AJA

Research Article

A Comparison of Two Methods for Measuring Listening Effort As Part of an Audiologic Test Battery

Jani Johnson,[a] Jingjing Xu,[a] Robyn Cox,[a] and Paul Pendergraft[a]

Purpose: We evaluated 2 measures of listening effort (a self-report measure and a word recall measure) regarding their suitability for inclusion in a comprehensive audiologic testing protocol. The relationship between the 2 measures was explored, and both measures were examined with regard to validity, sensitivity, and effect on speech intelligibility performance.
Method: Thirty adults with normal hearing participated. Speech intelligibility performance was evaluated at 4 signal-to-noise ratios by using keywords embedded in both high- and low-context sentences. Listening effort was evaluated at set intervals throughout the speech intelligibility task.
Results: Results obtained with the 2 measures were consistent with expected changes in listening effort. However, data obtained with the self-report method demonstrated greater sensitivity to these changes. The 2 measures were uncorrelated. Under certain conditions, speech intelligibility performance was more negatively affected when the word recall measure was used. Exploration of additional theoretical and practical considerations supported a conclusion that the self-report measure was preferable for measuring listening effort simultaneously with speech intelligibility.
Conclusion: The results of this study provide a rationale for preferring the self-report measure of listening effort over the word recall measure when testing audiologic outcomes.

Speech understanding is a complex task that includes use of peripheral hearing, but also involvement of higher-level auditory and cognitive processes such as auditory attention and memory (Davis, 1964; Marslen-Wilson, 1987). When processing speech, central (top-down) and peripheral (bottom-up) mechanisms function in tandem, providing listeners with normal hearing with redundant information. Because of this redundancy, understanding often remains possible even when one or more of the processing mechanisms are compromised. However, although understanding might be maintained, increased signal degradation and/or auditory processing deficit can result in a corresponding increase in reported exertion to understand speech, as well as other longer term subjective problems such as fatigue (Kramer, Kapteyn, & Houtgast, 2006; Nachtegaal et al., 2009). It is generally presumed that this reported exertion is a result of the increased mental effort needed for speech comprehension under difficult conditions. Thus, similar speech intelligibility scores might be obtained under different listening conditions if the listener applies different levels of mental effort (Broadbent, 1958; Rabbitt, 1966, 1968). The mental (or cognitive) exertion applied to assist speech understanding in difficult conditions is often referred to as "listening effort" (McGarrigle et al., 2014).

It is theoretically reasonable to suggest that listening effort measures might not represent the same underlying variable as speech intelligibility scores, and there is some evidence to support this idea. Research by Rudner, Lunner, Behrens, Thorén, and Rönnberg (2012) demonstrated that individuals rated listening as progressively less effortful as signal-to-noise ratio (SNR) increased, even though speech recognition performance did not improve. In another study by Humes (1999), an analysis of hearing aid outcome parameters established that reported listening effort and measured speech intelligibility were separate hearing aid outcome domains. These results, together with theoretical considerations, support the notion that listening effort should not be inferred from speech intelligibility scores.

Previous researchers have used a variety of strategies intended to measure listening effort. Interested readers can

[a]Hearing Aid Research Laboratory, University of Memphis, Memphis, TN

Correspondence to Jani Johnson: jajhns10@memphis.edu.

Paul Pendergraft is now at The National Center for Rehabilitative Auditory Research, Portland Veterans Affairs Medical Center, Portland, OR.

Editor and Associate Editor: Larry Humes

Received October 16, 2014
Revision received May 4, 2015
Accepted May 17, 2015
DOI: 10.1044/2015_AJA-14-0058

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

refer to McGarrigle et al. (2014) for a detailed review of these strategies and the advantages and disadvantages of each. Measures can be grouped into three broad categories of outcomes: self-report, physiologic, and behavioral. Typical audiologists attempting to choose among these for use in an audiologic test battery often are limited in time and equipment available for this purpose. These limitations narrow the choices of listening effort measures that might be implemented in a typical audiology setting. However, there is no clear rationale for choosing among the remaining viable options. The purpose of the present study was to evaluate two measures of listening effort regarding their suitability for inclusion in a protocol that implements a comprehensive set of audiologic measures including both listening effort and speech intelligibility. Required characteristics of the listening effort measures were as follows: (a) efficiency (i.e., capable of being measured simultaneously with a speech intelligibility test); (b) ease of administration; (c) no specialized test equipment or expertise different from that of a typical audiologist; and (d) sensitivity to changes in listening difficulty presumed to affect listening effort.

Many validated measures of listening effort possess one or more of these characteristics; however, several of these were not considered for use in this research because they did not meet all of the required conditions described above. For example, some self-report measures could not be administered simultaneously with a speech intelligibility test; and all of the physiologic and many of the behavioral measures required specialized test equipment and expertise that a typical audiologist is not likely to possess. We selected two measures for evaluation in this research. They included a measure of self-reported listening effort by using a rating scale and performance on a word recall task using a dual-task paradigm.

Self-report measures of listening effort are made by asking listeners to respond to a question about how much effort is required to complete the listening task. Listeners typically respond by using a rating scale (e.g., Larsby, Hällgren, Lyxell, & Arlinger, 2005). This subjective method for assessing listening effort has good face validity, is efficient, and can be administered without specialized test equipment. Possible disadvantages of self-reported listening effort stem from the subjective nature of this measure. It has been suggested that disadvantages might include self-report bias, inconsistency of internal scale, or lack of attention or cooperation during the test procedure (McGarrigle et al., 2014).

Word recall performance measures are made by asking listeners to respond to word lists or sentences and to retain key words in memory. Then, they are periodically asked to recall the key words. The ability to respond to the speech as well as to recall the key words typically is measured in listening environments with varying levels of difficulty. For this type of dual-task paradigm, it is hypothesized that more cognitive resources must be allocated to the speech intelligibility task in more difficult listening environments, leaving fewer resources for the word recall task. Therefore, recalling fewer key words is interpreted as reflecting increased listening effort (e.g., Pichora-Fuller, Schneider, & Daneman, 1995; Rabbitt, 1966; Sarampalis, Kalluri, Edwards, & Hafter, 2009). This method for assessing listening effort has ecological validity because individuals often need to perform multiple tasks while listening in their daily lives. Further, word recall performance measures are efficient and can be administered without specialized test equipment. A potential disadvantage of the word recall paradigm is possible sensitivity to floor and ceiling effects. In addition, there is uncertainty about whether measures of the behavioral consequences of listening in difficult environments (e.g., changes in word recall performance) are a direct measure of mental effort (McGarrigle et al., 2014).

In the present study, the two selected measurement methods were compared with the goal of evaluating which was most effective for assessing listening effort while simultaneously assessing speech intelligibility. The following questions were asked: (a) Is speech intelligibility score independent of the measure of listening effort? (b) Do self-reported ratings and word recall measures of listening effort provide the same information? (c) Are both methods valid measures of listening effort? (d) Which method is more sensitive to changes in listening difficulty presumed to affect listening effort?

