1   Abstract

2   Speech-on-speech listening involves selectively attending to a target talker while ignoring a

3   simultaneous competing talker. Spatially separating the talkers improves performance, a

4   phenomenon known as spatial release from masking (spatial RM). The same is true of

5   spectral separation, i.e., filtering the talkers into non-overlapping frequency bands (spectral

6   RM). The relative benefit of spatial vs. spectral RM is currently unknown. Furthermore, it is

7   unclear how listeners' ability to exploit spatial vs. spectral cues is related to individual

8   differences in cognition. The resource-limit account suggests that cognitive resources are

9   required to support the processing of degraded speech, implying the strongest

10   cognition/performance relationship when RM is limited or absent. However, an alternative

11   claim, referred to as the data-limit account, suggests that cognitive resources cease to be

12   useful when the target is severely degraded. In this study, participants (N=240) completed a

13   selective listening task in which they transcribed the speech of one of two simultaneously-

14   presented talkers. The speech was filtered into interleaved or overlapping frequency bands

15   (spectral RM vs. no spectral RM) and presented dichotically or collocated (a proxy for spatial

16   RM vs. no spatial RM). A battery of cognitive tasks was administered to assess working

17   memory/attention. Spectral RM provided at least as much benefit as spatial RM, with the best

18   performance when both RM types were present. Cognitive scores were significantly

19   positively correlated with RM benefits. However, the weakest correlation between cognitive

20   scores and performance was observed in the no-RM condition. The results therefore support

21   an account of speech-on-speech listening that lies on a continuum from data-limited to

22   resource-limited processing as a function of the quality of the target speech signal.

1. Introduction

Speech perception in the presence of irrelevant background sound can be difficult for many listeners. Specific challenges are posed by speech-on-speech listening, in which the listener must selectively attend to a target talker whilst ignoring a simultaneous competing talker. These challenges arise in part from bottom-up constraints: access to the target signal is limited by energetic masking (EM) – spectro-temporal overlap between target and masker, which results in direct competition at the cochlea (Brungart, 2001). Intelligibility depends on how much of the target can be "glimpsed" through spectro-temporal gaps in the masker (Cooke, 2003); in other words, performance is constrained by the quantity of available data.

However, speech-on-speech listening also imposes cognitive demands. Listeners must not only successfully parse the auditory scene into separate streams (segregation), but also allocate attention to the target stream and inhibit the masker (Shinn-Cunningham, 2008). They may also need to store degraded speech fragments in working memory (WM) for later repair and/or integration into the broader sentence context (Gatehouse et al., 2003). Indeed, selective attention, inhibition, and WM are often seen as closely connected: attentional control is modelled as a key component of WM (e.g., Baddeley, 2000; Engle & Kane, 2004; Unsworth & Spillers, 2010) and good WM is associated with better inhibition of irrelevant sound during selective attention tasks (Conway et al., 2001). Other cognitive resources beyond WM, such as crystallised intelligence, have also been suggested to play a role (e.g., Schneider et al., 2016; although see Dryden et al., 2017). Both contextual factors and individual differences may affect a listener's ability to draw on this network of relevant cognitive skills (as revealed by, for example, dual-task paradigms; e.g., Desjardins et al., 2013). Performance will therefore also be constrained by the extent to which cognitive resources are available to process the available data.

1.1 Data-limited vs. resource-limited processes

The distinction between data and resource availability is taken from the work of Norman and Bobrow (1975), as implemented in Mattys et al.'s (2025) Data-Resource-Language (DRL)_ framework. According to Norman and Bobrow's conception, a task is resource-limited if performance is determined primarily by the quantity of processing resources allocated to it. However, if performance is broadly independent of the resources allocated, and instead determined primarily by the quantity of data available, then the task is said to be data-limited. Crucially, the allocation of additional processing resources to data-limited tasks is unlikely to lead to performance improvements. For example, individual differences in working memory skills seem unlikely to affect performance on an audiometric test, which depends primarily on whether the presented tones are above or below listeners' detection thresholds (as suggested by Heinrich et al., 2020). In real-life situations, the majority of tasks are neither fully data-limited nor fully resource-limited, but lie somewhere along a spectrum. Nevertheless, it is possible to characterise any given task in terms of the relative contributions of the two.

