ORIGINAL ARTICLE
AUDIOLOGY
DOI: 10.5152/B-ENT.2021.20160

A Dutch version of a dual-task paradigm for measuring listening effort: a pilot study regarding its short-term test-retest reliability

Sofie Degeest, Paul Corthals, Hannah Keppler

Department of Rehabilitation Sciences, Ghent University, Ghent, Belgium

Cite this article as: Deegest S, Corthals P, Keppler H. A Dutch version of a dual-task paradigm for measuring listening effort: a pilot study regarding its short-term test-retest reliability. B-ENT 2021; 17(3): 135-44.

ABSTRACT

Objective: This study investigated the short-term test-retest reliability of a Dutch dual-task (DT) paradigm for measuring listening effort. Listening effort refers to the attention and cognitive resources necessary for understanding speech and can therefore provide information over and beyond the traditional speech audiometry outcomes. Such information can be beneficial in clinical practice as a part of the audiological test battery and therapeutic interventions such as hearing aids. To use this Dutch DT paradigm in further research and clinical practice, studies regarding its reliability are necessary.

Methods: A DT paradigm was used in which a primary speech-recognition task and a secondary visual memory task needed to be performed separately and simultaneously. Twenty-three young adults between the age of 18 and 31 years with normal hearing were tested at two moments.

Results: The intraclass correlation coefficient (ICC) showed a variation in reliability for both the primary and secondary tasks. In contrast, the coefficient of variation of the method error (CV_{ME}) showed good reliability for both the primary and secondary tasks. For listening effort, a large variation in ICC values as well as CV_{ME} values were found.

Conclusion: This study explored the short-term reliability of a DT paradigm for measuring listening effort. On the basis of these results, further studies to expand and refine this paradigm as well as studies regarding test-retest reliability are needed.

Keywords: Cognition, dual-task, listening effort, reliability, speech

Introduction

In complex listening situations, speech understanding can become a challenging and often exhausting experience because the information in the speech signal can be degraded by background noise and by hearing loss (1). Difficulties with understanding speech are typically quantified using standard pure-tone audiometry in combination with speech audiometry (in quiet or in noise). However, these standard audiological measurements are not suitable to detect and differentiate the degree of difficulty and effort related to understanding speech in a specific listening situation (2, 3). An individual, for example, may understand speech in a noisy listening situation, though this person may expend a large amount of effort to process these speech signals (4). The latter is important in the audiological diagnostic test battery, and especially in the assessment of difficulties in understanding speech in older adults. Furthermore, this is also important in terms of therapeutic interventions such as hearing aids, as it has been suggested that the amount of effort related to understanding speech can be affected by the hearing aid fitting (5). Specifically, different signal-processing features can lead to the same speech understanding scores, although the amount of difficulty and effort required to obtain this score can differ considerably between the different signal-processing features (4, 5). In this respect, there has been a surge of interest regarding the assessment of the amount of effort related to speech understanding. This particular effort, or the "listening effort," can be defined as the attention and cognitive resources necessary to understand speech (2, 6).

Until now, there is no "gold standard" measure of listening effort. Moreover, several measures of listening effort tap into multiple underlying mechanisms (7). In literature, several mea-

Corresponding Author: Sofie Degeest, sofie.degeest@ugent.be
Received: November 20, 2020 Accepted: July 17, 2021 Available Online Date: October 18, 2021
Available online at www.b-ent.be

sures of listening effort have been used. First, listening effort can be examined using self-reports (8). Second, listening effort can be evaluated through psychophysiological measures such as pupillometry (9), eye movement tracking (10), galvanic skin response, electromyographic activity, or heart rate variability (11). Finally, a dual-task (DT) paradigm can be used to investigate the listening effort (2, 3) and consists of simultaneously performing primary and secondary tasks. Specifically, this paradigm uses the limited extent of processing information and hypothesizes that this mental capacity is allocated to the sensory systems involved in a specific task (12). Typically, it is expected that the necessary mental capacity will be used to perform the primary task (12). When the primary task becomes more difficult, for example, due to background noise, less spare mental capacity will remain to accomplish the secondary task. A decrease in the secondary task performance, therefore, reflects more listening effort (13). An overview of the different types of primary and secondary tasks that are used within the DT paradigm for measuring listening effort is provided in a review by Gagne et al (4). From this review, it appears that the primary task typically consists of a listening assignment that includes sentence or word recognition in quiet or in background noise. The secondary task can take several forms, such as a memory task, a probe reaction-time task, or a tactile pattern recognition task (2, 3, 14). However, currently, there is no consensus on which type of secondary task may be the most suitable for measuring listening effort. Furthermore, a study conducted by Wu et al. (15) evaluated the effect of two different secondary tasks (i.e., a driving task in a simulator and a visual reaction-time task in a sound-treated booth) on the amount of listening effort. The amount of listening effort was generally consistent for both secondary tasks, which indicates that it is possible to measure listening effort by using different types of secondary tasks (15).

An advantage of using a DT paradigm to measure listening effort is that it provides information about realistic listening situations, as individuals often need to listen when performing other tasks, which, in turn, provides a type of ecological validity (4, 8). Nevertheless, the review conducted by Gagne et al. (4) concluded that although most paradigms are suitable for detecting changes in listening effort, the large number of different tests makes it difficult to conclude which tools are most appropriate for measuring listening effort (4). To increase the usefulness of DT paradigms in scientific research as well as clinical practice, more systematic analyses are necessary (4) in which the validity and reliability of a specific paradigm is evaluated.

Main Points:

- Until now, there is no "gold standard" for measuring listening effort, as several measures of listening effort tap into multiple underlying mechanisms.
- Dual-task (DT) paradigms can be used to behaviorally measure listening effort. To increase the usefulness of dual-task paradigms in scientific research as well as clinical practice, more systematic analyses are necessary.
- Although the intraclass correlation coefficient showed a variation in reliability for both the primary and secondary tasks that were used in this study, the coefficient of variation of the method error showed good reliability for both these tasks.
- A DT paradigm can provide additional information over and beyond the traditional speech audiometry outcomes that are used in audiological practice.
- For repeated measurements in one individual, the current DT paradigm should be further optimized.

