11. A battery-powered radio is essential.	M	1284	5
12. Cut a small corner off each edge.	M	93	7
13. Theirs is a sacrificial life by earthly standards.	M	544	5
14. No, they could kill him just as easy right now.	M	1691	8
15. There's still such a thing as mental illness.	M	1836	6
16. We'll work for our keep, the boy said eagerly.	M	1859	3
17. State numbering laws differ from each other in many ways.	M	718	5
18. Pledge to participate in Nevada's aquatic competition.	M	141	6

Written Sentence	Gender	TIMIT No.	Dialect
1. Tornados often destroy acres of farm land.	M	259	4
2. The dark, murky lagoon wound around for miles.	M	338	2
3. It was exposed to a high velocity gas jet.	M	1304	3
4. Do buy all-purpose mugs or cups.	M	743	3
5. Our key salesmen are in appliances and cosmetics.	M	2106	2
6. Draw each graph on a new axis.	M	197	7
7. He told me so today, coming up on the train.	M	2322	1
8. The boost is helpful but inadequate.	M	527	7
9. Shivering, he put on his clothes.	M	2175	3
10. She greeted her husband's colleagues with smiling politeness, offering nothing.	F	1472	4
11. Her classical performance gained her critical acclaim.	F	52	5
12. A large household needs lots of appliances.	F	92	2
13. They used pink, tan, or cream powder.	F	1779	1
14. His salary had reached the ten thousand mark.	F	1746	2
15. Milk appears twice a day.	F	901	3
16. This birth length seems to be typical.	F	1329	6
17. By that time, perhaps something better can be done.	F	1493	8
18. All his family was dead, except for his son.	F	2134	5

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 7: Lists M & N

Spoken Sentence	Gender	TIMIT No.	Dialect
1. He kissed her also, and with deep tenderness.	F	1872	3
2. The viewpoint overlooked the ocean.	F	228	4
3. Production may fall far below expectations.	F	943	1
4. Surely this is a reality we all acknowledge.	F	653	6
5. Several factors contributed to this change.	F	1387	4
6. Put the butcher block table in the garage.	F	223	5
7. The jaw operates by using antagonistic muscles.	F	268	5
8. The small boy put the worm on the hook.	F	216	2
9. Orange juice tastes funny after toothpaste.	F	154	4
10. A leather handbag would be a suitable gift.	M	322	4
11. The leagues are full of guys like that.	M	2239	7
12. They all agree that the essay is barely intelligible.	M	243	4
13. Straw hats are out of fashion this year.	M	192	6
14. My instructions desperately need updating.	M	83	2
15. Trespassing is forbidden and subject to penalty.	M	121	3
16. Her study of history was persistently pursued.	M	994	3
17. Boy, you're stirrin early, a sleepy voice said.	M	1526	2
18. Brush fires are common in the dry underbrush of Nevada.	M	288	5

Written Sentence	Gender	TIMIT No.	Dialect
1. Coconut cream pie makes a nice dessert.	M	33	8
2. Hastily the boy switched on a ceiling light.	M	1772	7
3. His head came up and he said it defiantly.	M	2069	5
4. On all sides doors were being slammed in his face.	M	1594	5
5. The trial balloons are afloat.	M	594	2
6. It's perhaps a mile from here where we sit.	M	1787	7
7. We are all involved in them, deeply involved.	M	668	2
8. That would make him mad enough if he was sober.	M	1613	5
9. Eternity is no time for recriminations.	M	1812	6
10. Let all projects dry slowly for several days.	F	773	5
11. Young children should avoid exposure to contagious diseases.	F	199	4
12. Every movement she made seemed unnecessarily noisy.	F	1919	5
13. Steve collects rare and novel coins.	F	326	5
14. Fill that canteen with fresh spring water.	F	75	2
15. Jane may earn more money by working hard.	F	4	1
16. A muscular abdomen is good for your back.	F	113	6
17. Both cars were slightly damaged.	F	485	6
18. His successors have adopted the opposite alternative.	F	1122	4

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 8: Lists O & P

Spoken Sentence	Gender	TIMIT No.	Dialect
1. At least, the wheels dug in.	F	2011	7
2. A huge tapestry hung in her hallway.	F	106	5
3. Most precincts had a third of the votes counted.	F	148	3
4. Cement is measured in cubic yards.	F	85	4
5. The sheriff's swivel chair tilted back.	F	1681	7
6. They all like long hot showers.	F	391	5
7. There are several sources of evidence on the micrometeorite environment.	F	1315	7
8. Just drop notices in any suggestion box.	F	400	5
9. Neglect means spiritual paralysis or death.	F	670	4
10. The singer's finger had a splinter.	M	59	3
11. Basketball can be an entertaining sport.	M	44	2
12. The nearest synagogue may not be within walking distance.	M	320	4
13. Oh, he'll be a plumber, came the answer.	M	1001	3
14. Oil-field workers were a rough-tough lot.	M	1929	7
15. Addition and subtraction are learned skills.	M	45	4
16. Beg that guard for one gallon of gas.	M	28	3
17. He said: the crazy fool, half aloud.	M	1754	3
18. The local drugstore was charged with illegally dispensing tranquilizers.	M	324	1