Method

Participants

Each potential participant responded to the question: "Overall, how much hearing difficulty do you have?" using the responses *none, mild, moderate, moderate-to-severe, severe*. To qualify for inclusion in the study, a self-rated hearing difficulty of *none* or *mild* was required. Persons with not more than mild self-reported hearing problems were selected as most appropriate for this study because their results allow for evaluation of the listening effort measures while avoiding issues related to differences in audibility. In addition, research has shown that mental exertion in difficult listening conditions occurs for normal hearers as well as for individuals with hearing impairment (Rakerd, Seitz, & Whearty, 1996; Sarampalis, Kalluri, Edwards, & Hafter, 2009). Thirty native speakers of American English participated. There were nine men and 21 women, ranging in age from 23 to 39 years (M = 26.6, SD = 4.8). After completion of the single 1.5-hr laboratory session, each participant received a $10 gift card. Procedures for this study were reviewed and approved by the institutional review board of the University of Memphis.

Test Material

Speech Intelligibility

The Revised Speech Perception in Noise Test (R-SPIN; Bilger, Nuetzel, Rabinowicz, & Rzeczkowski, 1984) was used for this study. The R-SPIN comprises eight lists of 50 sentences each. The participant is asked to repeat the last word of each sentence. For every list, half of the sentences contain contextual information that makes the last word

somewhat predictable (e.g., "The watchdog gave a warning growl."), whereas the other half do not contain contextual information (e.g., "I had not thought about the growl."). Each set of 25 key words is presented in a high-predictability (HP) context and a low-predictability (LP) context. A 12-talker babble track generated by Kalikow, Stevens, and Elliott (1977) is used as a masker. Details about R-SPIN test materials and implementation are provided by Bilger et al. (1984). For this study, the order of presentation of R-SPIN sentences was altered from the original test. Each original R-SPIN list was divided into two forms: one with HP sentences only and one with LP sentences only, so that measures of speech intelligibility and listening effort could be assessed separately for HP and LP sentences. This resulted in eight lists of 25 HP sentences and eight lists of 25 LP sentences. Two additional lists of 25 sentences containing both HP and LP sentences were extracted from a recording of unused sentences from the original Speech Perception in Noise Test (SPIN; Kalikow et al., 1977). These were used as practice lists. A portion of the 12-talker babble extracted from the R-SPIN noise track was used as the masker for the two practice lists. This babble also was saved and used for calibration purposes. Cuing phrases for all test sentences were removed. All audio files, including the reordered test lists, practice lists, and calibration signals, were saved and recorded onto a CD.

The difficulty of the speech intelligibility task was manipulated by varying key word context and SNR. It was expected that speech intelligibility scores would be higher for the HP key words and for each successively easier SNR condition.

Listening Effort

For the self-report rating of listening effort (RAT), participants used a seven-point scale (Figure 1) to rate how much effort it took for them to complete each list of R-SPIN sentences. This scale was based on that of Schulte et al. (2009), which was modified from Borg's general intensity Category scale with Ratio properties, numbered 0–10 (CR-10; Borg, 1990). For the word recall method of measuring listening effort (REC), participants were required to repeat groups of response words from the R-SPIN test. No additional materials were used.

Procedure

Participants completed a test of speech intelligibility in noise in a laboratory setting. Listening effort was assessed simultaneously. Two methods for measuring listening effort were used with every participant. These included a self-reported rating of listening effort and a word recall task. The procedures for the speech intelligibility test and word recall task were based on those described by Pichora-Fuller et al. (1995).

Test Environment

This work was conducted in the Hearing Aid Research Laboratory at the University of Memphis. Data were collected in a double-walled, sound-attenuating booth. The R-SPIN sentences and the 12-talker babble were presented by using a personal computer. Audio signals from the computer soundcard were routed through the two channels of a GSI-61 audiometer, amplified by a power amplifier, and then merged to one channel and delivered to a Boston Acoustics loudspeaker (CR55 or CR57). The loudspeaker was mounted on the wall of the sound booth with the center of its frontal surface at ear height.

Figure 1. Listening effort scale categories.

Listening Effort Scale
1. No effort
2. Very little effort
3. Little effort
4. Moderate effort
5. Considerable effort
6. Much effort
7. Extreme effort

Speech Intelligibility Test

Each listener was seated 1 m from a loudspeaker at 0° azimuth. The presentation level of R-SPIN sentences was fixed at 65 dB SPL (root-mean-square). The root-mean-square level of the 12-talker babble was varied to produce four SNRs: −4 dB, −2 dB, 0 dB, and 2 dB. All 16 R-SPIN test lists were administered to each participant. Eight lists (four HPs and four LPs) were administered with the RAT method, and the other eight lists (four HPs and four LPs) were administered with the REC method. List presentation alternated between HP and LP lists. The order in which the eight lists were presented within each contextual condition was randomized. In addition, the following conditions were counterbalanced across participants: HP/LP order, SNR order, and listening effort task order. The practice lists were administered at 2 dB SNR prior to the test lists to familiarize the participants with the speech intelligibility test and the two methods of measuring listening effort. The experimenter was seated outside the booth and could hear the participants' responses through a pair of headphones. After each sentence was presented, the participants repeated the last word as they believed they heard it. Participants were required to guess if they were unsure of the word. The experimenter then wrote down the response and marked it correct if it was identical to the presented word. The performance measure was number of words correct, giving a possible score of 0 to 25.

Listening Effort Rating

For the RAT measure, each participant was instructed as follows: "The sentences are divided into groups of 12

or 13. Every 5 minutes or so I will ask you to use the scale in front of you to rate how much effort it took for you to understand the group of words. If you think that the amount of effort was between two numbers on the scale, it is fine for you to pick a fraction." With this procedure, participants gave a rating score twice for each R-SPIN list. The two scores were averaged so that there was a single rating score per list. The possible score range for the RAT task was 1 to 7.

Word Recall Task

For the REC method, each participant was instructed as follows: "After every five sentences, you will be asked to repeat your last five answers. You may give your answers in any order." Responses were deemed correct when they were identical to the words previously reported for the speech intelligibility task. The possible score range for the REC task was 0 to 25.

Results

All statistical analyses were performed by using SPSS Version 21 software. General linear model repeated-measures analysis of variance (ANOVA) with planned contrasts was used to analyze these data except where noted. This approach has been shown to provide good statistical power while controlling the experiment-wise error rate (Rosenthal & Rosnow, 1985). For all analyses, any p value of .05 or lower was considered statistically significant, whereas p values between .05 and .1 were considered worthy of mention. Prior to all analyses, data were examined for distribution and outliers. Two of the total 1,024 data points were outliers. Following the recommendations of Tabachnick & Fidell (2007), each outlier value was adjusted to one unit more extreme than the next most extreme value in the distribution of that variable. All figures and analyses were based on these adjusted data.

Is Speech Intelligibility Score Independent of the Measure of Listening Effort?