Reliability has been defined as the extent to which a measurement is consistent over time and is free of errors (16). In literature, it has been stated that a comprehensive set of statistical measures is necessary for assessing reliability (17, 18). Hence, to use a DT paradigm in further studies and clinical practice, studies regarding its reliability are necessary. According to these results, further optimization and refinement of this paradigm can be assured. In this study, therefore, we aimed to examine the short-term reliability of a Dutch version of the DT paradigm for measuring listening effort. Specifically, the performances on the DT paradigm were assessed during two test moments in a group of young adults with normal hearing.

Methods

Participants

The study sample consisted of 23 young adults (four men and 19 women) aged between 18 and 31 years (mean 24.0 years, standard deviation [SD] 3.74 years). To evaluate the short-term test-retest reliability of the DT paradigm, all the participants were retested at a time interval between one and two weeks (mean 1.0 week, SD 0.37 weeks).

Each participant was a native speaker of Dutch and had no history of communication or learning problems, attention deficits, or known neurological disorders. Furthermore, a 226 Hz tympanometry with an 85 dB sound pressure level probe tone was performed to measure middle-ear function (AA222 audio traveler; Interacoustics, Assens, Denmark). For all participants, a normal middle-ear function was found at both test moments, according to the tympanometry results. The modified Hughson-Westlake technique was used to examine the hearing status at octave frequencies between 0.25 kHz and 8.0 kHz as well as half-octave frequencies of 3.0 kHz and 6.0 kHz (AA222 audio traveler; Interacoustics, Assens, Denmark). At both test moments, each participant had a normal hearing status, implying that the hearing thresholds were bilaterally equal to or better than 20 dB HL at each frequency tested.

This study was approved by the local ethics committee of the Ghent University Hospital (EC/2012/166). All the participants agreed with informed consent in accordance with the statements of the declaration of Helsinki.

Dual-task format

A primary task, consisting of a speech-recognition task in various listening conditions, and a secondary visual memory task were performed both separately (further denoted as "baseline condition") and simultaneously (further denoted as "DT condition"). Detailed information pertaining to the stimuli and test setup of the primary and secondary tasks can be found elsewhere (19, 20).

The stimuli of the primary speech-recognition task consisted of monosyllabic digits from zero to 12. Each spoken digit was

digitally mixed with a steady-state noise, whereby the intensity level of the digits was altered to create various signal-to-noise ratios (SNRs); +4 dB, +2 dB, 0 dB, −2 dB, −4 dB, −6 dB, −8 dB, and −10 dB. In addition, there was also a quiet listening condition without background noise. The speech stimuli were presented through two loudspeakers (Creative Inspire 265; Creative technology Ltd).

The secondary task was a visual memory task, in which geometric figures (identical blue-filled circles) appeared consecutively for one second in a raster. A series of five blue-filled circles were presented to the participants, whereby they had to follow the positions of these blue-filled circles in the raster. To ensure that each participant had normal or corrected-to-normal visual acuity, visibility was measured by using Sloan Letters (21) and by subjectively asking each participant whether the blue-filled circles could be distinguished on the computer screen.

Test procedure

The entire examination was carried out in a quiet, non-reverberant room. A summary of this procedure is outlined below, and further details can be found elsewhere (19, 20). A participant was excluded from the study if either the primary or secondary tasks could not be fully completed or if the score on the baseline secondary visual memory task was less than 50%.

Baseline condition

Baseline values for the primary speech-recognition task were determined by presenting two series of five digits in the quiet condition and at each SNR from +4 dB to −10 dB. On the basis of word scoring, each listening condition was scored at a total of 10 points. Baseline values for the secondary memory task were determined by presenting a series of five circles in the raster, post which, the participants had to indicate on a score form the exact position in the raster where each of the five circles had appeared. For the secondary task, one point was assigned for each circle that was indicated correctly.

Dual-task condition

The DT condition consisted of five digits that were presented simultaneously with five circles. Specifically, for each listening condition (the quiet condition and each SNR from +4 dB to −10 dB), two series of five digits and circles were offered to the participants. In each of the listening conditions, participants were instructed to give priority to the primary speech-recognition task (2, 3). To score the primary speech-recognition task as well as the secondary visual memory task, the same protocol as used in the baseline condition was applied.

Listening effort

For each participant, listening effort was determined as the performance shift of the secondary task from the baseline to the DT condition.

Formula to determine listening effort is as follows (22):

For each of the conditions, the occurrence of digits across each of the series was randomized, as well as the order of presentation of the different listening conditions. During the test-retest data collection, each participant performed the test twice, whereby each participant had identical test conditions at both test moments.

Statistical analysis

Statistical analysis was performed using the Statistical Package for Social Sciences version 21 (IBM SPSS Corp.; Armonk, NY, USA). Descriptive parameters were established, and tests of normality (Shapiro-Wilk, histograms, QQ-plots, and box and whisker plots) were applied to the different outcome variables, that is, the primary and secondary task outcomes in both baseline and DT conditions, as well as the calculated amount of the listening effort. Subsequently, for each listening condition, paired Student's t-tests were performed to evaluate the assumption that a participant's performance on the speech-recognition task remained stable between the baseline and DT condition (12).