Support for the differential contributions of data and resource limits during speech-in-noise listening can be found in a number of studies. For example, Knight et al. (2023) found that cognitive abilities predict speech-on-speech listening performance most robustly when EM is reduced or absent. In other words, when data availability is high, performance is more strongly determined by the availability of processing resources than when data availability is lower. Similarly, Janse and Andringa (2021) showed that older adults' hearing sensitivity was a better predictor of performance for filtered (i.e., perceptually degraded) speech than unfiltered speech, whereas working memory capacity showed the opposite pattern. In other words, when data availability is high (unfiltered speech), performance is constrained by the available cognitive resources, whereas when data availability is low (filtered speech), performance is constrained by the amount of additional data that can be gleaned through good

72   hearing acuity. This chimes with the review conducted by Humes (2007), which suggests that

73   the importance of cognition is most pronounced for older adults once audibility is above a

74   given threshold. O'Leary et al. (2023), meanwhile, found that inserting silent pauses into

75   rapid speech improved recall performance when speech was moderately, but not severely,

76   degraded. This finding implies that a data limit was present in the latter case which prevented

77   additional cognitive processes (i.e., the extra processing time provided by the pauses) from

78   aiding performance. A similar reduction in benefit from silent pauses was also demonstrated

79   for CI users with poorer, but not better, hearing sensitivity (O'Leary et al., 2023).

80   The precise position of any given speech-on-speech task on the data- to resource-limit

81   spectrum will depend critically on a number of factors. As outlined above, EM – and hence

82   data availability – during speech-on-speech listening is determined by spectro-temporal

83   overlap between target and masker speech. However, speech-on-speech listening also

84   requires the ability to break down the acoustic scene into separate streams and selectively

85   direct attention to the target stream. Cues which support this streaming process and

86   subsequent maintenance of attention do not always overlap with cues affecting EM, and they

87   can exert an independent influence on performance. Thus, the position of a speech-on-speech

88   task on the data- vs. resource-limit spectrum depends on (a) the degree of EM (i.e., glimpsing

89   opportunities) and (b) the presence of streaming cues (such as fundamental frequency

90   differences between the two voices, e.g., Darwin et al., 2003).

91   When EM is moderate and/or there are acoustic cues available to aid stream segregation,

92   there is some disruption to data availability, but this can be ameliorated by the application of

93   cognitive resources (e.g., Rönnberg et al., 2022b). Within this portion of the data- to

94   resource-limit spectrum, the relationship between cognitive ability and listening performance

95   is positive (i.e., additional cognitive resources result in a performance benefit). The task

96   therefore falls towards the resource-limit end of the spectrum. However, as EM becomes

4

97   increasingly high and cues to streaming become less available, it seems reasonable to assume

98   that performance starts to tend towards a data limit: there is simply not enough of the target

99   speech available, or enough reliable streaming cues, for the application of cognitive resources

100   to provide a substantial benefit. At this point, then, one might expect a weaker (or even

101   absent) relationship between cognition and performance. In other words, the task falls

102   towards the data-limit end of the spectrum.

103   **1.2 Release from masking**

104   One well-documented way to increase the amount of data available to listeners during

105   speech-perception-in-noise tasks is to spatially separate the target and the masker (Culling &

106   Stone, 2017). As the two are moved apart, there is an increase in the number and duration of

107   spectro-temporal regions in which the target energy exceeds that of the masker for a given ear

108   (Edmonds & Culling, 2006), thus affording more glimpses. In other words, there is a change

109   in the degree of EM. In addition to this increase in glimpsing potential, strong location-based

110   streaming cues are also introduced, with the target and masker being reliably associated with

111   different positions in space (Allen et al., 2011; Kidd et al., 2005). These additional glimpses

112   and enhanced streaming cues can give rise to large improvements in performance, with some

113   studies reporting as much as a 12 dB reduction in the signal-to-noise ratio (SNR) required to

114   achieve a set performance level (Freyman et al., 1999) – a phenomenon known as spatial

115   release from masking (spatial RM; Litovsky, 2012).

116   Similar results have been obtained when a target and masker are presented not just in

117   different real or perceived spatial locations, but also when they are presented dichotically

118   over headphones (i.e., with the target and masker presented to separate ears; e.g., Calcus,

119   2015; Knight et al., 2023). Although dichotic presentation precludes the kind of central

120   auditory processing involved in true spatial listening – such as the use of interaural time

121   differences (ITDs) – it nevertheless provides both a reduction of EM (to effectively zero at

122   each ear) and strong location-based cues to streaming (derived from attending to one or the

123   other ear specifically). Such location-based cues also induce extreme interaural level

124   differences (ILDs), with the signal at a given ear being entirely absent from the opposite ear.

125   In that sense, dichotic release from masking can be seen as a special case of spatial release

126   from masking.

127   Release from masking can also be obtained by separating the target and masker in the

128   spectral domain (spectral RM). As with spatial RM, presenting a target and masker in non-

129   overlapping frequency bands makes more glimpses of the target data available to the listener

130   and provides an additional cue to streaming, and has been shown to improve speech

131   perception performance compared to when the target and masker are spectrally overlapping

132   (Apoux and Healy, 2010; Kidd et al., 2010; O'Neill et al., 2019). Other types of release from

133   masking have been shown, such as temporal release from masking (e.g., Gifford et al., 2012);

134   however, the focus of this study is on spatial (specifically dichotic) and spectral release.