One reason for selecting the RAT and the REC methods of measuring listening effort for evaluation in this study was that both measures allow for simultaneous assessment of speech intelligibility and listening effort. Thus, they are potentially efficient for inclusion in an audiologic testing protocol. However, because the testing protocol under development was intended to assess both variables simultaneously, it was possible that the method of listening effort assessment would affect measured speech intelligibility performance. Therefore, the first research aim was to explore whether the two measures of listening effort were associated with similar speech intelligibility performance. We investigated this by comparing speech intelligibility scores for the RAT and the REC methods of measuring listening effort. Substantial differences in speech intelligibility scores for the two listening effort measures would suggest that speech intelligibility data were affected by the listening effort measure.

Figure 2 shows mean speech intelligibility scores as a function of SNR, for HP and LP sentences. Data were obtained by using the RAT and REC methods of measuring listening effort. For both conditions of the listening effort measure, it was observed that mean speech intelligibility scores improved with increasing SNR. Also, mean scores were better when words could be predicted from contextual information provided in the sentence. For the HP condition, mean speech intelligibility performance was slightly better when listening effort was assessed by using the RAT method compared with using the REC method. For the LP condition, this trend was reversed for the three easiest SNRs.

These trends were explored statistically by using a separate within-subjects repeated measure ANOVA for each context condition. Speech intelligibility score was used as the dependent variable. The main effects, listening effort method (RAT and REC) and SNR (−4, −2, 0, and 2 dB), were treated as categorical variables. The results of these analyses are summarized in Table 1. For the HP sentences, the main effect of method was statistically significant, $F(1, 29) = 4.25$, $p < .05$, indicating that speech intelligibility performance was significantly better when listening effort was assessed by using the RAT method compared with using the REC method. In addition, the main effect of SNR was significant, $F(2.115, 61.337) = 304.291$, $p < .001$. To investigate the significance of differences between adjacent SNR means, mean speech intelligibility scores at adjacent SNRs were compared by using planned contrasts. These three contrasts were driven by the a priori hypothesis that speech intelligibility scores would be better at each easier SNR condition. The results of these comparisons showed that mean data obtained for each SNR condition were significantly different from mean data obtained at adjacent SNRs ($p < .001$). The interaction between method and SNR was not significant for the HP lists, $F(2.02, 58.593) = 0.396$, $p > .05$, indicating that differences in speech intelligibility

Figure 2. Average speech intelligibility scores (total number of words correct) for high-predictability (HP) and low-predictability (LP) lists under four signal-to-noise ratio (SNR) conditions when listening effort was measured using the listening effort rating (RAT) method and word recall (REC) method ($N = 30$). The score range is from 0 to 25. Note that a higher score indicates better performance.

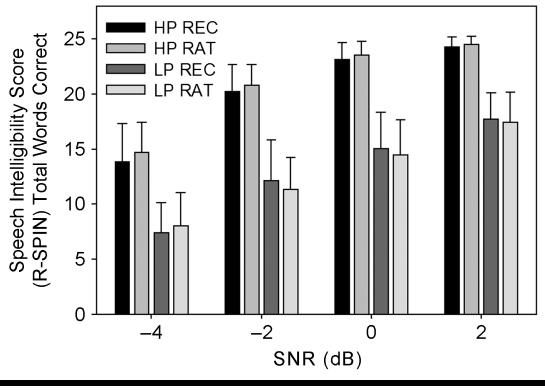


Table 1. Summary of two separate repeated-measures analyses of variance of speech intelligibility scores for high-predictability (HP) and low-predictability (LP) sentences when using the RAT and REC listening effort methods at four signal-to-noise ratios (SNRs).

Sentence type	Variables	*df*	*F*	*p*
HP	Method	1	4.25	.048*
	Error (Method)	29	(3.891)	
	SNR[a]	2.115	304.291	< .001*
	Error (SNR)[a]	61.337	(5.799)	
	Method × SNR[a]	2.02	0.396	.677
	Error (Method × SNR)[a]	58.593	(3.639)	
LP	Method	1	0.553	.463
	Error (Method)	29	(7.237)	
	SNR	3	164.574	< .001*
	Error (SNR)	87	(6.523)	
	Method × SNR	3	0.581	.629
	Error (Method × SNR)	87	(9.687)	

Note. Values enclosed in parentheses represent mean square errors. RAT = self-report rating of listening effort; REC = word recall method of measuring listening effort.

[a]With Greenhouse-Geisser adjustment.

*Statistically significant at .05 level.

score for the two listening effort measures were maintained across the four SNRs.

For the LP sentences, the main effect of method was not significant, $F(1, 29) = 0.553$, $p > .05$. Again, the main effect of SNR was statistically significant, $F(3, 87) = 164.574$, $p < .001$, and planned contrasts showed that mean data obtained at each SNR were significantly different from mean data obtained at adjacent SNRs ($p < .001$). As seen for the HP analysis, the interaction between method and SNR was not significant, $F(3, 87) = 0.581$, $p > .05$.

Do Self-Reported Ratings and Word Recall Measures of Listening Effort Provide the Same Information?

Both self-reported ratings and word recall measures have been used in previous research studies as methods of measuring listening effort. If both methods assess the same theoretical construct (underlying listening effort), then one would expect similar results from both measures. To investigate the relationship between the two listening effort measures, correlations were completed between RAT and REC scores for each of the eight combinations of SNR and context. If both methods quantify the same underlying variable, we would expect them to be inversely correlated. In other words, correlations should show that sentences that were rated as requiring higher levels of listening effort using the RAT method also should show fewer words correctly recalled by using the REC method. Table 2 shows the Pearson *r* and Spearman ρ correlation values for all conditions. All correlation coefficients indicate a weak or no relationship between scores obtained with the two measurement strategies. We examined the scatterplots of RAT and REC scores to gain a deeper understanding of these relationships. Although these figures are not presented in this article, they demonstrated a reasonable range of scores for both variables, with no patterns suggesting that the data were associated in a way that might not be detected by using a correlation coefficient. Furthermore, none of the adjusted distributions were significant for skewness, kurtosis, or influential outlier data. Therefore, it is reasonable to interpret the low correlation results as indicating that the two measures did not provide equivalent information with regard to listening effort.

Table 2. Correlation between the two methods of measuring listening effort across high-predictability (HP) and low-predictability (LP) lists. None of the correlation analyses was statistically significant.

Sentence type	Statistic	SNR (dB)			
		–4	–2	0	2
HP	Pearson's *r*	–.026	–.111	–.025	.058
	Spearman's ρ	–.096	–.072	–.049	.123
LP	Pearson's *r*	–.21	–.061	.114	.018
	Spearman's ρ	–.199	–.123	.135	.062

Note. SNR = signal-to-noise ratio.