For each of the outcome variables, the test-retest reliability was evaluated using a comprehensive set of statistical measures. First, repeated measurement analysis of variance (ANOVA) was used to investigate possible changes between the test and retest conditions with the listening condition (i.e., quiet and SNR from +4 dB to −10 dB) and test moment (test versus retest) as within-subject factors. Changes in the mean values between two test moments can consist of either a random change or a systematic change. Random change results from inherent variations within the actual test situation (e.g., variability in the equipment or test environment, and any other unmeasured variability in the subject's response), whereas systematic changes result from non-random variations (e.g., learning effects) (17, 18). Second, a two-way random model single-measures intraclass correlation coefficient (ICC) was used to determine the consistency of the position of individual scores relative to others between the two test moments (23). According to Fleiss (24), ICC values > 0.75 represent excellent reliability, values between 0.4 and 0.75 fair to good reliability, and values < 0.4 represent poor reliability. Nevertheless, ICC can be misleading if the sample is homogeneous, which is reflected by the between-subject variability not reaching statistical significance ($p > 0.05$). As only young adults with normal hearing were included in the study, the primary and secondary task outcomes as well as the amount of listening effort can be homogeneously distributed and can therefore lead to lower ICC values than in a more heterogeneous group (17). Hence, the method error (ME) was determined in addition to the ICC. ME is not affected by a lack of variability and expresses test-retest reliability in terms of the percentage variation from trial to trial (25). Specifically, ME is calculated using the SD of the difference scores (SD_{diff}) between the test and retest conditions by means of the following formula (25):

ME is often converted to a percentage as it must be interpreted relative to the size of the mean difference. This conversion has been described as the coefficient of variation of the ME (CV_{ME}) and can be calculated using the following formula (17):

By using the ME and CV_{ME}, the reliability of the different outcome variables can be compared, with lower CV_{ME} values reflecting higher reliability.

Finally, the standard error of measurement (SEM) and the minimal detectable difference (MDD) were calculated as these reliability parameters can be used for clinical applications by providing a reference for evaluating test outcomes over time. The SEM evaluates the reliability of repeated measures in one subject and was estimated by taking the square root of the mean

Table 1. Mean values and standard deviations of the primary speech-recognition and the secondary visual memory tasks in baseline and dual-task conditions as well as the amount of listening effort for both the test and retest conditions (n = 23).

	Primary speech-recognition task (raw score)		Secondary visual memory task (raw score)		
Listening condition	Baseline	Dual-task	Baseline	Dual-task	Listening effort (%)
Test Condition					
No listening condition	NA	NA	9.87 (0.34)	NA	NA
Quiet	9.91 (0.42)	9.78 (0.52)	NA	8.13 (1.32)	18.50 (12.40)
SNR					
+4 dB	9.57 (0.66)	9.74 (0.45)	NA	8.83 (1.53)	11.45 (14.93)
+2 dB	9.83 (0.39)	9.65 (0.57)	NA	8.65 (1.40)	13.29 (13.06)
0 dB	9.43 (0.79)	9.61 (0.72)	NA	8.48 (1.31)	14.15 (12.61)
−2 dB	9.48 (0.85)	9.26 (0.69)	NA	8.13 (1.49)	18.50 (14.12)
−4 dB	8.65 (1.03)	8.91 (0.95)	NA	9.04 (0.88)	9.28 (8.09)
−6 dB	8.13 (0.87)	8.26 (0.81)	NA	8.43 (1.88)	15.36 (18.52)
−8 dB	7.65 (1.34)	7.70 (0.97)	NA	7.57 (1.44)	23.29 (14.67)
−10 dB	6.13 (1.18)	6.13 (1.55)	NA	8.39 (1.41)	15.80 (13.78)
Retest Condition					
No listening condition	NA	NA	9.91 (0.29)	NA	NA
Quiet	9.80 (0.67)	9.91 (0.29)	NA	8.83 (1.37)	11.88 (13.09)
SNR					
+4 dB	9.74 (0.54)	9.87 (0.34)	NA	9.09 (1.16)	9.23 (11.26)
+2 dB	9.83 (0.49)	9.78 (0.52)	NA	8.91 (1.78)	10.97 (17.56)
0 dB	9.52 (0.67)	9.70 (0.56)	NA	9.04 (1.02)	9.66 (9.77)
−2 dB	9.65 (0.57)	9.30 (0.63)	NA	8.57 (1.56)	13.48 (16.13)
−4 dB	9.09 (0.99)	9.04 (0.97)	NA	9.30 (0.93)	7.05 (8.80)
−6 dB	8.83 (1.07)	8.57 (0.95)	NA	9.00 (1.17)	10.97 (11.24)
−8 dB	7.70 (1.15)	7.78 (0.95)	NA	7.65 (1.50)	22.66 (15.70)
−10 dB	5.83 (1.44)	6.04 (1.52)	NA	8.48 (1.16)	15.27 (11.61)

NA: not applicable; SNR: signal-to-noise ratio

square error term from the ANOVA (SEM = $\sqrt{MS_E}$). This specific method of calculating the SEM was used to exclude the influence of the range of measured values (18). Subsequently, the SEM was used to calculate the minimum detectable difference (MDD). The MDD can be defined as the amount of change in the outcome variables that must exist to conclude that there is a true test-retest difference. To indicate the 95% confidence interval (CI) to detect a real difference, the following equation was used (25):

In addition to the statistical measures of the test-retest reliability, cumulative frequencies of the absolute score differences between the test and retest conditions were calculated for the baseline primary speech-recognition task, the baseline and DT secondary visual memory tasks, and the amount of listening effort. The cumulative frequency is used to determine the number of observations that lie above or below a particular value in the data set. For this study, a cumulative frequency distribution can indicate how frequently a particular test-retest difference occurs in the sample and, therefore, can be used to explore the individual variation in scores across the test-retest interval. Specifically, the cumulative frequency of the absolute score difference between test and retest was calculated as the frequency of occurrence of that score difference plus the sum of the frequencies of all scores differences with a lower value. In addition to the cumulative frequency, the cumulative percentage was calculated as:

Results

Baseline and dual-task performance

For each listening condition, descriptive analyses were conducted, and the average baseline and DT speech-recognition scores in both the test and retest conditions are shown in Table 1. In both the test and retest conditions, speech-recognition scores decreased when the SNR became more negative. Paired Student's t-tests were conducted to assess the scores on the speech-recognition task between baseline and DT conditions. For both the test and retest conditions and at each listening condition, speech-recognition scores did not differ

significantly between the baseline and DT conditions (Student's t-tests, $p > 0.05$). Hence, for the speech-recognition task, test-retest reliability measures will only be based on the baseline speech-recognition scores.

