Original Article

A Flexible Question-and-Answer Task for Measuring Speech Understanding

Trends in Hearing
2016, Vol. 20: 1–8
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2331216516678706
tia.sagepub.com
SAGE

Virginia Best[1], Timothy Streeter[1], Elin Roverud[1], Christine R. Mason[1], and Gerald Kidd Jr.[1]

Abstract

This report introduces a new speech task based on simple questions and answers. The task differs from a traditional sentence recall task in that it involves an element of comprehension and can be implemented in an ongoing fashion. It also contains two target items (the question and the answer) that may be associated with different voices and locations to create dynamic listening scenarios. A set of 227 questions was created, covering six broad categories (days of the week, months of the year, numbers, colors, opposites, and sizes). All questions and their one-word answers were spoken by 11 female and 11 male talkers. In this study, listeners were presented with question-answer pairs and asked to indicate whether the answer was true or false. Responses were given as simple button or key presses, which are quick to make and easy to score. Two preliminary experiments are presented that illustrate different ways of implementing the basic task. In the first experiment, question-answer pairs were presented in speech-shaped noise, and performance was compared across subjects, question categories, and time, to examine the different sources of variability. In the second experiment, sequences of question-answer pairs were presented amidst competing conversations in an ongoing, spatially dynamic listening scenario. Overall, the question-and-answer task appears to be feasible and could be implemented flexibly in a number of different ways.

Keywords

speech comprehension, dynamic conversations, realistic tests

Date received: 19 May 2016; revised: 20 October 2016; accepted: 20 October 2016

Introduction

A typical conversation consists of a sequence of exchanges between two and more people, often relying on frequent questions and answers to elicit and convey information and clarify meaning. In ordinary social settings, this communication exchange often occurs in a background of noise or competing conversations. To participate successfully in a conversation, one must hear what is said, understand what is said, resist distraction from competing sounds, and give appropriate responses. When the conversation involves several participants, there may be the extra challenge of following the thread of the conversation despite frequent and sometimes unpredictable changes in voice and location of the current "target" talker.

Speech *recognition* tests used in the laboratory and in the clinic (e.g., HINT, Nilsson, Soli, & Sullivan, 1994; QuickSIN, Killion, Niquette, Gudmundsen, & Banerjee, 2004) typically differ from real-world listening during conversations in several important ways. First, they usually are trial based, consisting of a single word or short sentence followed by a silent gap for responding. Second, they typically require the listener to repeat back from memory what was heard (e.g., by speaking or typing), without the need for any consideration of or response to the *content* of the message. These tests can provide accurate and repeatable measures of speech intelligibility in quiet or in masked conditions and are used widely for both research and clinical purposes (e.g., for optimizing hearing aid fittings). However, there is increasing interest in more realistic speech tests that capture some of the more complex aspects of listening that are part of natural conversations.

[1] Department of Speech, Language and Hearing Sciences, Boston University, MA, USA

Corresponding author:
Virginia Best, Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Ave, Boston, MA 02215, USA.
Email: ginbest@bu.edu

Speech *comprehension* tests typically require listeners to follow a passage of discourse (e.g., a lecture or story) and then to answer a series of content-related questions about what was heard (e.g., Gordon, Daneman, & Schneider, 2009; Murphy, Daneman, & Schneider, 2006; Schneider, Daneman, Murphy, & See, 2000; Sommers et al., 2011; Tye-Murray et al., 2008). For longer passages, this testing format introduces a significant memory requirement, which may not be the aspect of the communication process of greatest interest to the tester if the primary goal is to assess natural listening abilities. One way to minimize the influence of memory load is to use short-duration passages (e.g., up to 1 min; Kei et al., 2003; Kei & Smyth, 1997) or to query the listener during the stimulus instead of at the end of the presentation (e.g., Best, Keidser, Buchholz, & Freeston, 2016; Best, Keidser, Freeston, & Buchholz, 2016; Hafter, Xia, & Kalluri, 2013).