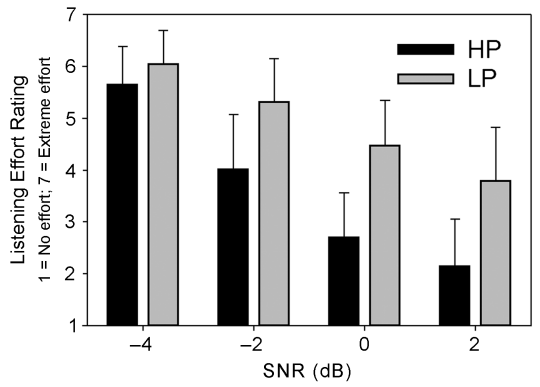
Are Both Methods Valid Measures of Listening Effort?

A valid measure of listening effort should be sensitive to changes in the degree of mental exertion expended in varying listening conditions. In this study, we manipulated task difficulty through changes in SNR and linguistic context. The R-SPIN scores confirmed our expectation that sentences presented at poorer SNRs were more difficult to understand and sentences with reduced linguistic context were more difficult to understand. Therefore, we would expect a measure of listening effort to reflect more effort (higher ratings or fewer words recalled) at more difficult SNRs and with lower context sentences.

Self-Reported Rating (RAT)

Figure 3 shows the average self-report rating of listening effort as a function of SNR, when tested by using HP and LP sentences. As predicted, participants reported greater listening effort on average as SNR became more difficult and for sentences with lower-predictability. Further, it can be seen that the differences in self-reported effort between HP and LP sentences were greater at easier SNRs. These data were explored by using a within-subjects repeated-measures ANOVA with listening effort rating as the dependent variable, and context (HP and LP) and SNR (−4, −2, 0, and 2 dB) as categorical variables. A summary of the ANOVA is shown in Table 3. Results of this analysis showed that the main effect of context was statistically significant, $F(1, 29) = 167.095$, $p < .001$. This indicated that participants reported significantly more listening effort when attempting to understand sentences with less linguistic context. The main effect of SNR also was statistically

Figure 3. Average scores on listening effort rating (RAT) for high-predictability (HP) and low-predictability (LP) lists under four signal-to-noise ratio (SNR) conditions. The score range is from 1 to 7. Note that a higher score indicates more listening effort.



significant, $F(3, 87) = 202.374$, $p < .001$. Planned contrasts were driven by the a priori hypothesis that listening effort would be rated lower at easier SNR conditions. The results of these three contrasts showed that mean data obtained for each SNR condition were significantly different from data obtained at adjacent SNRs ($p < .001$). The interaction between SNR and context also was statistically significant, $F(3, 87) = 15.025$, $p < .001$. This suggested that the amount of linguistic context available in the sentences had a different effect on mean listening effort ratings depending on the SNR at which sentences were presented. Despite this significant interaction, follow-up t test comparisons showed statistically significant differences between rating scores for HP and LP sentences for all four SNRs ($p < .005$). To further explore the interaction, the differences in listening effort ratings between HP and LP conditions were computed for each participant at each of the four SNRs. Mean difference scores were as follows: at −4 dB SNR, X = 0.785 (SD = 1.3); at −2 dB SNR, X = 2.62 (SD = 2.18); at 0 dB SNR, X = 3.54 (SD = 1.69); at +2 dB SNR, X = 3.29 (SD = 2.17). These difference scores were compared by using a one-way ANOVA with follow-up pairwise comparisons using Šidák correction (Šidák, 1967; experiment-wise $\alpha = .05$). The main effect of SNR was significant, $F(3, 67.536) = 20.606$, $p < .001$,[1] indicating that the mean difference scores varied across SNR conditions. Follow-up testing was conducted by using post hoc pairwise comparisons with Šidák correction to control for experiment-wise error (Šidák, 1967). These results indicated that the mean difference at −4 dB SNR was significantly smaller than the mean difference at any of the other SNRs ($p < .05$). However, the mean differences for the three easiest SNRs were not significantly different from one another. These results demonstrated that, on average, the addition of linguistic context reduced rated listening effort at each of the four tested SNRs. However, at the most difficult SNR, the benefit obtained from the addition of linguistic context was significantly smaller than at the three easier SNRs.

Table 3. Summary of two separate repeated measure analyses of variance of listening effort scores for the RAT and REC listening effort methods, using high predictability and low predictability sentences at four signal-to-noise ratios (SNRs)

Listening effort task	Variables	*df*	*F*	*p*
RAT	Context	1	167.095	< .001*
	Error (Context)	29	(2.35)	
	SNR	3	202.374	< .001*
	Error (SNR)	87	(1.895)	
	Context × SNR	3	15.025	< .001*
	Error (Context × SNR)	87	(1.547)	
REC	Context	1	19.629	< .001*
	Error (Context)	29	(1.997)	
	SNR	3	10.152	< .001*
	Error (SNR)	87	(3.724)	
	Context × SNR	3	1.803	.153
	Error (Context × SNR)	87	(4.138)	

Note. Values enclosed in parentheses represent mean square errors. RAT = self-report rating of listening effort; REC = word recall method of measuring listening effort.

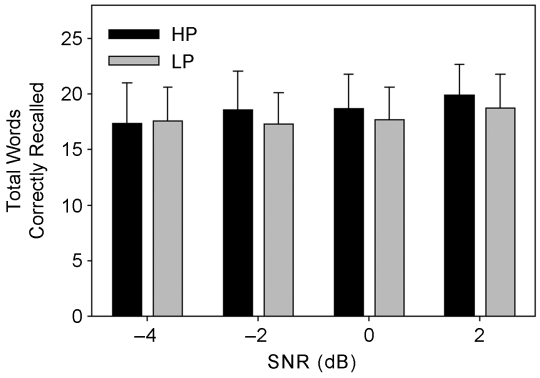
*Statistically significant at .05 level.

Word Recall Method (REC)

Figure 4 shows the average number of words correctly recalled as a function of SNR, for HP and LP sentences. It can be observed that mean word recall performance improved slightly with increasing SNR. Also, at the three easiest SNRs, word recall performance was slightly better for HP words. A within-subjects repeated-measures ANOVA was performed with total number of words correctly recalled as the dependent variable and context (HP and LP) and SNR (−4, −2, 0, and 2 dB) as categorical variables. Results are summarized in Table 3. Results of this analysis showed that the main effect of context was statistically significant, $F(1, 29) = 19.629$, $p < .001$). As with the RAT method, this finding demonstrated that average participants experienced more listening effort when attempting to understand sentences with less linguistic context. The main effect of SNR also was statistically significant, $F(3, 87) = 10.152$, $p < .001$. Planned contrasts of word recall scores obtained at adjacent SNRs were driven by the a priori hypothesis that word recall scores would be higher at easier SNR conditions. The results of these three contrasts indicated that mean word recall scores obtained at 2 dB SNR were significantly different from mean word recall scores obtained at 0 dB SNR ($p < .005$). However, comparisons of mean scores at −4 and −2 dB SNR, and −2 and 0 dB SNR were not statistically different. It is worth noting that the small differences in mean scores across the four SNRs trended in the same direction as with the RAT method. However, with the RAT method, listening effort was significantly different for all three adjacent SNR pairs, whereas with the REC method

[1]Welch's F was used because Levene's test of homogeneity was significant.