Table 1 further shows the average baseline and DT visual memory task scores as well as the calculated amount of listening effort in both the test and retest condition. According to these descriptive results, it can be seen that the mean performance on the visual memory task generally remained stable across the listening conditions.

Test-retest reliability

Test-retest reliability measures were first performed for the speech-recognition task at each listening condition in the baseline conditions. For the visual memory task, test-retest reliability was determined for the baseline condition and at each listening condition in the DT condition. In addition, test-retest reliability was determined for the amount of listening effort. The reliability measures are outlined in Table 2 (repeated measures ANOVA, the mean differences [$Mean_{diff}$] with their standard deviations [SD_{diff}], ICCs, CV_{MEs}, SEMs, and MDDs).

Baseline speech-recognition task

The reliability measures for the baseline speech-recognition scores are presented in Table 2a. Repeated measures ANOVA revealed no significant difference between the two test moments for each listening condition ($p > 0.05$). The ICC values vary between the different listening conditions, with excellent reliability for the listening condition with a SNR of 0 dB, fair to good reliability for the listening conditions with a SNR of +4 dB and −10 dB, and low to even negative reliability for the remaining listening conditions. In contrast, the CV_{ME} values are low for each listening condition (CV_{ME} range 3.11–14.01), thus reflecting good reliability. The SEMs and MDDs ranged from 0.29 to 1.05 and from 0.82 to 2.92 for the different listening conditions, respectively.

Baseline visual memory task

For the visual memory task scores, repeated measures ANOVA showed no significant difference between the test and retest conditions ($p > 0.05$). As shown in Table 2b, ICC for the baseline visual memory task yielded a negative value. However, the CV_{ME} value was low (CV_{ME} = 3.39), which indicates good reliability. The SEM and MDD yielded a value of 0.34 and 0.93, respectively.

Dual-task condition

In terms of the baseline visual memory task scores, no significant difference between the test- and retest conditions was found for the visual memory task in the DT condition (repeated measures ANOVA, $p > 0.05$). On the basis of ICC values, fair to good reliability was found for all listening conditions, except the listening condition with a SNR of −4 dB, which showed low reliability. The CV_{ME} values were low for all listening conditions (CV_{ME} range 7.78–12.73), indicating good reliability (Table 2c). The SEMs and MDDs ranged from 0.71 to 1.06 and from 1.98 to 2.94 for the different listening conditions, respectively.

Listening effort

According to the repeated measures ANOVA, no significant difference in listening effort was found between the test and retest conditions for each listening condition ($p > 0.05$). As seen in Table 2d, ICC displayed fair to good reliability for all listening conditions, except the listening condition with a SNR of −4 dB, which showed low reliability. Furthermore, for all listening conditions, the CV_{ME} values were higher than the CV_{ME} values at baseline and DT visual memory scores separately (CV_{ME} range 37.75–83.84), which indicates poorer reliability. The SEMs and MDDs ranged from 6.85 to 10.15 and from 18.97 to 28.13 for the different listening conditions, respectively.

In addition to the statistical measures of the test-retest reliability, the absolute score differences between the test and retest conditions were calculated for each participant for the baseline speech-recognition task (Figure 1), baseline and DT visual memory task (Figure 2), and the amount of listening effort (Figure 3). On the basis of these score differences between test and retest conditions, the cumulative frequency and cumulative percentage were determined. Both the speech-recognition task and the visual memory task were scored out of 10

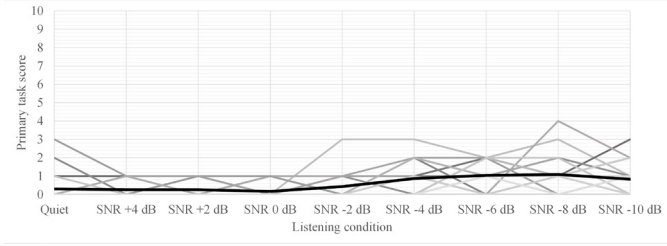

Figure 1. Overview of the absolute score differences between the test and retest conditions for the baseline primary speech-recognition task for each individual participant. The black filled line represents the average speech-recognition score of all the participants (n = 23) for each listening condition.

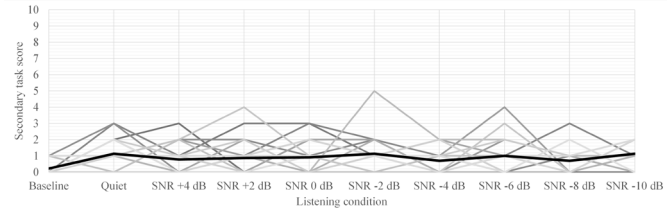

Figure 2. Overview of the absolute score differences between the test and retest conditions for the baseline and dual-task secondary visual memory task for each individual participant. The black filled line represents the average visual memory task score of all the participants (n = 23) for each listening condition.

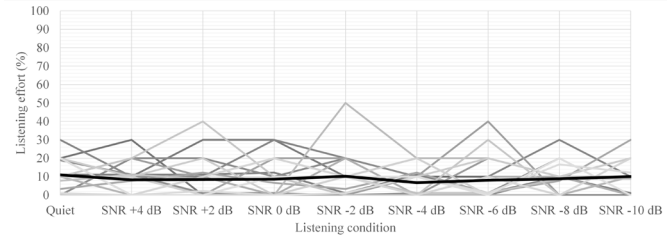

Figure 3. Overview of the absolute score differences between the test and retest conditions for the amount of listening effort for each individual participant. The black filled line represents the average amount of listening effort of all the participants (n = 23) for each listening condition.