135   Little is known about the relationship between dichotic and spectral RM. Specifically, the

136   questions of whether one type of RM is more beneficial than the other, and whether any extra

137   benefit is accrued when both are present, remain open. The underlying mechanisms driving

138   each type of RM and the role of individual differences are also currently unclear. As

139   described above, both dichotic and spectral RM make more of the target data perceptually

140   available by directly reducing EM; they also both provide a clear cue to streaming (location-

141   and frequency-based, respectively). However, dichotic and spectral RM may rely on these

142   two mechanisms (EM reduction vs. streaming) to different degrees. Furthermore, these

143   mechanisms may be underpinned by different cognitive functions, leading to individual

144   differences in the ability to exploit one type of RM compared to the other.

145   The question of which type of RM is more beneficial may depend on: (a) whether the

146   portions of the signal unmasked through EM reduction provide more data in one case than the

147   other; and (b) whether location or frequency provides a more salient cue for streaming – in

148   other words, whether listeners are more or less able to direct and maintain their attention to a

149   spatial vs. a spectral region. Existing empirical work using speech stimuli to address these

150   questions is relatively sparse, but there is some suggestion that listeners rely more heavily on

151   spatial cues than on the spectral cues provided by, e.g., talker gender (Allen et al., 2008).

152   However, results from a study by Best et al. (2013) using overlapping vs. non-overlapping

153   frequency bands showed robust effects of both spatial and spectral separation. It also appears

154   to be the case that spatial and spectral cues affect different aspects of the segregation and

155   streaming process during speech-on-speech listening, and that these effects depend on

156   whether attention is explicitly oriented to the spatial or spectral domain (Ihlefeld & Shinn-

157   Cunningham, 2008). In short, the picture is mixed.

158   The question of whether extra benefit is gained when both types of RM are present may again

159   depend on whether one type of RM unmasks additional useful information compared to the

160   other. However, it may also depend on whether streaming is facilitated by the presence of

161   multiple cues compared to a single cue. Best et al. (2013) found the highest performance

162   when both spatial and spectral RM were present. However, this study used concatenated

163   individual words as target sentences and nonsense maskers. Additionally, the spatial

164   separation involved only a lateralisation of the masker. The generalisability of these findings

165   to a situation involving naturally-produced, meaningful targets and maskers with larger

166   degrees of spatial separation is unclear.

167   **1.3 Cognitive contributions to RM effects**

168   In addition to questions about the nature of RM effects *per se*, it is also unclear how the

169   bottom-up, perceptual benefits offered by RM interact with listeners' deployment of

170   cognitive resources. It has been suggested that cognitive resources support speech perception

171   when the signal is degraded (Rönnberg et al., 2022), implying that the relationship between

172   individual differences in cognitive abilities and performance should be strongest when target

173   and masker overlap spatially or spectrally (i.e., when there is no RM). However, the data-

174   limit account described earlier (cf. Norman & Bobrow, 1975; Knight et al., 2023) suggests

175   that cognitive resources have relatively lower utility when the available target information is

176   severely reduced by the masker, implying that the cognition/performance relationship should

177   in fact be weakest when there is no RM. Furthermore, and as discussed above, the different

178   mechanisms implicated in RM (EM reduction vs. streaming) may recruit different cognitive

179   skills, and do so differently for different types of RM. We may therefore see a relatively weak

180   correlation between dichotic and spectral RM. In other words, any given listener may vary in

181   their ability to exploit one type of RM compared to the other. We may also see a different

182   relationship between cognition and each type of RM, with particular cognitive abilities

183   contributing differentially to the ability to exploit dichotic vs. spectral cues.

184   **1.4 The current study**

185   In the current study, we addressed these questions via a selective listening paradigm in which

186   participants were asked to track and transcribe the speech of one of two simultaneously-

187   presented talkers. On each trial, participants heard a male talker and a female talker

188   simultaneously over headphones. Each talker was uttering a meaningful but low-

189   predictability sentence (e.g., M: *The box was thrown beside the parked lorry*; F: *Glue the*

190   *paper to the dark blue background*). Listeners were instructed before the trial to report the

191   sentence spoken by either the male or the female talker. Perceived talker location was

192   manipulated, with the talkers presented via headphones either dichotically or collocated

193   (diotically). The speech was also filtered into frequency bands such that the talkers were

194   either spectrally overlapping or interleaved. There were thus four listening conditions: no

195   RM, dichotic RM only, spectral RM only, or both dichotic and spectral RM (see Figure 1).

196   With respect to implementing spatial RM in our experiments, dichotic listening was used as a

197   way of simulating maximal spatial separation. As discussed above, true spatial hearing is best

198   implemented using a combination of interaural time differences (ITDs) and interaural level

199   differences (ILDs), reflecting the cues used during central auditory processing to determine

200   sound source location. Although dichotic presentation does not allow for ITDs, it

201   nevertheless does create ILDs, which have been shown to be a dominant spatial cue at high

202   frequencies and to contribute to the localization of low-frequency sounds (Middlebrooks &

203   Green, 1991; Middlebrooks & Onsan, 2012). As such, dichotic listening meets our goals of

204   reducing energetic masking and improving streaming through increased perceived left vs.