Figure 4. Average scores on the word recall (REC) method for high predictability (HP) and low predictability (LP) lists under four signal-to-noise ratio (SNR) conditions. The score range is from 0 to 25. Note that a higher score indicates less listening effort.



the significant main effect of SNR was due to the difference in mean word recall scores at 0 and 2 dB SNR only. The interaction between context and SNR did not reach statistical significance, $F(3, 87) = 1.803$, $p > .05$.

Which Method Is More Sensitive to Changes in Listening Conditions That Are Presumed to Affect Listening Effort?

To answer the question of which method was more sensitive to changes in listening demand, it was necessary to directly compare the two measures. However, the RAT method used a seven-point scale, and the REC method used a 26-point scale. Thus, direct comparison of performance differences for the two measures might be misleading. Direct comparison across RAT and REC data required a standardized measure of the performance differences obtained across contexts and SNR conditions with each measuring method. This was achieved by using an effect size analysis (Cohen, 1988). Effect sizes provide a standardized answer to the question "How big is the difference between two conditions?" Thus, effect size values allow for direct comparison of performance differences across measures.

There are a variety of ways to compute an effect size. Interested readers can refer to Lipsey and Wilson (2001) for details. In the present study, the effect size known as Cohen's *d* was used. The method used for this calculation was not unique to this study; however, a review of this computation is provided here. The equation for computing the value is shown in Equation 1.

$$d = \frac{\overline{X_1} - \overline{X_2}}{S} \tag{1}$$

In this equation $\overline{X_1}$ and $\overline{X_2}$ are the means of the two groups being compared. S is the pooled standard deviation of the two groups. There are several ways to compute S. A common method for computing S, and the one used for this study, is shown in Equation 2.

$$S = \frac{(N_1 - 1) \times SD_1 + (N_2 - 1) \times SD_2}{N_1 + N_2 - 2} \tag{2}$$

In this equation, N_1 and N_2 are the number of observations in the two groups. In our case, they are the number of research participants. SD_1 and SD_2 are standard deviations of the two groups.

An example is provided here to demonstrate how to use the two equations to compute Cohen's *d* effect sizes. The data used in this demonstration are RAT scores obtained with HP and LP sentences at −4 dB SNR. In this comparison, rating values for LP sentences are $\overline{X_1} = 6.05$, $SD_1 = .64$, and $N_1 = 30$. Rating values for HP sentences are $\overline{X_2} = 5.66$, $SD_2 = .72$, and $N_2 = 30$. With Equation 2, the pooled standard deviation (S) is computed and the result is .68. Then, the computed pooled standard deviation (S) and the two means ($\overline{X_1}$ and $\overline{X_2}$) are used in Equation 1. The resulting Cohen's *d* is 0.57.

Table 4 shows the effect sizes for score differences between HP and LP sentences at each SNR for each listening effort measurement method. It was observed that the effect sizes for the four SNRs for data obtained by using the REC method were between .07 and .41, whereas the corresponding values for data obtained by using the RAT method were between .57 and 2.07. Table 5 shows the computed effect sizes for score differences between adjacent SNRs for HP and LP sentences for each listening effort measurement method. It was observed that for both types of sentences, the effect sizes for data obtained by using the REC method were between .03 and .42, whereas the corresponding values for data obtained by using the RAT method were between .64 and 1.85. Although there is not a consensus in interpreting values of Cohen's *d*, conventional interpretations suggest that values less than 0.2 represent a small effect; values around 0.5 are considered a moderate effect; and values of 0.8 or more represent a large effect. Using these conventions, it can be seen that data from the RAT method reflected primarily large effects as listening demand varied across contexts and

Table 4. Effect sizes (Cohen's *d*) of the score differences between high predictability and low predictability sentences when using the two measures of listening effort. All effect sizes are absolute values.

Listening effort task	SNR (dB)			
	−4	−2	0	+2
REC	0.07	0.41	0.33	0.4
RAT	0.57	1.39	2.07	1.71

Note. SNR = signal-to-noise ratio; REC = word recall method of measuring listening effort; RAT = self-report rating of listening effort.

Table 5. Effect sizes (Cohen's *d*) of the score differences between adjacent signal-to-noise ratios (SNRs) for high predictability (HP) and low predictability (LP) sentences when using the two measures of listening effort. All effect sizes are absolute values.

Sentence type	Listening effort task	SNR (dB)		
		–4 vs. -2	–2 vs. 0	0 vs. 2
HP	REC	0.35	0.03	0.42
	RAT	1.85	1.37	0.64
LP	REC	0.10	0.14	0.36
	RAT	0.99	1.00	0.72

Note. REC = word recall method of measuring listening effort; RAT = self-report rating of listening effort.

adjacent SNRs, whereas data from the REC method demonstrated primarily small effects (Tables 4 and 5). This finding suggests that the RAT method was more sensitive to changes in listening difficulty presumed to affect listening effort.

Discussion

Individuals with hearing impairment report fatigue and stress more frequently than individuals with no hearing impairment (Hétu, Riverin, Lalande, Getty, & St-Cyr, 1988; Kramer et al., 2006). It has been presumed that these real-world complaints are the result of increased listening effort necessary for individuals with hearing impairment to communicate successfully in day-to-day listening environments. However, there has been a lack of consensus about the definition of listening effort. An article from the British Society of Audiology proposed to define listening effort as "the mental exertion required to attend to, and understand, an auditory message" (McGarrigle et al., 2014). At this time, there is no agreement on the best method for measuring mental exertion. Both self-report and behavioral measures are presumed to be indicative of the mental exertion underlying the clinical presentation of listening effort. Interpretation of results from these measures requires that researchers rely on assumptions about each methodology. For self-report measures, it is assumed that the experience of listening effort will be accurately perceived and recognized; and for behavioral measures such as the dual-task paradigm, that mental exertion during listening is a product of burdening a limited-capacity system, and that this exertion will result in poorer performance on a secondary task (McGarrigle et al., 2014). Although it is still unclear how these measures are related to the construct of listening effort or mental exertion, both types of measures have the potential to serve as a clinical evaluation tool. This research adds to the listening effort literature by comparing results from examples of self-report and behavioral listening effort measures under varying levels of listening demand, and evaluating the two measures regarding their suitability for inclusion in an audiologic test battery that simultaneously assessed listening effort and speech intelligibility performance.

Effects of Listening Effort Measure on Speech Intelligibility Performance

Both the RAT and the REC methods measured listening effort during pauses at set intervals throughout a speech intelligibility task. Simultaneous administration of the speech intelligibility and listening effort measures results in reduced total administration time compared with obtaining results for each domain consecutively. A reduction of administration time is desirable when numerous outcome domains are under investigation in a single research session. Further, such time-saving methods are key factors for clinicians when they consider how to evaluate audiologic outcomes cost-effectively.