In the present article, we present an alternative approach in which the speech material is reduced to brief question-answer pairs, which represent one of the basic components of typical conversations (Stivers et al., 2009). This approach was motivated by the "Helen test," a clinical test developed for assessing the speech-reading abilities of individuals with profound hearing loss (Kei, Smyth, Murdoch, & McPherson, 2000; Ludvigsen, 1974; Plant, Phillips, & Tsembis, 1982). In the original form of this test, a simple question is spoken by the clinician, and the listener gives a one-word verbal answer. Here, we describe a variation in which the stimuli comprise both a question and an answer, and the listener is required to indicate whether the answer is true or false.

The new design has several useful features that potentially provide advantages over either of the approaches mentioned earlier. First, the use of sentence-length stimuli means that the task has very low-memory requirements. Second, because the response depends on a simple true or false decision, the responses may be obtained very rapidly (each judgment requires only a single button or key press). This structure allows multiple questions and answers to be presented in succession so as to emulate an ongoing, continuous conversation to which the listener must maintain attention. The simple motor responses required for pressing buttons or keys do not require the engagement of vision. Thus, this response format may be useful in situations where directing attention to a visual interface—such as a monitor or touch screen—in order to register a response would interfere with the task itself (e.g., an audiovisual task). Another practical advantage of these binary responses is that they can be scored immediately and automatically. Finally, because the information on each "trial" is distributed across two parts (the question and the answer), it is possible to introduce intratrial transitions in voice or location. This provides the opportunity to assess speech understanding in the presence of dynamic source variations, which can be useful for studying attention switching, or for evaluating adaptive hearing aid algorithms. Also, the ability to vary parameters of the sources is well suited to experiments that incorporate varying degrees of listener uncertainty into the design.

In this report, we briefly describe the question-and-answer materials and procedures, present two representative experiments intended to determine the basic feasibility, and illustrate possible implementations of the new task.

Materials

A set of 227 questions was created, using the original Helen test lists as a guide (Ludvigsen, 1974; Plant et al., 1982). Each question is simple and unambiguous and is associated with a single-word answer. The questions cover six broad categories, as shown in Table 1. The table lists an example question (and correct answer) from each category, as well as the number of questions in that category, and the number of valid answers available for each question in that category. Although each question had only one correct answer, incorrect but valid answers were chosen from the answers to other questions in the corpus. For example, valid answers to the question "What day comes after

Table 1. Description of the Six Question Categories.

Category	Number of questions	Number of valid answers	Example question	Correct answer	Example incorrect answer
Days	14	7	What day comes after Monday?	Tuesday	Friday
Months	24	12	What month comes before April?	March	October
Colors	20	8	What color is the sky?	Blue	Green
Opposites	20	20	What is the opposite of up?	Down	Open
Sizes	20	2	Which is bigger, an elephant, or a mouse?	Elephant	Mouse
Numbers	129	23	What is two plus two?	Four	Eight

Monday?'' would be any of the seven days of the week, while valid answers to the question ''What is two plus two?'' would be any numeric answer. On the other hand, the question ''Which is bigger, an elephant or a mouse?'' has only two valid answers, ''elephant'' and ''mouse.''

Each question and answer was spoken by each of 24 talkers (12 female, 12 male) and recorded by Sensimetrics, Inc. (Malden, MA).[1] The recordings were made in an audiometric booth using an Edirol R-44 digital recorder (24 bit or 48 kHz) and a Rode NT1-A condenser microphone. Talkers read the materials from a scrolling list that was presented on a monitor to set the overall pace. During offline editing, the recordings were trimmed, and each question and answer from each of the talkers was normalized to the same broadband root-mean-square level. The decision was made to exclude one male talker who made a number of errors during the recording session. To maintain even numbers, one female talker was also excluded, chosen on the basis of informal listening to be somewhat of an outlier in the group. To illustrate the acoustic variability across the remaining set of 22 talkers, Table 2 lists the F0 for each talker, extracted from the entire set of questions spoken by that talker using PRAAT software (Boersma & Weenink, 2016). Also listed in Table 2 are the average word durations per talker, calculated by dividing the duration of each question by the number of words in that question and averaging across all questions.