205   right (L-R) separation. A similar ILD-only method was used successfully by Knight et al.

206   (2023) to simulate a four-step change in spatial location using L-R mixtures with graded

207   intensity differences, and by Calcus et al. (2015), who compared various dichotic listening

208   conditions with a diotic baseline to evaluate spatial masking release. Nevertheless, it must be

209   noted that dichotic listening was used as a proxy for spatial hearing in the current study, and

210   that key differences – particularly with regard to the role of central auditory processing –

211   exist between the two.

212   To address the contribution of cognitive abilities to performance in general, and to RM

213   benefits in particular, a battery of six cognitive tasks was administered. These tasks were

214   chosen to cover a complementary range of relevant dimensions of WM and attention (e.g.,

215   Baddeley, 2000; Diamond, 2013). To assess short-term phonological memory, we used Non-

216   word Repetition and Rhyme Judgement tasks, both of which have been used to demonstrate a

217   relationship between speech perception in noise and the ability to manipulate phonological

9

218   information in WM (e.g., Millman & Mattys, 2017; Dryden, 2018). To assess attention, we

219   used visual versions of Rapid Serial Visual Presentation (RSVP) and Singleton tasks. These

220   two tests were included to reflect the role of visual attention as a reliable predictor of spatial

221   listening (e.g., Clayton et al., 2016). To assess executive function, we used the Reading Span

222   and Letter-Number Sequencing (LNS) tasks. The former represents a widely-used index of

223   complex WM in speech-in-noise studies known to tap into both storage and executive control

224   of non-auditory verbal information (e.g., Rönnberg et al., 2022). The LNS, meanwhile,

225   provides a measure of simple WM, with a greater emphasis on storage and lower demands on

226   executive function, but which has nevertheless been shown to predict speech-in-noise

227   performance in a variety of contexts and listener groups (e.g., Knight et al., 2023; Heinrich &

228   Knight, 2016).

229   Although these tasks were chosen to cover a range of potentially relevant cognitive

230   dimensions, it was unclear which skills might contribute most strongly to performance on our

231   listening task. Each individual cognitive measure also involves task-specific demands, which

232   are not necessarily relevant to either the cognitive construct under assessment or to speech-in-

233   noise perception. Additionally, it was desirable to keep the analyses relatively parsimonious

234   to aid interpretation. We therefore used factor analysis to derive a single cognitive score for

235   each participant from their individual scores on the six tasks, with the aim of reducing task-

236   specific noise and ensuring the relative contributions of each task were suitably represented.

237   In terms of listening performance, we expected to see main effects of both spectral and

238   dichotic RM, with performance being higher in the spectrally interleaved than the spectrally

239   overlapping conditions and in the dichotic than in the collocated conditions. If combining

240   both types of RM provides a simple additive benefit, then we would expect the best

241   performance of all in the condition featuring both dichotic and spectral RM but no 2-way

242   interaction. However, it seems plausible to assume that dichotic and spectral RM provide

243   similar benefits in terms of EM reduction – albeit in combination with qualitatively different

244   streaming cues – and thus that we would see only a relatively small extra benefit of adding a

245   second type of RM. If that is the case, we would expect to see a 2-way interaction of dichotic

246   RM x spectral RM, with the improvement in performance resulting from the addition of a

247   second type of RM being somewhat smaller than the improvement resulting from the addition

248   of either dichotic or spectral RM alone.

249   In terms of the relationship between transcription performance and cognitive abilities, if

250   cognitive abilities are most likely to contribute to speech perception in conditions in which

251   EM is moderate and/or there are cues to streaming, then we expect to see a strong positive

252   relationship between listening task scores and cognitive task scores when one or both types of

253   RM are present. However, when no RM is present – in other words, when EM is severe and

254   there are no cues to streaming – we expected the data limit to attenuate – or perhaps even

255   fully suppress – the contribution of cognition to performance, thus leading to the weakest

256   relationship between cognition and performance in this listening condition.