When speech intelligibility and listening effort measures are administered concurrently, listeners are typically instructed to prioritize the speech intelligibility task. Varying the difficulty of the speech intelligibility task is presumed to influence performance on the listening effort measure. In fact, the REC and other dual-task methods of measuring listening effort are based on this premise. However, researchers have demonstrated that attempting two tasks simultaneously can interfere with performance on both tasks, even when one is prioritized over the other (e.g., Johnston, Greenberg, Fisher, & Martin, 1970; Trumbo & Noble, 1970). This suggests that the inclusion of a listening effort measure that is administered simultaneously with a speech intelligibility measure might negatively affect performance in the speech intelligibility domain. In an application where both speech intelligibility and listening effort are measured concurrently, it is important to evaluate the accuracy of both types of data.

The REC method requires silent word rehearsal throughout the speech intelligibility task; however, the RAT method requires only that participants reflect on and rate their perceived effort subsequent to the speech intelligibility task. Therefore, it would be reasonable to hypothesize that the REC method might result in more interference with speech intelligibility performance than the RAT method. This outcome was observed in this study when sentences had higher levels of linguistic context (Figure 2, Table 1). However, the type of method used to measure listening effort had only a small effect on speech intelligibility performance. Although statistically significant, this small difference might not be of practical importance when evaluating audiologic outcomes. Note that we did not evaluate speech intelligibility performance compared with a baseline score using no listening effort measure. As a result, we do not know the absolute effect of these two measures on speech intelligibility performance, although it seems unlikely that the RAT method would have any effect. To our knowledge, this is the first study that has included an evaluation of the comparative effect of different listening effort measures on speech intelligibility performance as a way to

determine the methods' suitability for use in a comprehensive set of audiologic outcome measures.

Relationship Between Self-Report Ratings and Word Recall Measures of Listening Effort

In this study, we manipulated mental exertion through changes in SNR and linguistic context. Both the RAT and REC measures of listening effort reflected more effort (higher ratings, or fewer words recalled) at more difficult SNRs, and with lower context sentences (Figures 3 and 4, Table 3). This finding confirmed that both methods assessed a trait related to listening effort. However, the weak association between RAT and REC data indicated that the two measures did not assess the same underlying variable (Table 2). This finding is consistent with other studies that have included both self-report and behavioral measures of listening effort (e.g., Downs & Crum, 1978; Feuerstein, 1992; Fraser, Gagne, Alepins, & Dubois, 2010; Hicks & Tharpe, 2002; Larsby et al., 2005). Because these types of measures have been shown to assess different aspects of listening effort, some researchers have recommended the inclusion of both self-report and behavioral measures when attempting to obtain a complete picture of outcomes in the listening effort domain (e.g., Anderson Gosselin & Gagne, 2011; Larsby et al., 2005). However, including multiple methods to measure listening effort in an already lengthy testing protocol might not be optimal. For our purposes, it was necessary to select between the two methods on the basis of a comparison of each measure's suitability for use in such a protocol.

Validity for Measuring Listening Effort

A valid measure of listening effort should reflect changes in the degree of mental exertion expended in varying listening conditions. Because speech intelligibility scores confirmed our expectation that speech presented at poorer SNRs was more difficult to understand and sentences with reduced linguistic context were more difficult to understand, we expected the measures of listening effort to reflect more effort at more difficult SNRs, and with lower context sentences. These expectations were realized for the RAT method (Figure 3, Table 3) and, to a lesser extent, for the REC method (Figure 4, Table 3). Based on the result that mean listening effort data were in the same direction for both the RAT and REC methods across contexts and SNRs, some might assert that, statistically, both methods demonstrated expected changes, and therefore both were valid measures of listening effort. However, despite the statistical similarity, it was observed that, compared with the RAT method, the differences in mean performance on the REC measure were relatively small across contexts and SNRs.

Sensitivity to Changes in Listening Demand

The procedures for the speech intelligibility test and the REC method were based on those described by Pichora-Fuller et al. (1995), and reproduced by Sarampalis et al. (2009). Our research extended the design of these studies by also including a self-report measure of perceived listening effort. Similar to the present study, both of those research groups explored the effects of linguistic context and SNR on listening effort by using a word recall task. Like those studies, scores obtained by using the REC method in the present research showed increased listening effort when SNR decreased and with reduction of linguistic context (Figure 4, Table 3). The self-report data demonstrated results in the same direction as the REC method (Figure 3, Table 3). However, our results indicated that the RAT method was considerably more sensitive than the REC method to changes in listening demand across context levels and SNRs (Tables 4 and 5). This suggests that, compared with scores obtained by using the REC method, scores obtained by using the RAT self-report method are more likely to reflect differences in listening effort when differences exist.

Additional Considerations About Subjective Self-Report and Dual-Task Measures of Listening Effort

Taken together, the results of these analyses suggest that RAT data are better for inclusion in an audiologic testing protocol compared with REC data. However, additional theoretical considerations have been raised concerning the use of subjective self-report and dual-task methods to measure listening effort. Also, some practical concerns were observed during data collection for this study. These additional considerations were worthy of further exploration to inform our evaluations of each of the measurement options.

Theoretical Considerations

First, although subjective reports have been recognized as valuable and practical for providing important information about listening effort, especially in a hearing clinic (e.g., Hällgren, Larsby, Lyxell, & Arlinger, 2005; Humes, 1999; Rudner et al., 2012; Wingfield, 2014), there is some evidence that self-report measures of listening effort might not be reliable for between-group comparisons, such as comparison of data from young and old adults. This possibly is due to groups having differing concepts about how to rate effort in a listening task. For example, Larsby et al. (2005) and Anderson Gosselin and Gagné (2010) both found evidence that older adults might underestimate the perceived effort needed for a listening task compared with younger adults, even when they performed similarly or even worse on the task than the younger listeners. However, Anderson Gosselin and Gagné (2010) also demonstrated that subjective rating scales were reliable when making within-subject comparisons of listening effort for the older adults. This within-subject comparison is the application that a researcher or clinician would most likely implement when measuring listening effort as an audiologic outcome.

Second, despite the straightforward approach and seeming suitability of the self-report measure of listening

effort, some researchers mistrust the results of self-report measures because they are susceptible to issues such as bias and inconsistency of internal scale. In this view, behavioral measures of listening effort have greater value because they are assumed to avoid some of these potential problems. As a result, weak correlations between behavioral and self-report measures have led to conclusions that the self-report measures of listening effort did not measure listening effort, or mental exertion (e.g., Anderson Gosselin & Gagné, 2011; Downs & Crum, 1978; Feuerstein, 1992). For example, in a study by Downs and Crum (1978), researchers compared results of a dual-task measure and self-reported rating of effort for a learning task. They found that the self-reported measure correlated with performance on the task, but not with the dual-task results. These researchers concluded that the participants based their self-reported ratings on how well they thought that they performed on the learning task, rather than judging their mental exertion during the task. Similar findings prompted Feuerstein (1992) to conclude that self-report measures of listening effort might indicate individuals' perceptions of effort in a given listening situation, but that these perceptions did not reflect actual demand on cognitive resources. If true, this would reduce the value of self-reported listening effort data.