Table 2. Statistical measures of the short-term test-retest reliability of the baseline speech-recognition task (a), baseline visual memory task (b), dual-task visual memory task (c) and the amount of listening effort (d)

	Listening condition	Repeated measures ANOVA		$Mean_{diff}$	SD_{diff}	ICC	CV_{ME}	SEM	MDD
		F-value	p						
a.	Baseline primary speech-recognition task								
	Quiet	0.59	>0.05	−0.13	0.81	−0.06	5.85	0.58	1.60
	SNR								
	+4 dB	2.89	>0.05	−0.17	0.49	0.67	3.60	0.35	0.96
	+2 dB	0.01	>0.05	0.00	0.52	0.30	3.76	0.37	1.02
	0 dB	1.0	>0.05	0.09	0.42	0.84	3.11	0.29	0.82
	−2 dB	1.0	>0.05	0.17	0.83	0.33	6.17	0.59	1.64
	−4 dB	3.02	>0.05	0.43	1.20	0.30	9.56	0.85	2.35
	−6 dB	9.11	>0.05	0.70	1.11	0.36	9.22	0.78	2.17
	−8 dB	0.02	>0.05	0.04	1.49	0.28	13.75	1.05	2.92
	−10 dB	1.52	>0.05	−0.30	1.18	0.59	14.01	0.84	2.32
b.	Baseline secondary visual working memory task								
	NA	0.19	>0.05	0.04	0.47	−0.12	3.39	0.34	0.93
c.	Dual-task secondary visual memory task								
	Quiet	6.64	>0.05	0.70	1.29	0.54	10.80	0.92	2.54
	SNR								
	+4 dB	1.13	>0.05	0.26	1.18	0.63	9.29	0.83	2.31
	+2 dB	0.81	>0.05	0.26	1.39	0.63	11.18	0.98	2.72
	0 dB	4.80	>0.05	0.57	1.24	0.45	9.98	0.87	2.42
	−2 dB	1.93	>0.05	0.43	1.50	0.51	12.73	1.06	2.94
	−4 dB	1.54	>0.05	0.26	1.01	0.37	7.78	0.71	1.98
	−6 dB	4.08	>0.05	0.57	1.34	0.63	10.89	0.95	2.63
	−8 dB	0.15	>0.05	0.09	1.08	0.73	10.07	0.77	2.12
	−10 dB	0.10	>0.05	0.09	1.38	0.43	11.56	0.98	2.70
d.	Listening effort								
	Quiet	6.83	>0.05	−6.62	12.15	0.55	56.53	8.59	23.81
	SNR+4 dB	0.83	>0.05	−2.22	11.71	0.61	80.10	8.28	22.95
	+2 dB	0.67	>0.05	−2.32	13.54	0.62	78.97	9.58	26.54
	0 dB	3.09	>0.05	−4.49	12.27	0.41	72.83	8.67	24.04
	−2 dB	2.82	>0.05	−5.02	14.35	0.55	63.48	10.15	28.13
	−4 dB	1.21	>0.05	−2.22	9.68	0.34	83.84	6.85	18.97
	−6 dB	2.60	>0.05	−4.40	13.08	0.64	70.24	9.25	25.63
	−8 dB	0.06	>0.05	−0.63	12.26	0.67	37.75	8.67	24.03
	−10 dB	0.04	>0.05	−0.53	13.35	0.45	60.80	9.44	26.17

$Mean_{diff}$: mean differences; SD_{diff}: standard deviations of the differences; ICC: intraclass correlation coefficient; CV_{ME}: coefficient of variation of the method error; SEM: standard error of measurement; MDD: the minimal detectable difference to determine a confidence interval of 95%; NA: not applicable; SNR: SNR: signal-to-noise ratio

points, so that the differences in the scores between the test and retest conditions can range between 0 and 10. Considering the baseline speech-recognition score, 90% of the subjects had a score difference between test and retest conditions ≤ 1 for the quiet listening condition as well as the listening condition with a SNR of +4 dB, +2 dB, 0 dB, and −2 dB. For the other listening conditions, 90% of the subjects had a score difference ≤ 2. In the case of the visual memory task in baseline and DT conditions, 90% of the subjects had a score difference ≤ 1 and ≤ 2, respectively. The amount of listening effort was ex-

pressed as a percentage, so that the minimum and maximum difference in listening effort that could occur between the test and retest conditions was 0% and 100%, respectively. For listening effort, 90% of the subjects had a difference in listening effort ≤ 20% in all listening conditions.

Discussion

In recent years, several types of DT paradigms have been used to measure listening effort during understanding speech in different populations, particularly in the aging population. To increase the usefulness of DT paradigms in scientific research as well as clinical practice, the validity and reliability of such paradigms should be evaluated (4). The goal of this study was therefore to assess the short-term test-retest reliability of a Dutch DT paradigm for measuring listening effort.