Written Sentence	Gender	TIMIT No.	Dialect
1. Your leg muscles and back muscles feel weary.	M	1646	3
2. The news agency hired a great journalist.	M	213	3
3. Mosquitoes exist in warm, humid climates.	M	152	7
4. Land-based radar would help with this task.	M	1276	4
5. Withdraw all phony accusations at once.	M	191	2
6. Penguins live near the icy Antarctic.	M	97	5
7. We are open every Monday evening.	M	257	1
8. A doctor was in the ambulance with the patient.	M	238	2
9. This is followed by a boom in conferences.	M	911	2
10. The overweight charmer could slip poison into anyone's tea.	F	433	2
11. Youngsters love common candy as treats.	F	94	4
12. Good service should be rewarded by big tips.	F	82	8
13. Chocolate and roses never fail as a romantic gift.	F	61	7
14. The golfing fathers ruled in his favor.	F	474	6
15. Nevertheless, she continued to move upward.	F	1917	5
16. Objects made of pewter are beautiful.	F	124	6
17. First go over the type of coverage you now have.	F	874	5
18. It had gone like clockwork.	F	1804	2

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 9: Lists Q & R

Spoken Sentence	Gender	TIMIT No.	Dialect
1. A flame would use up air.	F	1289	5
2. That'll be plenty to eat by.	F	1810	4
3. Help celebrate your brother's success.	F	35	6
4. John cleaned shellfish for a living.	F	374	2
5. An official deadline cannot be postponed.	F	74	1
6. The cow wandered from the farmland and became lost.	F	363	2
7. The roof of the command post began to buckle.	F	1886	2
8. This tragic lad had forged his own shackles.	F	2123	8
9. Their basis seems deeper than mere authority.	F	1211	5
10. They were chasing a rain cloud.	M	1066	7
11. Then we'd really have someplace to go.	M	466	4
12. We welcome many new students each year.	M	381	3
13. In fact, nobody saw us, cop or citizen.	M	1803	2
14. Even a simple vocabulary contains symbols.	M	53	7
15. Blockade is one answer offered by experts.	M	1275	5
16. George is paranoid about a future gas shortage.	M	382	5
17. Place work on a flat surface and smooth out.	M	774	1
18. Pam gives driving lessons on Thursdays.	M	379	3

Written Sentence	Gender	TIMIT No.	Dialect
1. We experience distress and frustration obtaining our degrees.	M	58	2
2. They also want money, prestige, and security.	M	1220	4
3. Combine all the ingredients in a large bowl.	M	118	5
4. Do not draw yarn too tight.	M	776	3
5. The coalition was to prove inadvisable.	M	961	7
6. Don't plan meals that are too complicated.	M	738	2
7. We must be ready for any needed sacrifice.	M	489	3
8. Oh, no, not again, he said aloud.	M	1668	3
9. When they got home at midnight she was tired out.	M	2147	2
10. Remember to allow identical twins to enter freely.	F	260	8
11. Scientific progress comes from the development of new techniques.	F	136	3
12. The essay undeniably reflects our view ably.	F	258	8
13. My ideal morning begins with hot coffee.	F	255	4
14. They own a big house in the remote countryside.	F	389	5
15. Herb's birthday occurs frequently on Thanksgiving.	F	244	5
16. Calcium makes bones and teeth strong.	F	394	1
17. His sudden departure shocked the cast.	F	111	5
18. Dolphins are intelligent marine mammals.	F	355	5

Perceptually Robust English Sentence Test – Revised (PRESTO-R) Lists

Pair 10: Lists S & T

Spoken Sentence	Gender	TIMIT No.	Dialect
1. You saw them always together those years.	F	2189	7
2. The former scout's alibi couldn't be shaken.	F	1944	5
3. Cable confirmation, it said translated.	F	487	7
4. The system works as an impersonal mechanism.	F	1083	2
5. His sweet whisper came after great effort.	F	1602	5
6. Norwegian sweaters are made of lamb's wool.	F	198	2
7. The blow encountered silky hair and hard bone.	F	1505	3
8. Differences were related to social, economic, and educational backgrounds.	F	921	1
9. Choose carefully between contributory or non-contributory pension plans.	F	858	7
10. With this no loyal citizen can quarrel.	M	606	7
11. He was, thus, an early and spectacular victim.	M	1246	4
12. If dark came they would lose her.	M	1033	6
13. He stole a dime from a beggar.	M	104	5
14. Elderly people are often excluded.	M	43	2
15. The hallway opens into a huge chamber.	M	241	7
16. Let us now give some thought to the soul.	M	1216	4
17. The third crawling man forced himself erect.	M	1600	7
18. This coat looks like a rag heap.	M	1636	3

Written Sentence	Gender	TIMIT No.	Dialect
1. As a rule, part-time farmers hire little help.	M	946	7
2. In the long run, it pays to buy quality clothing.	M	287	4
3. Add things as you find you need 'em.	M	2271	3
4. Aim to balance your employee benefit package.	M	853	7
5. It all takes place in the eighteenth century.	M	628	7
6. We can get it if we dig, he said patiently.	M	2000	4
7. To put it bluntly, you are getting out-moded.	M	910	5
8. The drug is also incorporated in supplements.	M	823	8
9. We've got plenty of time to think about that.	M	1635	7
10. They used an aggressive policeman to flag thoughtless motorists.	F	437	6
11. Gently place Jim's foam sculpture in the box.	F	76	3
12. Turbulent tides rose as much as fifty feet.	F	956	6
13. One thing, he thought, nobody knows about it yet.	F	1673	4
14. He picked up nine pairs of socks for each brother.	F	390	5
15. Development requires a long-term approach.	F	1252	5
16. These always contain metallic inclusions.	F	1438	8
17. The kid has no manners, boys.	F	2048	1
18. We always thought we would die with our boots on.	F	908	3