Table 2. Mean Word Duration and Mean F0, Calculated Across all Questions, for Each of the 22 Talkers. Also Shown Are the Across-Talker Mean and Standard Deviations for Females and Males.

	Female		Male	
Talker	Duration (ms)	F0 (Hz)	Duration (ms)	F0 (Hz)
1	394	231	343	121
2	378	210	337	110
3	351	227	298	139
4	341	204	290	117
5	357	214	350	133
6	328	200	363	99
7	334	220	349	132
8	306	193	360	153
9	353	226	339	115
10	402	207	367	127
11	418	223	409	120
Mean (±*SD*)	360 ± 34	214 ± 12	346 ± 32	124 ± 15

Experiment 1

The purpose of the first experiment was to evaluate the basic feasibility of the question-and-answer concept under relatively simple and controlled conditions and to examine how performance varies across subjects, across the different categories of questions, and across time.

Methods

Eight young adult listeners with normal hearing participated (age 19–30 years). A trial consisted of one question-answer pair separated by 0.5 s (measured from offset of question to onset of answer). The answer was correct on 50% of trials and incorrect (but valid) on 50% of trials. The listeners were informed about the a priori probability of a correct answer. Two different voices were chosen randomly on each trial for the question and the answer. Each of the 227 questions was presented in speech-shaped noise at each of four signal-to-noise ratios (SNRs: −14, −12, −10, and −8 dB). The SNR was set by varying the level of the target relative to the noise which was presented at 65 dB SPL. The noise was shaped based on the average magnitude spectrum of the entire set of questions (averaged across all talkers) and was ramped on and off with the start of the question and the end of the answer using 10-ms raised-cosine windows. Stimuli were controlled in MATLAB (MathWorks Inc., Natick, MA) and presented diotically via a 24-bit soundcard (RME HDSP 9632) through a pair of headphones (Sennheiser HD280 Pro). The listener was seated in a double-walled sound-treated booth (Industrial Acoustics Company) in front of a computer monitor. The task was a two-alternative forced choice in which the listeners indicated whether the answer given was true or false by clicking with a mouse on one of two buttons displayed on the monitor. For each listener, the order of presentation of the trials (908 in total) was randomized, and no feedback was given. The total testing time was around 2 hours, and short breaks were enforced every 30 min.

Results and Discussion

Figure 1(a) shows psychometric functions for each listener (averaged across all questions) as well as the average performance at each SNR. Mean performance for the chosen SNRs ranged from near chance (65% correct) to near perfect (92%). The error bars show across-subject standard deviations, which were 4.5 percentage points on average. Logistic fits were obtained for each listener and used to estimate 75% thresholds and slopes. Threshold SNRs ranged from −13 to −11 dB (mean −12 dB). Slopes ranged from 4%/dB to 10%/dB (mean