The DT paradigm presumes that an increase in the difficulty of the primary speech-recognition task will result in a greater need for cognitive resources to accomplish this speech assignment, which, in turn, will result in less cognitive resources to complete the secondary visual memory task. In this study, the mean performances on the secondary visual memory task and, as a result, the amount of listening effort generally remained rather stable across the listening conditions (the scores were close to each other with no clear decrease toward the more unfavorable listening conditions). The presentation order of the different listening conditions was randomized to exclude a systematic learning effect. However, it should be pointed out that the performance on both the primary speech-recognition and the secondary visual memory tasks was high in each listening condition (i.e., more than five out of 10). This can be attributed to the homogeneous group of young adults with normal hearing included in this study, who were expected to perform well on both tasks because of normal hearing as well as normal cognitive capacities. Hence, such high-performance levels for each of the listening conditions may have been the reason for the lack of a gradual decrease in the scores on the secondary visual memory task from the favorable to the unfavorable listening conditions. In addition, previous studies have shown that measuring listening effort can be affected by some factors apart from the task itself (26). DT outcomes may, for example, be less sensitive if the primary task is too easy (i.e., quiet listening condition) or too difficult (i.e., listening condition with high levels of background noise) (27, 28). Hence, the results found in this study may also be explained by the participants' degree of attention during their performance in each of the listening conditions. It can be suggested that the participants' degree of attention varied from one condition to another, leading to a variation in the performance on the visual memory task. According to the framework for understanding effortful listening (6), the amount of listening effort can also be influenced by the demands of the listening condition and the motivation of the subject to keep engaged in the listening task. In this study, motivation was not evaluated during the task; therefore, its impact on the results cannot be ruled out. Specifically, as was described by Wu et al. (28) and Zekveld et al. (29), the participants may at times have experienced cognitive overload in the more difficult listening conditions, tending to give up on the primary speech-recognition task. As a result, the participants might have exerted more effort on the secondary task to pursue reward (28). Moreover, studies have also indicated that listening and the allocation of effort during listening in daily life can differ between individuals with normal hearing and those with hearing impairment (30). As suggested by Alhanbali et al. (7), it will be important to also consider listening conditions that represent real-life situations. Therefore, as also described in the review by Gagne et al. (4), further studies are required to explore the relationship between performance on the speech-recognition task and the amount of listening effort. Specifically, the current DT paradigm can be expanded to include listening conditions that are common in daily life as well as listening conditions where the primary speech understanding score decreases to 50% or less. As a result, it will be possible to investigate the amount of listening effort in real-life listening conditions and to investigate which listening conditions will be most sensitive to measure listening effort. Such information will be important for clinical practice, in particular the use of listening effort within the audiological diagnostic test battery as well as therapeutic interventions such as hearing aids.

The performances on the DT test were evaluated during two test moments. As the DT paradigm implies performing a primary task and a secondary task both separately (i.e., the baseline condition) and simultaneously (i.e., the DT condition), the reliability of both tasks was evaluated in both conditions. In addition, test-retest reliability of the amount of listening effort was evaluated.

First, reliability was assessed by evaluating the changes in the mean scores between the test and retest conditions. As mentioned above, changes in the mean values between the two test moments can consist of either a random change or a systematic change. Random changes might be attributed to variability in the equipment that is used or variation related to the test environment (e.g., ambient noise). Furthermore, inherent biological variability, such as changes in hearing status, can also lead to changes between the two test moments. For example, especially in listening conditions with background noise, hearing status can negatively affect the speech-recognition outcomes and, therefore, the performance on the visual memory task and the resulting amount of listening effort (14). No statistically significant differences were found between the test and retest conditions for each of the conditions of both the speech-recognition task and the visual memory task, as well as for listening effort. These results corroborate the accuracy of the test procedure that was used, in which the equipment as well as the test environment were controlled and where the hearing status of the participants did not differ between the two test moments. However, a shift in the mean scores between the test and retest conditions can also be associated with systematic errors such as learning effects. The fact that no statistically significant differences were found for both the speech-recognition task and the visual memory task as well as for listening effort indicates the absence of a major learning effect. Moreover, during the development of the current DT paradigm, there was a careful selection of the primary and secondary tasks whereby several factors, such as learning effects, were taken into account.

A second category of reliability measures concerns the ICC and the CV_{ME}. ICC values showed a variation in reliability, rang-

ing from "fair to good," for the different conditions of both the speech-recognition task and the visual memory task, as well as for listening effort. These variations in ICC values may raise questions about the reliability of the DT paradigm that was used, though the values were similar to the study of Giuliani et al. (31) It is important to note that ICC outcomes can be misleading as only young adults with normal hearing were included, which could lead to a lack of inter-subject variability. As mentioned above, the participants in this study had overall high performance levels for the speech-recognition task; hence, DT was insensitive to changes in SNR (28, 31). Specifically, these good performance levels may indicate that the cognitive load required to accomplish the speech-recognition task in each of the conditions is not high enough to create a variation in performance levels of the visual memory task and the amount of listening effort derived from it. Indeed, negative ICC values were obtained for the baseline speech-recognition scores in the quiet listening condition and the baseline visual memory scores as the majority of the subjects achieved the maximum score in both the test and retest conditions. Nevertheless, the authors have deliberately chosen to include only young adults with normal hearing in this pilot study to exclude the confounding influence of age and hearing loss, as it is well known that such factors can negatively affect the listening effort (1, 3, 14, 19, 32).

Consequently, an extra reliability parameter was calculated to assess the reliability of the DT paradigm for measuring listening effort. Specifically, the ME and CV_{ME} were used, which are not affected by a lack of between-subject variability because ME is based on the SD of the difference between the test and retest measurements (25). Furthermore, the ME and CV_{ME} express test-retest reliability in terms of the percentage variation from trial to trial (17, 25). In contrast with the ICC values, the CV_{ME} values were less variable and were low for each of the listening conditions of the baseline speech-recognition task as well as the visual memory task in both the baseline condition and the different listening conditions in the DT condition, thus indicating good reliability. The CV_{ME} values are, however, slightly lower (higher reliability) for the speech-recognition task than for the visual memory task. A possible reason for this small difference in CV_{ME} values is related to the design of the DT paradigm. First, the primary task is a speech-recognition task in which digits are used as speech stimuli. In literature, it has been described that such a test can be administrated multiple times with a low risk of familiarity as it is difficult for the participants to remember the different digit combinations that were already used (33). Second, CV_{ME} values may be better for the speech-recognition task because of task prioritization. Participants were instructed to give priority to the primary speech-recognition task (2), meaning that they were instructed to optimize their performance on the primary task. No significant differences in the primary speech-recognition scores were observed between the baseline and DT condition, which was an indication that the participants did mostly pay attention to this task in both the baseline and DT conditions. However, as also mentioned earlier, although the scores on the primary speech-recognition task were equal between the baseline and DT conditions, it should be noted that it is not possible to rule out whether the participants allocated their attention predominantly to the secondary task. Previous studies with children have shown that instructions alone may not be adequate to ensure that a participant will primarily focus on the primary task (34). Furthermore, the present study used different fixed listening conditions to evaluate listening effort; thus, the speech-recognition performance in each listening condition will differ for each participant. Further studies are, therefore, necessary to evaluate how adult individuals prioritize their attention between the primary and secondary tasks in different listening conditions. The present DT paradigm can be expanded by adding one or more conditions where the performance on the baseline speech-recognition task is equalized (e.g., 50% and 80% speech understanding) so that the variance in the performance on this task across the listening conditions can be taken into account (31). In addition, the relative change in performance associated with performing a DT can be calculated not only for the secondary task but also for the primary task, which makes it possible to investigate how much capacity an individual allocates to each task (35).