257   **2. Methods**

258   **2.1 Participants**

259   Participants were recruited via the Human Participant Pool System at the University of

260   (managed by SONA; www.sona-systems.com) and the online recruitment platform Prolific

261   (www.prolific.com). All participants provided informed consent before taking part. Testing

262   was conducted via the online experiment builder Gorilla (www.gorilla.sc; Anwyl-Irvine et al.,

263   2020) between 8/11/23 and 24/04/24. Participants recruited through SONA were rewarded with

264   course credit, and participants recruited through Prolific were paid 12 GBP for their time. All

265   participants were self-reported monolingual native speakers of British English with normal

266   hearing and vision, had no self-reported language or reading related disorders, and were aged

267   between 18 and 35 (mean age = 29 [SD = 4.17], female/male = 123/117). Recruitment filters

268   were applied on Prolific, so that only people with a Prolific approval rating of 90% and above

269   could access the study. Approval ratings reflect the number of Prolific studies for which a

270   participant has not been rejected by the researcher (e.g., because of attention check failures),

271   and therefore indicate level of compliance and engagement in previous studies.

272   Before starting the experiment, participants heard a brief segment of white noise played at the

273   same RMS level as the stimuli in the main listening task and were instructed to adjust the

274   volume to an audible and comfortable listening level. They then completed a headphone check

275   based on a task designed and validated by Woods et al. (2017). On each trial, participants were

276   asked to identify the quietest of three tones, and the task consisted of 6 trials in total. The use

277   of antiphase audio for some tones meant that this task could only be completed correctly if

278   participants wore stereo headphones. If a participant responded incorrectly on 2 or more trials

279   out of 6, the headphone check was repeated. If they again responded incorrectly on 2 or more

280   trials, they were not allowed to proceed to the study.

281   Power analyses based on pilot studies run in our lab using comparable methodology suggested

282   that the study would require approximately 240 participants (Knight et al., 2023). This sample

283   size should allow detection of both an effect of dichotic/spectral RM (assuming effect sizes

284   equivalent to those in the pilot data) and any significant correlations between our continuous

285   measures (i.e., cognitive task scores) and performance across conditions (assuming correlation

286   coefficients of .25, power of .90, and alpha levels corrected for multiple comparisons).

287   **2.2 Overall Procedure**

288   Participants took part in two sessions. Session 1 consisted of the listening task. At least 24

289   hours after the completion of the listening task, participants completed Session 2. Session 2

290   consisted of six cognitive tasks, designed to tap into three key components of WM/attention:

291   two phonological short-term memory tasks (Non-Word Repetition [Phon_NW] and Rhyme

292   Judgement [Phon_RJ]), two attention tasks (Rapid Serial Visual Presentation [Att_RSVP] and

293   Singleton Task [Att_Singleton]), and two executive function tasks (Letter-Number Sequencing

294   [Exec_LNS] and Reading Span [Exec_RSpan]). In total, Session 1 and Session 2 took around

295   two hours to complete.

296   **2.3 Session 1: Listening Task**

297   *2.3.1 Materials*

298   The stimuli for the listening task consisted of 184 meaningful sentences drawn from a set of

299   750 sentences (IEEE sentences; Rothauser et al., 1969) modified to fit modern British English

300   (see Stubbs & Summerfield, 1990). From the selected sentences, 160 were used to form the 80

301   sentence pairs used in the main listening task, and 24 were used to form 16 practice trials (eight

302   trials with one sentence and eight trials with sentence pairs). All sentences were spoken by the

303   same native male talker and native female talker of British English. The energy spectrum of

304   each sentence was root-mean-square (RMS) equalised to 0.06 Pa and its long-term average

305   spectrum (LTAS) was matched to the average LTAS of all the sentences before being band-

306   pass filtered into four non-overlapping frequency bands (see Table 1). The number of bands

307   and centre frequencies were chosen via informal piloting to maximise intelligibility. Each band

308   had a fixed smoothing window of 50 Hz. The bands had logarithmically spaced centre

309   frequencies. The upper and lower cut-off frequencies of the bands were determined via

310   estimations of the equivalent rectangular bandwidths (ERBs): estimations were calculated for

311   centre frequencies equivalent to the upper and lower cut-offs, and cut-offs chosen which

312   resulted in no overlap of stimulation on the basilar membrane between the bands after

313   smoothing. The bands were then recombined in pairs to create two types of sentences: those

314   made of odd bands (Bands 1 and 3) and those made of even bands (Bands 2 and 4). The

intelligibility of these sentences was verified through pilot testing. Pilot participants (N = 21)

were recruited and tested via Prolific and Gorilla, and the sentences were presented using the

same general procedure as for the main task, including a headphone check. Each participant

heard 80 sentences, with counterbalancing ensuring that (a) each sentence was heard in its male

and female versions, and in odd bands vs. even bands, an equal number of times and (b) each

participant heard both an equal number of male and female sentences and an equal number of

odd-band and even-band sentences. Mean transcription performance for each sentence type

(MaleOdd, MaleEven, FemaleOdd, FemaleEven) was 91.95%, 85.75%, 78.50%, and 76.55%

respectively. The sentences were then paired to match duration as closely as possible, with each

pair consisting of one female and one male talker. The sentences were paired to create two

spectral conditions: with spectral overlap ("Overlapped"; i.e., FemaleOdd + MaleOdd or

FemaleEven + MaleEven) and without spectral overlap ("Interleaved"; i.e., FemaleOdd +

MaleEven or MaleOdd + FemaleEven).