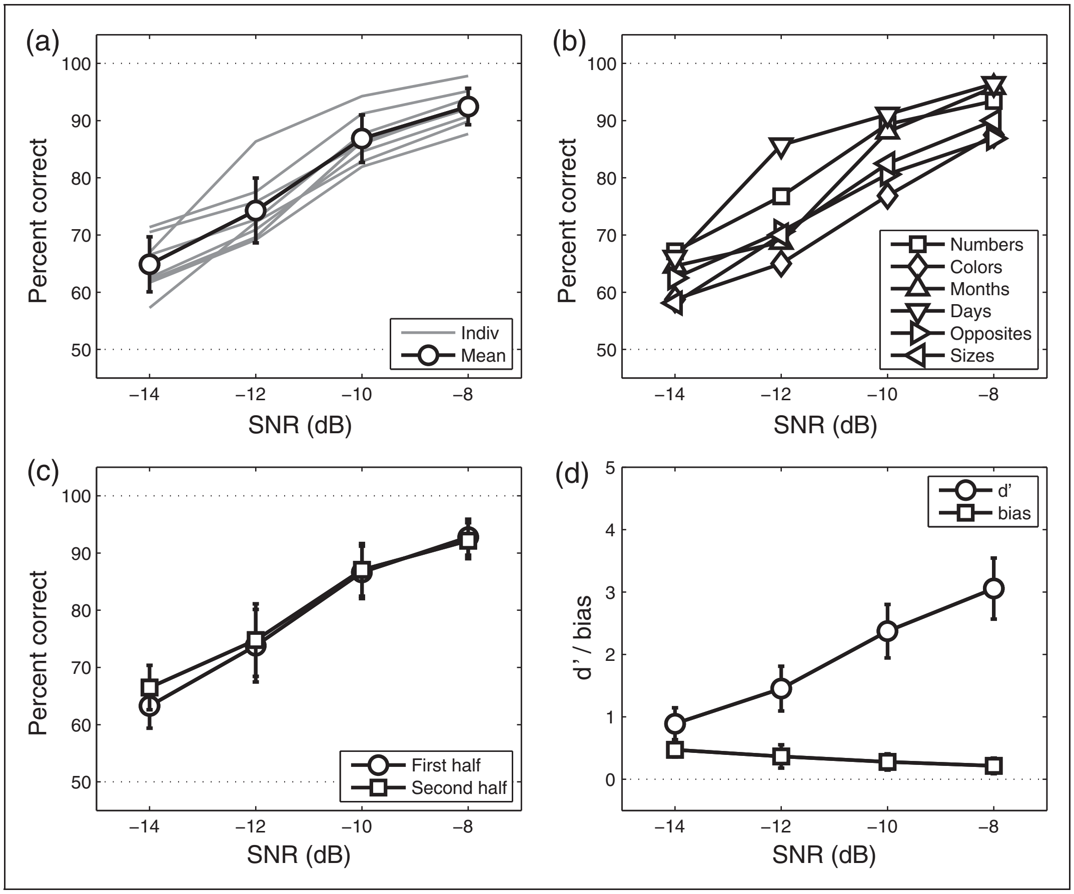

Figure 1. (a) Psychometric functions for each listener (gray lines) and the mean psychometric function (black lines and circles). (b) Mean psychometric functions for each category type. (c) Mean psychometric functions based on the first half (circles) and second half (squares) of the trials completed by each subject at each SNR. (d) Mean psychometric function in units of d′ (circles) and bias (squares). Error bars, where shown, represent across-subject standard deviations.

7%/dB). Figure 1(b) shows psychometric functions for each category (averaged across all questions in that category and all listeners). There were systematic differences in performance across the different question categories, with the best performance observed for the "days" category and poorest performance observed for the "colors" category. A repeated-measures analysis of variance (ANOVA) confirmed that the main effects of category, $F(5,35) = 10.5$, $p < .001$, and SNR, $F(3,21) = 125.0$, $p < .001$, were significant and did not significantly interact, $F(15,105) = 1.2$, $p = .3$.

To examine the data for possible learning effects across the session, the 227 trials per subject per SNR were divided into two halves (the first 114 trials and the second 113 trials). Scores were computed for each half, and the across-subject means are shown in Figure 1(c). A repeated-measures ANOVA found only the expected effect of SNR, $F(3, 21) = 203.1$, $p < .001$, but no significant effect of half, $F(1, 7) = 1.3$, $p = .3$, and no interaction between SNR and half, $F(3, 21) = 0.8$, $p = .5$. To obtain an estimate of test–retest reliability, the standard deviation of the differences between halves was calculated for each SNR. This value came to 4.9 percentage points on average, which compares favorably with the across-subject standard deviations.