Compared with the speech-recognition and the visual memory tasks in both baseline and DT conditions, the CV_{ME} values of the calculated amount of listening effort were more variable and were clearly higher, indicating lower reliability. This result may be explained by the between-trial variability in both the baseline and DT conditions as well as the formula that was used to calculate listening effort (36). The DT paradigm used in this study included two series of five circles for each listening condition, which resulted in a total score of 10 points for each listening condition and, therefore, a small measurement scale. Hence, this small measurement scale will probably be reflected in the amount of listening effort, which was calculated as the performance shift on the visual memory task between the baseline and DT conditions. Furthermore, the differences in listening effort across the two test moments can be further increased by variations in the baseline visual memory performance, as this is the denominator of the equation. For example, if a participant's baseline visual memory performance at the test moment yielded a score of 8 (out of a total score of 10), and DT performance yielded a score of 7 (out of a total score of 10), the amount of listening effort will be 12.5%. However, if the participant's baseline visual memory score increased to 9 (out of a total of 10) at the retest moment, and the DT performance remained constant, the resulting amount of listening effort would be 22.2%, which means a doubling in the amount of listening effort between the test and the retest. Hence, a small difference in the performance of the secondary visual memory task between test and retest in either the baseline or DT conditions can result in a notable difference in the listening effort. Although the primary and secondary tasks were carefully selected and evaluated for floor and ceiling effects during previous laboratory work, using this small number of trials for each listening condition leads to the possibility of under sampling. Therefore, the current DT paradigm can be adjusted, whereby the number of test trials in each of the listening conditions will be increased (more series of five digits).

This study also calculated the SEM and MDD, which provide a reference for using this DT paradigm over time. Specifically, the use of the MDD enables detection of a significant change in the scores on both the speech-recognition task and visual memory task as well as the amount of listening effort. In addition to the SEM and MDD, the cumulative frequencies of

the absolute score differences were calculated for both the speech-recognition task and the visual memory task, as well as for listening effort. Cumulative frequency distributions provide the number of values in the sample that are at or below a given value and can therefore be used to compare the results of future studies as no reliability data are yet available in this area. On the basis of the MDD, a 95% CI could be derived to reflect the interval in which 95% of the observations of a person can be found. For example, if this interval is exceeded during a second measurement, the difference observed will probably be because of a real or genuine difference. For the cumulative frequencies, a cutoff was determined at which 90% of the score differences were included. Such intervals are essential when two results of a person are compared to discover any changes in the performance of both the primary and secondary tasks because of, for example, alterations in hearing sensitivity or cognitive capacity. Furthermore, these intervals can be used to evaluate the effects of therapeutic interventions such as hearing aid fitting or auditory training programs.

In conclusion, this study is the first to explore the short-term test-retest reliability of a Dutch DT paradigm for measuring listening effort. On the basis of the outcomes of the ICC measures, a variation in reliability was found across the different listening conditions as well as between the primary and secondary tasks and the calculated amount of listening effort. These findings may be attributed to the high performance levels that were found, as only young adults with normal hearing were included in the study, leading to a lack of between-subject variability and a lack of variation in performance levels. Furthermore, a large variation as reflected by large standard deviations for the calculated amount of listening effort was observed in this study. As mentioned before, the calculation method that was used to calculate the amount of listening effort as well as the small number of trials for each listening condition will probably be related to the large standard deviations that were found. Besides, other factors, such as the degree of vigilance and motivation during the task, could have influenced the outcome of the DT paradigm as well as the sensitivity of the listening conditions that were used to evaluate listening effort. In contrast to the ICC, CV_{ME} values, which are not affected by a lack of variability in the measurement, showed good reliability. However, the CV_{ME} values of the calculated amount of listening still showed a larger variation. Notwithstanding, the Dutch DT paradigm has demonstrated the ability to show differences in the amount of listening effort between different groups, that is, the effect of age on the amount of listening effort as well as differences in the listening effort between young adults with and without tinnitus can be used to measure listening effort (19, 20). A DT paradigm can provide additional information over and beyond the traditional speech audiometry outcomes used in audiological practice. However, to be useful for repeated measurements in the same individual, the current DT should be further optimized. In this respect, further studies should take into account the following factors; increasing the measurement scale to provide more test trials, the inclusion of real-life listening conditions as well as individualized SNRs (e.g., SNR with 50% speech understanding), and evaluation of the reliability of including individuals of different age groups as well as subjects with and without hearing loss to increase inter-subject variability. Ultimately, further studies investigating long-term test-retest reliability can be performed to evaluate the use of this DT paradigm over the long term, which could be beneficial in, among others, the evaluation of therapeutic interventions such as hearing aids. Outcomes from the DT paradigm can help clinicians to better verify complaints regarding difficulties in understanding speech, especially when the traditional speech audiometry outcomes are within normal limits.

Ethics Committee Approval: This study was approved by Ethics committee of Ghent University, (Approval No: EC/2012/166).