Table 1

*Lower and upper cut-offs of the four frequency band filters and their logarithmic centre
frequencies*

| Band	Lower cut-off (Hz)	Upper cut-off (Hz)	Lower cut-off with smoothing window (Hz)	Upper cut-off with smoothing window (Hz)	Centre frequency (Hz)
1	138	218	88	268	173
2	375	713	325	763	517
3	927	2579	877	2629	1546
4	3006	7120	2956	7170	4626

Two versions of each pair were generated to also create two location conditions: collocated

(both sentences combined to make mono stimuli played in both channels) and dichotic (each

sentence played in a separate stereo channel, one to each ear). Prior to mixing, each band-pass

filtered sentence was root-mean-square (RMS) equalised to the same level (0.06 Pa). Following

mixing, all stimuli were once again RMS-equalised to the same overall level. The mixing

338   process therefore created sentence pairs in four conditions: Collocated_Overlapped,

339   Dichotic_Overlapped, Collocated_Interleaved, and Dichotic_Interleaved.

340   **Error! Reference source not found.** presents a schematic illustration of the four conditions.

341   The most challenging condition, Collocated_Overlapped, includes neither spatial nor spectral

342   RM. In contrast, the least challenging condition, Dichotic_Interleaved, includes both spatial

343   and spectral RM. The intermediate conditions have either spatial RM (Dichotic_Overlapped),

344   or spectral RM (Collocated_Interleaved), but not both.

345   **Figure 1**

346   *Schematic of the four experimental conditions in the listening task*

Collocated_Overlapped   **Dichotic_Overlapped**

Collocated_Interleaved   **Dichotic_Interleaved**
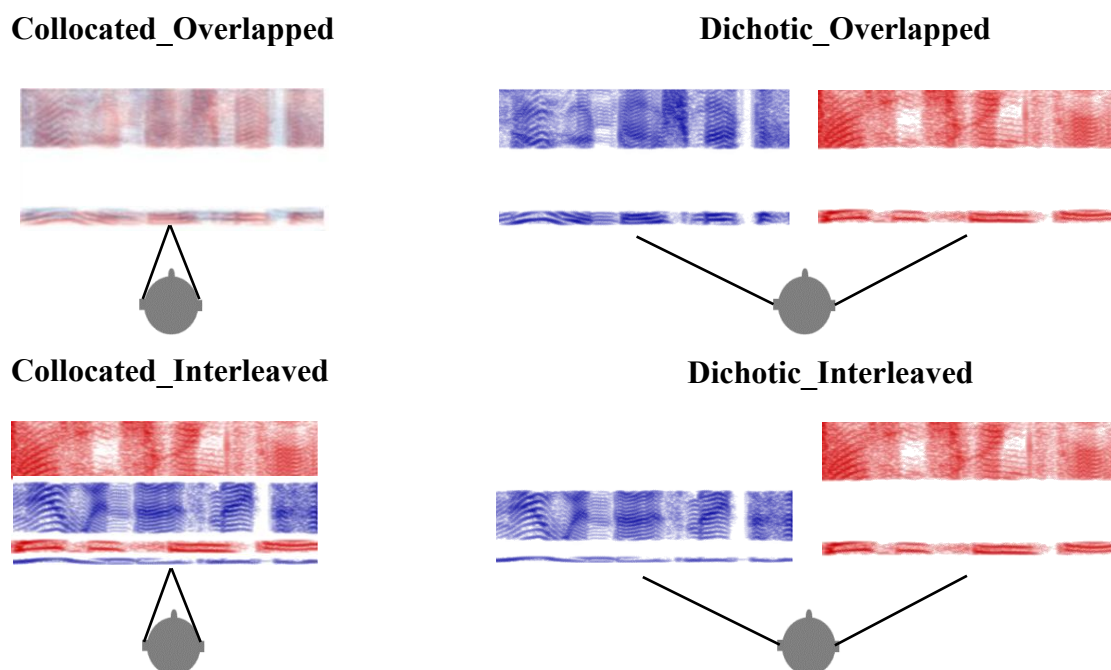
347   *Note.* Participants were asked to track one of the two voices. The spectral overlap (overlapped,
348   interleaved) and spatial location of the two voices (collocated, dichotic) were manipulated,
349   leading to listening conditions with no RM, either dichotic or spectral RM, or both types of
350   RM. Note that this figure is a schematic only and does not reflect the relative width of the actual
351   frequency bands used. Please see Table 1 for boundary values of the bands.