Because the subjects' responses take the form of true or false judgments, it is possible to consider the data in terms of signal detection theory (e.g., Green & Swets, 1966) such that the response "true" when the answer is correct is a hit while a response "true" when the answer is incorrect is a false alarm. Such an analysis could be useful if it were desirable to analyze performance in terms of observer bias separate from sensitivity. Figure 1(d) shows the group mean results plotted as sensitivity (d′) and bias. Although the listeners were informed that correct and incorrect answers were equally likely, the observed bias was slightly positive, indicating that subjects had more of a tendency to respond "false" than "true" in this task, especially at the lower SNRs. At low SNRs, there would be trials in which the question or answer would be partly inaudible, and it may simply be that listeners were more likely to map their uncertainty to the negative "false" rather than the affirmative "true." It is also possible, even though there was an equal probability of true and false answers, that subjects were biased by the knowledge that the set of all possible false answers is generally large, whereas there is only one true answer .

Experiment 2

The purpose of Experiment 2 was to demonstrate a more sophisticated implementation of the question-and-answer task that includes competing talkers and spatial dynamics. The intent was to capture aspects of normal conversation in which two or more participants located at different points in the sound field take turns speaking while competing conversations take place at other locations. Two conditions were contrasted, one in which the target conversation participants occupied a single location and a second in which the participants were spatially distributed.

Methods

The experiment was completed by eight young adult listeners with normal hearing (age 21–24 years) who participated as part of a larger study but were not involved in Experiment 1. The testing environment, stimulus control, and delivery were similar to those described for Experiment 1, but the specific procedures differed in some respects.

A run consisted of 12 question-answer pairs ("trials") with a gap of 0.5 s between the question and answer within a trial and also between trials (i.e., between the end of one answer and the beginning of the next question). The answer was correct on 50% of the trials and incorrect (but valid) on the remaining 50% of the trials. The questions and answers (targets) were spoken by three randomly selected talkers in a run such that the voice changed between each consecutive item (question or answer) in the sequence.

The spatial and temporal layout of an example stimulus is illustrated in Figure 2. The three target talkers were located at −30°, 0°, and +30° azimuth. The targets were presented simultaneously with three maskers, each of which consisted of a conversation between a male and a female at two different locations. The masker conversations were located at −60°/−45°, −15°/+15°, and +45°/+60° azimuth. Six scripted "everyday" conversations were recorded from three pairs of nontarget talkers (two conversations per pair), and on each run one conversation from each pair was drawn at random and assigned at random to the three pairs of masker locations. The maskers were ramped on 1 s before the first question and ramped off 1.5 s after the final answer.

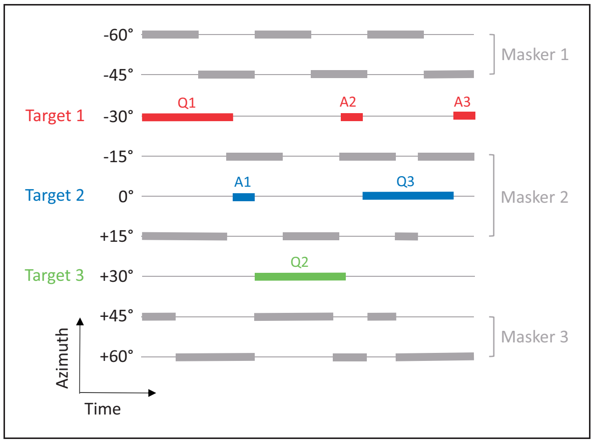

Figure 2. Spatial and temporal configuration of the targets and maskers in the dynamic listening condition. The nine horizontal lines represent the nine stimulus positions (from −60° to +60° azimuth), of which three were potential target locations (−30°, 0°, +30°), and the remaining six were occupied by three pairs of masker talkers engaged in conversation. The shaded bars indicate the times during which a particular talker was speaking (colors: targets; gray: maskers). In this example, three questions and answers (labeled Q1, A1, etc.) out of the sequence of 12 are shown. The fixed condition was identical except that all questions and answers came from one of the three target locations.