Informed Consent: Written informed consent was obtained from the patients who agreed to take part in the study.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – S.D., P.C., H.K.; Design – S.D., P.C., H.K.; Supervision – H.K.; Resources – S.D., H.K.; Materials – S.D., H.K.; Data Collection and/or Processing – S.D.; Analysis and/or Interpretation – S. D.; Literature Search – S. D., H.K.; Writing Manuscript – S. D.; Critical Review – P.C., H.K.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

References

1. Schneider BA, Daneman M, Pichora-Fuller MK. Listening in aging adults: from discourse comprehension to psychoacoustics. Can J Exp Psychol 2002; 56: 139-52. [Crossref]
2. Bourland-Hicks C, Tharpe AM. Listening effort and fatigue in school-age children with and without hearing loss. J Speech Lang Hear Res 2002; 45: 573-84. [Crossref]
3. Gosselin P, Gagne JP. Older adults expend more listening effort than young adults recognizing speech in noise. J Speech Lang Hear Res 2011; 54: 944-58. [Crossref]
4. Gagne J-P, Besser J, Lemke U. behavioral assessment of listening effort using a dual-task paradigm: a review. Trends Hear 2017; 21: 2331216516687287. [Crossref]
5. Kestens K, Degeest S, H K. The effect of cognition on the aided benefit in terms of speech understanding and listening effort obtained with digital hearing aids: a systematic review. Am J Audiol 2021; 30: 190-210. [Crossref]
6. Pichora-Fuller MK, Kramer SE, Eckert MA, et al. Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). Ear Hear 2016; 37: 5S-27S. [Crossref]
7. Alhanbali S, Dawes P, Millman RE, Munro KJ. Measures of listening effort are multidimensional. Ear Hear 2019; 40: 1084-97. [Crossref]
8. McGarrigle R, Munro KJ, Dawes P, et al. Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. Int J Audiol 2014; 53: 433-40. [Crossref]
9. Zekveld AA, Kramer SE, Festen JM. Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. Ear Hear 2011; 32: 498-510. [Crossref]
10. Ben-David BM, Chambers CG, Daneman M, Pichora-Fuller MK, Reingold EM, Schneider BA. Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. J Speech Lang Hear Res 2011; 54: 243-62. [Crossref]

11. Mackersie CL, MacPhee IX, Heldt EW. Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. Ear Hear 2015; 36: 145-54. [Crossref]
12. Kahneman D. Attention and effort. Englewood Cliffs, NJ: Prentice-Hall; 1973.
13. Downs DW. Effects of hearing aid use on speech discrimination and listening effort. J Speech Hear Disord 1982; 47: 189-93. [Crossref]
14. Desjardins JL, Doherty KA. Age-related changes in listening effort for various types of masker noises. Ear Hear 2013; 34: 261-72. [Crossref]
15. Wu Y-H, Aksan N, Rizzo M, Stangl E, Zhang X, Bentler R. Measuring listening effort: driving simulator vs. simple dual-task paradigm. Ear Hear 2014; 35: 623-32. [Crossref]
16. Portney L, Watkins M. Reliability of measurements. In: Portney L, Watkins M, editors. Foundations of clinical research: Applications to practice. New Jersey: Pearson Education Inc.; 2008. p. 77-96.
17. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. Am J Phys Med Rehabil 2005;84(9):719-23. [Crossref]
18. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports medicine 1998; 26: 217-38. [Crossref]
19. Degeest S, Keppler H, Corthals P. The effect of age on listening effort. J Speech Lang Hear Res 2015; 58: 1592-600. [Crossref]
20. Degeest S, Keppler H, Corthals P. The effect of tinnitus on listening effort in normal-hearing young adults: a preliminary study. J Speech Lang Hear Res 2017; 60: 1036-45. [Crossref]
21. Sloan LL, Rowland WM, Altman A. Comparison of three types of test target for measurement of visual acquity. Rev Ophthalmol 1952; 8: 4-17.
22. Kemper S, Schmalzried R, Herman R, Leedahl S, Mohankumar D. The effects of aging and dual task demands on language production. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 2009; 16: 241-59. [Crossref]
23. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005; 19: 231-40. [Crossref]
24. Fleiss J. The Design and Analysis of Clinical Experiments. New York: John Wiley Sons; 1986.
25. Portney L, Watkins M. Statistical measures of reliability. In: Portney L, Watkins M, editors. Foundations of clinical research Applications to practice. New Jersey: Pearson Education Inc.; 2008. p. 585-618.
26. Zekveld AA, van Scheepen JA, Versfeld NJ, Veerman EC, Kramer SE. Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. Hear Res 2019; 381: 107768. [Crossref]
27. Picou EM, Ricketts TA, Hornsby BW. How hearing aids, background noise, and visual cues influence objective listening effort. Ear Hear 2013; 34: e52-e64. [Crossref]
28. Wu Y-H, Stangl E, Zhang X, Perkins J, Eilers E. Psychometric functions of dual-task paradigms for measuring listening effort. Ear Hear 2016; 37: 660-70. [Crossref]
29. Zekveld AA, Kramer SE. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. Psychophysiology 2014; 51: 277-84. [Crossref]
30. Ohlenforst B, Zekveld AA, Jansma EP, Wang Y, Naylor G, Lorens A, et al. (2017). Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. Ear and Hearing 38: 267. [Crossref]
31. Giuliani NP, Brown CJ, Wu Y-H. Comparisons of the sensitivity and reliability of multiple measures of listening effort. Ear Hear 2020; 42: 465-74. [Crossref]
32. Gosselin P, Gagne JP. Use of a dual-task paradigm to measure listening effort. Can J Speech Lang Pathol Audiol 2010; 34: 43-51.
33. Smits C, Kapteyn TS, Houtgast T. Development and validation of an automatic speech-in-noise screening test by telephone. Int J Audiol 2004; 43:15-28. [Crossref]
34. Choi S, Lotto A, Lewis D, Hoover B, Stelmachowicz P. Attentional modulation of word recognition by children in a dual-task paradigm. J Speech Lang Hear Res 2008; 51: 1042-54. [Crossref]
35. Plummer P, Eskes G. Measuring treatment effects on dual-task performance: a framework for research and clinical practice. Front Hum Neurosci 2015; 9: 225. [Crossref]
36. Yang L, He C, Pang MYC. Reliability and validity of dual-task mobility assessments in people with chronic stroke. PloS One 2016; 11: e0147833. [Crossref]