Similar to Downs and Crum (1978) and Feuerstein (1992), our data demonstrate patterns of speech intelligibility and listening effort rating results that are in the same direction. Given our study's design, with speech intelligibility and listening effort being assessed in situations with varying degrees of difficulty, this result was expected. To explore the hypothesis that the self-reported estimates of listening effort are instead self-reported estimates of speech intelligibility performance, we compared the patterns of the mean RAT data with those of the corresponding speech intelligibility data. If the RAT results were based solely on estimates of speech intelligibility performance, then we would expect the RAT data to follow the same patterns as the speech intelligibility data. On the other hand, if the two measures assessed different variables, then we would expect instances where the patterns of the results might diverge. For this exploration, results of a within-subjects repeated-measures ANOVA of speech intelligibility scores were compared with the results of the previously described analysis of self-reported listening effort (see Table 3). The two analyses yielded parallel results for significance of the two main effects and the interaction. However, the pattern of differences between HP and LP conditions for the four SNRs was different for RAT data and speech intelligibility data. RAT data showed that for the three best SNR conditions (when the speech was sufficiently audible to support the use of top-down linguistic cues), listeners reported a consistent improvement in listening effort as a result of contextual cues. For the speech intelligibility data, improved performance due to contextual cues was consistent for the two middle SNRs but substantially reduced for the easiest SNR. This finding is consistent with the report by Rudner et al. (2012), who demonstrated that perceived listening effort continued to improve with increasing SNR even when speech intelligibility performance had reached a ceiling. Taken together, our findings and those of Rudner et al. (2012) suggest that self-reported listening effort ratings are not merely surrogate estimates of speech intelligibility performance. Additional research is needed that clearly demonstrates differences between listening effort and speech intelligibility measures.

Third, it has been hypothesized that changes in listening effort that result from varying audibility are more likely to occur when top-down cognitive processes are engaged (Sarampalis et al., 2009). The use of R-SPIN (Bilger et al., 1984) materials and the design of our study allowed us to investigate whether the two measures of listening effort produced results in alignment with this hypothesized outcome. By evaluating changes in listening effort across SNRs separately for sentences high in linguistic context and low in linguistic context, we were able to control at least one component of higher-level processing involvement for speech intelligibility. When the sentences were low in linguistic context, listeners were forced to rely on audibility-based bottom-up processing to understand the speech signal. When the sentences were rich in context, top-down processing also was engaged to utilize the contextual information and assist in speech intelligibility (Janse & Ernestus, 2011). Utilizing a combination of bottom-up and top-down resources can make keywords easier to predict across all SNRs compared with when there is linguistic context. This was illustrated through better speech intelligibility scores for HP sentences at all SNRs (Figure 2). Further, the significant main effects of context for both the RAT and REC methods demonstrated less listening effort for HP sentences, providing additional support for this notion (Table 3). However, for the higher-context sentences, as the speech signal became increasingly degraded and fewer bottom-up resources were available, listeners theoretically should have allocated increasingly more top-down resources to process the signal by using available contextual information. Pichora-Fuller et al. (1995) hypothesized that as a signal becomes less audible and reliance on contextual information increases, listening becomes more effortful. Sarampalis et al. (2009) extended this concept to suggest that, because the use of linguistic context is a top-down and effortful process, changes in performance on a listening effort measure would be more likely to occur when linguistic context was available. We explored this idea by using results obtained from the RAT and REC listening effort methods. Based on the suggestion of Sarampalis et al. (2009), we expected the measures to reflect not only greater listening effort at more difficult SNRs, but also greater differences in listening effort between the easier and more difficult SNRs when contextual information was available to process the speech signal (HP sentences) compared with when there was minimal context (LP sentences).

Data from the REC method, which had no significant interaction between context and SNR, did not support this hypothesis (Table 3). This replicated the findings of Pichora-Fuller et al. (1995) and Sarampalis et al. (2009), who also evaluated REC results. However, listening effort

data from the RAT method did demonstrate the hypothesized relationship (Figure 3, Tables 3 and 5). Prior to this study, this relationship had not been explored with the RAT results. This outcome lends further support for the notion that self-reported listening effort is reflective of mental exertion at the resource level.

Practical Considerations

During data collection and analysis for this research, it was noted that the REC method was subject to some practical concerns that have not been noted in previous research.

First, discrepancies in word recall and speech intelligibility performance were observed for the REC method at the most adverse SNRs, especially for sentences with low linguistic context. In these unfavorable listening conditions, some participants seemed unable to utilize adequate contextual or audibility cues to speculate about the keyword. This was demonstrated by the poor speech intelligibility scores that often were obtained under these conditions (Figure 2). This result was not surprising given that ecologically realistic speech intelligibility testing is expected to result in poor scores for some listeners. However, for this protocol, listeners were required to provide a response for the key word for each sentence, regardless of their level of certainty about what that word might be. For the REC method, credit was given for accurately recalling responses whether they had been correct or not. Careful scrutiny of the raw data suggested that guessed words might sometimes have been words that the participants could easily remember. As a result, for these participants, in these conditions, scores for the recall task were relatively high, which did not reflect the difficulty of the speech intelligibility task. This tendency can be observed in Figure 2 for the REC method at −4 dB SNR. At this unfavorable SNR, participants had poorer speech intelligibility scores on average for LP sentences, correctly identifying approximately seven out of 25 total words, compared with approximately 14 total words for HP sentences. However, at this SNR, scores on the word recall task indicated that average listeners were able to correctly recall even more of their response words for LP sentences ($x = 17.5$) than for HP sentences ($x = 17$). Therefore, using a memory task as the secondary task in a listening effort measure was not valid when speech intelligibility scores were very low.