Spatialization of the questions, answers, and masker conversations was achieved using binaural impulse responses measured on a KEMAR manikin in a mildly reverberant room at a distance of 5 ft. The target stimuli were presented at 55 dB SPL (as measured at the headphones for a frontal sound), and the level of each masker conversation was varied to set the target-to-masker ratio (TMR) to one of four values (−10, −5, 0, and +5 dB). Note that since TMR was defined relative to each competing conversation, at 0 dB TMR, all of the competing talkers were equal in level. At all other TMRs, negative or positive, it is worth noting that there was a "level cue" that potentially could be used to help distinguish the target talkers from the masker talkers.

Two conditions were examined. In the dynamic condition (shown in Figure 2), the location of the questions and answers moved unpredictably across the three target locations with a forced transition on every question and every answer. On each run in this condition, each of the three target voices was associated with one of the three target locations. In the fixed condition, everything was identical except that the three target talkers occupied a single location throughout a run.

The task was a two-alternative forced choice in which the listener indicated "true" or "false" after each question-answer pair by pressing one of two buttons on a hand-held keypad. Thus, although the run was continuous, 12 key presses were elicited by the end of a run. Because the questions and answers were expected to be inaudible on some trials, a visual cue was provided on the monitor so that the listener knew when to listen and respond. The visual cue consisted of the letters "Q" and "A," which were presented at the same time as the questions and answers and were mapped to one of three locations on the screen corresponding to the azimuths of the talkers. Trials in which a response was not registered were rare but were excluded from the analysis when they occurred.

The listeners completed five sessions, each lasting about half an hour, which were distributed across four visits to the laboratory. The first session was counted as training and was not included in the analysis. In a single session, the fixed condition was tested for each of the target locations, and the random condition was tested 3 times to ensure that each of the three target locations was sampled as often as in the fixed condition. These six runs were tested at each of the four TMRs in a random order, for a total of 24 runs per session. This design resulted in a total of 36 trials per condition per TMR for each listener.

Results and Discussion

Figure 3(a) shows psychometric functions for the fixed and dynamic spatial conditions (averaged across listeners and target locations). Performance was very good for this set of TMRs, ranging from 73% to 97% correct. A repeated-measures ANOVA found a significant main effect of TMR, $F(3,\ 21) = 204.9$, $p < .001$, but no main effect of condition, $F(1,\ 7) = 0.4$, $p = .6$, and no interaction, $F(3,\ 21) = 0.1$, $p = 1.0$, suggesting that switching attention at this rate under these conditions did not compromise performance. When examined in terms of d′ (Figure 3(b)), the close relationship between the fixed and dynamic conditions remained, and again a slightly positive bias was observed, especially at the lower TMRs, indicating that subjects had more of a tendency to respond "false" than "true."