352   Each target sentence in the listening task contained five keywords which were used for scoring

353   transcription accuracy expressed as a proportion of keywords correctly transcribed. A typed

354   letter string was scored as correct if it omitted one letter, if it added one letter, if two consecutive

355 letters were transposed (e.g., 'lihgt' rather than 'light'), or if it was phonetically the same as

356 the keyword (e.g., 'lite' rather than 'light'). Strings of words in which spaces were omitted

357 were separated if they made up two correct consecutive words in the sentence. Practice trials

358 were excluded from the analysis.

359 ***2.3.1 Procedure***

360 Participants first adjusted their volume and completed the headphone check task described

361 above. Participants who passed the headphone check proceeded to the main listening task, in

362 which they heard the sentence pairs described above. Participants were instructed to pay

363 attention to one of the two talkers in each pair and report what the talker said at the end of each

364 trial. They were not told that the sentences contained keywords or how they would be scored.

365 The assignment of target talker (male or female) remained the same throughout the task for

366 each participant and was counterbalanced across participants. Thus, a given participant was

367 asked to report the same voice throughout the experiment. On each trial, participants were

368 presented with a fixation cross on the display screen whilst the audio stimulus played. At the

369 end of the audio stimulus, participants were prompted to type the target sentence into a response

370 box and press the enter key to commence the following trial. They were given a maximum of

371 60 seconds to respond before the next trial started.

372 The listening task consisted of 80 trials, with a break after 40 trials. The task used a two-by-

373 two factorial design, with location (collocated/dichotic) and spectral overlap

374 (overlapped/interleaved) as two fully-crossed independent variables, creating four conditions

375 of 20 trials each.

376 All participants heard the same 80 sentence pairs, but the allocation of each unique sentence

377 pair to conditions was counterbalanced across participants. Within each of the 4 conditions,

378 there were 2 blocks of 10 trials each, with 8 blocks in total. Within each overlapped condition,

16

379   one block consisted of FemaleOdd + MaleOdd sentences, and the other block consisted of

380   FemaleEven + MaleEven sentences. Within each interleaved condition, one block consisted of

381   FemaleOdd + MaleEven sentences, and the other block consisted of FemaleEven + MaleOdd

382   sentences. The presentation order of blocks, and the presentation order of sentence pairs within

383   each block, were randomised. For each unique sentence pair, the two sentences were presented

384   in all possible combinations of male/female and band manipulation (odd- or even-band filtered)

385   an equal number of times across all participants. Each participant heard an equal number of

386   sentences from male and female talkers, as well as an equal number of sentences filtered by

387   odd and even bands.

388   The 16 practice trials were presented at various points throughout the experiment. Two practice

389   trials were played prior to the listening task, and two practice trials were played between the

390   blocks of the listening task. For each pair of practice trials, the first trial consisted of the target

391   talker heard in isolation in the same bands (odd or even) as in the block which followed. The

392   second practice trial consisted of a sentence pair with the same band and spatial manipulation

393   as the sentence pairs in the following block. The practice trials ensured that participants had

394   experienced the spatial location and masking configuration of the following block before the

395   block commenced.

396   **2.4 Session 2: Cognitive Tasks**

397   Session 2 consisted of six cognitive tasks as described in **Overall Procedure**.

398   *2.4.1 Materials*

399   **Non-word Repetition Task (Phon_NW).** Participants were required to decide if two

400   auditorily-presented sequences of monosyllabic non-words were identical or not. The

401   sequences were presented in two conditions: deviant, where the two sequences differed by one

402   vowel phoneme positioned randomly within the sequence, and non-deviant, where identical

403 sequences were played. There were 28 trials (14 deviant, 14 non-deviant), with the sequence

404 increasing by one non-word every four trials, from a minimum of two to a maximum of eight

405 non-words per sequence. Within a set of four trials, trial presentation order was randomised

406 across participants. After each trial, participants responded by clicking the button labelled

407 'same' or 'different' displayed on their screen with no time limit.

408 The stimuli were generated from the English Lexicon Project (Balota et al, 2007). The initial

409 search criteria consisted of non-words with no orthographic neighbours which were 3-6

410 characters in length. From these, monosyllabic non-words which did not bear a strong

411 similarity to an identifiable word or name were selected. The non-words were created using

412 text-to-speech software (the female talker "Amy" from ttsmp3.com). Deviant non-words were

413 prosodically matched to their non-deviant (identical) counterparts as far as possible. There was

414 a 500-ms gap between non-words within a sequence and a 2000-ms gap between sequences.