Second, it was observed that some participants did not do well on the word recall task regardless of the listening condition, whereas some performed well in all test conditions. This is not consistent with the dual-task rationale. Like other dual-tasks, the REC paradigm relies on the assumption that working memory has a limited capacity (Kahneman, 1973). It is assumed that the increased mental exertion needed to process speech while listening under difficult conditions consumes a large proportion of that limited capacity, leaving less capacity available for word storage. Therefore, listening under increasingly difficult conditions is expected to result in successively poorer word recall performance. Although that pattern was observed for the mean scores, it was not seen for all participants. There were no obvious reasons why some participants performed poorly and some performed well across conditions. We speculate that this might be attributed to differences in working memory capacity among the participants, or in the way that participants used their working memory capacities for the REC task. Research has demonstrated that individuals differ substantially in their working memory capacity (e.g., Conway & Engle, 1996). Individuals with greater working memory capacity have a greater ability to store and process information. In the present study, we did not obtain information about participants' cognitive capacities. However, it is reasonable to speculate that those who did well on the word recall task regardless of the listening condition might have working memory capacities that exceeded the requirements for both the listening and the word recall tasks, even for the most difficult listening condition that we tested. In contrast, those who did poorly on the word recall task in all test conditions might have working memory capacities that were insufficient to manage the word recall task in any condition after allocating most of their capacity to the listening task. Note also that the dual-task paradigm for measuring listening effort relies on the assumption that a listener will divide all of their cognitive resources between the two concurrent tasks (McGarrigle et al., 2014). However, this assumption has not been proved and might not be tenable for every listener. After devoting a certain amount of cognitive resources to the primary speech intelligibility task, it is possible that some listeners might only use a portion of the remaining cognitive resources available for the secondary word recall task while reserving some for other mental or physical activities not related to the recall task. This might contribute in part to the systematically poor performance observed for some participants on the word recall task. So far, there is no way to manage or measure the amount of cognitive resources that are dedicated to the two simultaneous tasks. Evaluation of these theoretical and practical considerations strengthened the rationale for using the RAT method for measuring listening effort in a hearing aid outcome testing protocol.

Conclusions

Individuals with hearing impairment report that increased effort is needed to listen carefully and to concentrate in their daily listening environments. It is this perception of effort that leads to clinical complaints, and is associated with increased stress and fatigue in daily living. Because one goal of audiologic testing and intervention is to address challenges in daily living, it seems reasonable that reduced listening effort would be an important and valuable outcome. The purpose of the present study was to evaluate two types of listening effort measures (a self-report measure [RAT] and a word recall measure [REC]) regarding their suitability for inclusion in an audiologic testing protocol that included measures of listening effort and speech intelligibility. Both listening effort measures were evaluated with

regard to validity, sensitivity, and effect on speech intelligibility performance. Our findings revealed that, on a statistical level, both types of listening effort measures were capable of demonstrating results that reflected expected changes in effort with corresponding changes in listening demand. However, our results indicated that data obtained by using the RAT method were considerably more sensitive to these changes than data obtained by using the REC method. This finding not only has an important effect on developing a clinical audiological test battery, but also has significant implications for evaluation of audiologic interventions such as hearing aids. Given that typical outcome differences with different hearing aids and hearing aid technologies are subtle, a sensitive measure is required to detect differences when they exist. Furthermore, our findings demonstrated that, under some conditions, implementation of the REC method might result in more interference with speech intelligibility performance than the RAT method when both outcomes are measured simultaneously. In addition to the research questions, some theoretical concerns were explored. The results of these explorations provided additional support for the validity of subjective ratings of listening effort for use in the context of audiologic assessment. It was observed that, for some participants, listening effort scores obtained by using the REC method were complicated by floor and ceiling effects. Based on these results, the RAT method was deemed more appropriate for inclusion in an audiologic testing protocol than the REC method, especially when simultaneously assessing speech intelligibility.

Acknowledgments

This research was funded in part by National Institute on Deafness and Other Communication Disorders Grant RO1 DC011550, awarded to the third author.

References

Anderson Gosselin, P., & Gagné, J.-P. (2010). Use of a dual-task paradigm to measure listening effort. *Canadian Journal of Speech-Language Pathology and Audiology, 43,* 43–51.

Anderson Gosselin, P., & Gagne, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54,* 944–958.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research, 27,* 32.

Borg, G. (1990). Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian Journal of Work, Environment, and Health, 16*(Suppl. 1), 55–58.

Broadbent, D. E. (1958). *Perception and communication*. London, England: Pergamon.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: more evidence for a general capacity theory. *Memory (Hove, England), 4,* 577–590.

Davis, H. (1964). International audiology. In H. Davis & R. Silverman (Eds.), *Hearing and deafness* (3rd ed., p. 76). New York, NY: Holt, Rinehart & Winston.

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research, 21,* 702–714.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort. *Ear and Hearing, 13,* 80–86.

Fraser, S., Gagne, J. P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research, 53,* 18–33.

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *International Journal of Audiology, 44,* 574–583.

Hétu, R., Riverin, L., Lalande, N., Getty, L., & St-Cyr, C. (1988). Qualitative analysis of the handicap associated with occupational hearing loss. *British Journal of Audiology, 22,* 251–264.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research, 45,* 573–584.

Humes, L. (1999). Dimensions of hearing aid outcome. *Journal of the American Academy of Audiology, 10,* 26–39.

Janse, E., & Ernestus, M. (2011). The roles of bottom-up and top-down information in the recognition of reduced speech: Evidence from listeners with normal and impaired hearing. *Journal of Phonetics, 39,* 330–343.

Johnston, W. A., Greenberg, S. N., Fisher, R. P., & Martin, D. W. (1970). Divided attention: A vehicle for monitoring memory processes. *Journal of Experimental Psychology, 88,* 164–171.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kalikow, D., Stevens, K., & Elliott, J. (1977). The speech perception in noise (SPIN) test. *The Journal of the Acoustical Society of America, 61,* 1337–1351.

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International Journal of Audiology, 45,* 503–512.

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology, 44,* 131–143.

Lipsey, M. W., & Wilson, D. (2001). *Practical meta-analysis (applied social research methods)*. Thousand Oaks, CA: Sage.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1), 71–102.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology, 53,* 433–440.

Nachtegaal, J., Kuik, D. J., Anema, J. R., Goverts, S. T., Festen, J. M., & Kramer, S. E. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an Internet-based national survey on hearing. *International Journal of Audiology, 48,* 684–691.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America, 97,* 593–606.

Rabbitt, P. M. (1966). Recognition: Memory for words correctly heard in noise. *Psychonometric Science, 6,* 383–384.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *Quarterly Journal of Experimental Psychology, 20,* 241–248.

Rakerd, B., Seitz, P., & Whearty, M. (1996). Assessing the cognitive demands of speech listening for people with hearing losses. *Ear and Hearing, 17*(2), 97–106.

Rosenthal, R., & Rosnow, R. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, England: Cambridge University Press.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology, 23,* 577–589.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research, 52,* 1230–1240.

Schulte, M., Vormann, M., Wagener, K. C., Buchler, M., Dillier, N., Dreschle, W., . . . Wouters, J. (2009). *Listening effort scaling and preference rating for hearing aid evaluation*. Paper presented at the HearCom Workshop on Hearing Screening and Technology, Brussels, Belgium. PowerPoint slides retrieved January 20, 2015, from http://hearcom.eu/lenya/hearcom/authoring/about/DisseminationandExploitation/Workshop/S2B-3_Michael-Schulte_Hearing-Aid-Scaling-Rating.pdf

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson/Allyn & Bacon.

Trumbo, D., & Noble, M. (1970). Secondary task effects on serial verbal learning. *Journal of Experimental Psychology, 85,* 418–424.

Wingfield, A. (2014). Comment from Dr. Arthur Wingfield. *International Journal of Audiology, 53,* 442–444; discussion 444–445.