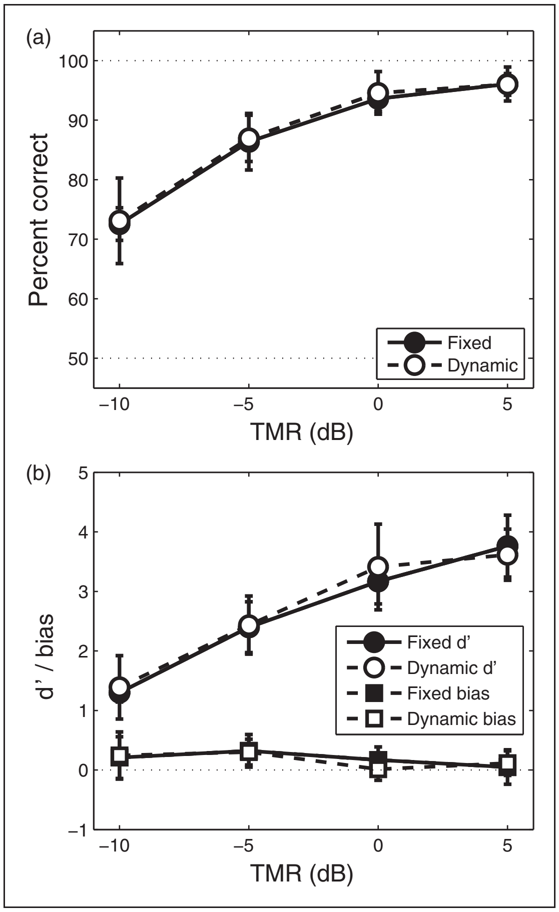

Figure 3. (a) Psychometric functions for the fixed and dynamic listening conditions (averaged across all locations and all listeners). (b) Psychometric functions in units of d′ (circles) and bias (squares) for the fixed and dynamic listening conditions (averaged across all locations and all listeners). Error bars in both panels show across-subject standard deviations.

The similarity in performance for fixed and dynamic conditions is somewhat surprising in light of previous studies, using tasks based on the recall of sentences or digit sequences, in which substantial costs of having to redirect spatial attention within or between trials have been observed (e.g., Best, Ozmeral, Kopčo, & Shinn-Cunningham, 2008; Brungart & Simpson, 2007; Jensen, Johannesson, Laugesen, & Hietkamp, 2012; Kidd, Arbogast, Mason, & Gallun, 2005). There are several potential reasons why a similar cost of spatial uncertainty was not observed in our study. First, visual cues accompanied each question and answer, which completely eliminated the uncertainty about the location where attention should be directed (although one could still expect a cost to be apparent if a listener was not able to switch quickly enough). Second, most of the previous studies involved tasks that were characterized by a high degree of "informational masking" (e.g., Kidd, Mason, Richards, Gallun, & Durlach, 2008). Unlike our study, those tasks employed targets and maskers that were drawn from the same set of items meaning that explicit masker confusions were possible when registering responses. This might be a prerequisite for observing large costs of switching attention, where an inaccurate switch can bring the wrong source into the foreground (e.g., Best et al., 2008). Third, after a spatial transition, the target talker always appeared at a new location where previously there was no sound source. As such, the novelty of the new onset at that location may have minimized any associated costs. However, this explanation is somewhat unlikely, since the maskers consisted of conversations that alternated between two locations and thus also contained frequent new onsets. Fourth, it is worth remembering that the information in these question and answer stimuli is distributed across time.

So, even if listeners are not perfectly focused at the onset of a question or answer, it might be that this is often not critical to performance. This is consistent with another recent study (Best, Keidser, Freeston, et al., 2016) in which listeners were presented with either a monologue (at a single location) or conversations involving two or three talkers (at different locations). In that study, there was no effect of increasing the spatial dynamics on the ability of listeners to follow along and answer content-related questions.

Conclusions

This report described a new speech task in which the listener monitors the accuracy of short answers given by one talker in response to simple questions asked by another talker. Our initial evaluation of the task, in which the questions and answers were presented in speech-shaped noise, found variations across subjects and across question categories but no evidence of learning effects and reasonable reliability. The question-and-answer task may be useful for investigations concerned with real-world listening abilities, as it can be configured to tap into aspects of natural conversation that usually are not captured by traditional speech tests such as the need to comprehend the meaning of speech, to process and respond continuously, and to cope with dynamic variations in the voice and location of the target talker often in competition with task-irrelevant conversations. Here, we provided one example of an implementation that incorporated all of these aspects to investigate the effect of natural spatial dynamics on speech comprehension. Overall, the question-and-answer test appears to be extremely flexible, with numerous possible implementations that may be tailored to suit a range of applications.

Acknowledgments

The portions of the work were presented at the ARO MidWinter Meeting (San Diego, February 2016). The authors would like to acknowledge Lorraine Delhorne for help with subject recruitment and Todd Jennings for technical help.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by NIH-NIDCD award DC04545.

Note

1. The recordings are freely available and can be obtained by contacting the first author at ginbest@bu.edu.

References

Best, V., Keidser, G., Buchholz, J. M., & Freeston, K. (2016). Development and preliminary evaluation of a new test of ongoing speech comprehension. *International Journal of Audiology*, *55*, 45–52.

Best, V., Keidser, G., Freeston, K., & Buchholz, J. M. (2016). A dynamic speech comprehension test for assessing real-world listening ability. *Journal of the American Academy of Audiology*, *27*, 515–526.

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*, 13173–13177.

Boersma, P., & Weenink, D. (2016, September 4). *Praat: doing phonetics by computer [Computer program]. Version 6.0.20*. Retrieved from https://github.com/praat/praat

Brungart, D. S., & Simpson, B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Perception & Psychophysics*, *69*, 79–91.

Gordon, M. S., Daneman, M., & Schneider, B. A. (2009). Comprehension of speeded discourse by younger and older listeners. *Experimental Aging Research*, *35*, 277–296.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley and Sons.

Hafter, E. R., Xia, J., & Kalluri, S. (2013). A naturalistic approach to the cocktail party problem. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, & H. E. Gockel (Eds.), *Basic aspects of hearing: Physiology and perception* (pp. 527–534). New York, NY: Springer.

Jensen, N. S., Johannesson, R. B., Laugesen, S., & Hietkamp, R. K. (2012). Measuring speech-in-speech intelligibility with target location uncertainty. In T. Dau, M. L. Jepsen, J. Christensen-Dalsgaard, & T. Poulsen (Eds.), *Speech perception and auditory disorders. Proceedings of the international symposium on audiological and auditory research (ISAAR)* (pp. 135–142). Denmark: The Danavox Jubilee Foundation.

Kei, J., & Smyth, V. (1997). Measuring the ability of hearing impaired children to understand connected discourse: A comparison of two methods. *British Journal of Audiology*, *31*, 283–297.

Kei, J., Smyth, V., Burge, E., Fernando, S., Fiteni, R., Haslam, S., . . . McMahon, S. (2003). Measuring the ability of children to understand everyday speech using computer technology: A normative study. *Asia Pacific Journal of Speech Language and Hearing Research*, *8*, 235–242.

Kei, J., Smyth, V., Murdoch, B., & McPherson, B. (2000). Measuring the understanding of sentences by hearing-impaired children: Comparison with connected discourse ratings. *Audiology*, *39*, 38–49.

Kidd, G. Jr., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, *118*, 3804–3815.

Kidd, G. Jr, Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–189). New York, NY: Springer Handbook of Auditory Research.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & 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. *Journal of the Acoustical Society of America*, *116*, 2395–2405.

Ludvigsen, C. (1974). Construction and evaluation of an audio-visual test (The Helen test). *Scandinavian Audiology Supplementum*, *4*, 67–75.

Murphy, D. R., Daneman, M., & Schneider, B. A. (2006). Why do older adults have difficulty following conversations? *Psychology and Aging*, *21*, 49–61.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, *95*, 1085–1099.

Plant, G. L., Phillips, D., & Tsembis, J. (1982). An auditory-visual speech test for the elderly hearing impaired. *Australian Journal of Audiology*, *4*, 62–28.

Schneider, B., Daneman, M., Murphy, D., & See, S. (2000). Listening to discourse in distracting settings: The effects of aging. *Psychology and Aging*, *15*, 110–125.

Sommers, M. S., Hale, S., Myerson, J., Rose, N., Tye-Murray, N., & Spehar, B. (2011). Listening comprehension across the adult lifespan. *Ear and Hearing*, *32*, 775–781.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., . . . Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*, 10587–10592.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology*, *47*, S31–S37.