415 A *d'* score was calculated to measure participants' ability to discriminate between sequences

416 whilst also accounting for any bias towards choosing either 'Same' or 'Different'. *d'* was

417 measured using the proportion of 'Different' trials that received a 'Different' response and the

418 proportion of 'Same' trials that received a 'Different' response. The *d'* scores were then

419 standardised (*z*-score) and used as the Phon_NW score in further analyses.

420   **Rhyme Judgement Task (Phon_RJ).** Participants were instructed to determine as

421 quickly and accurately as possible if two visually-presented words rhymed with each other.

422 Word pairs were taken from Johnston and McDermott (1987) and formed four conditions:

423 rhyme/orthographically similar (R+O+), rhyme/orthographically dissimilar (R+O-), does not

424 rhyme/orthographically similar (R-O+), does not rhyme/orthographically dissimilar (R-O-).

425 Potentially unfamiliar words and words with possible individual variability in phonological

426 representation were replaced. All words were four letters long and were presented in

18

427 monospace font to ensure that word length was not indicative of rhyming status. Participants

428 used their keyboard to respond by pressing the A key for 'no' and L key for 'yes', with visual

429 labels on screen to remind participants which button referred to which response. There was no

430 time limit for participants to respond. Following the button press, there was an 800-ms gap

431 before the next word pair. Trial order was randomised across conditions. There were 92 trials

432 in total (23 trials per condition).

433 Both response accuracy and reaction time were measured, but only the R+O- and R-O+

434 conditions were used in the analyses due to likely ceiling effects in the R+O+ and R-O- trials

435 (Dryden, 2018). Response accuracy was the proportion of correct incongruent trials, and

436 reaction time was the average reaction time of correct incongruent trials. Both measures were

437 *z*-transformed and integrated using the Balanced Integration Score method (BIS; Liesefeld &

438 Jancyzk, 2019). In this way, each participant received a score which took into account speed-

439 accuracy trade-offs which may occur when participants are asked to answer 'as quickly and as

440 accurately as possible'.

441 **Rapid Serial Visual Presentation Task (Att_RSVP).** Participants were presented

442 with two rapid streams of sequentially presented letters on either side of a central fixation cross.

443 In the selective-attention condition, participants were instructed to fixate on the cross whilst

444 monitoring one of the two streams (indicated by an arrow prior to each trial). In most trials, a

445 number (3, 4, 6, 7, 8, or 9) appeared at some point in one of the streams, and participants

446 indicated whether or not the number was presented on the cued side by pressing the 'Y' key

447 for 'Yes' or the 'N' key for 'No'. In the divided-attention condition, participants were

448 instructed to fixate on the cross whilst monitoring both streams. A number appeared at some

449 point in one of the streams, and participants indicated on which side the number was presented

450 by pressing 'A' for left, 'L' for right, or spacebar for neither. For both conditions, visual

451 reminders remained on screen to indicate which button referred to which response. Five trials

19

452 within each condition did not include a number. Participants were informed of this to avoid

453 tactical responding in the divided attention condition (i.e., using the absence of a number in

454 one stream to infer its presence in the other).

455 The target number appeared in each stream (left/right) an equal number of times in each

456 condition. The position of the number within each sequence was randomised but was never the

457 first or the last in the sequence. There were 10 letters in each letter stream, with each pair of

458 letters (left/right) presented for 50 ms and no gaps between pairs. Letters I, Z, S and numbers

459 1, 2, 5 were excluded due to visual similarity. Conditions were blocked, but block presentation

460 order was counterbalanced across participants. There were 70 trials in total (35 selective

461 attention, 35 divided attention).

462 The difference in performance accuracy between the divided and selective conditions was

463 calculated for each participant. This allowed for the cost of dividing attention between two

464 sources of information to be assessed. This difference was z-transformed and multiplied by -1

465 to ensure larger values reflected better performance as in the other cognitive tasks.

466   **Singleton Task (Att_Singleton).** Participants were presented with eight outlined

467 circles evenly positioned around the perimeter of a larger imaginary circle. Seven circles

468 contained a letter, and one circle contained a number (1-4), which served as the target. The

469 circles were presented for one second. The larger imaginary circle was divided in two

470 vertically, with each half containing four of the smaller circles. Participants were instructed to

471 pay attention to only the set of circles on one half of the perimeter (left/right), indicated prior

472 to each trial. The target number was always shown in this half, and participants had to press

473 the corresponding number on their keyboard as quickly as possible when they detected it. In

474 the "with-singleton" condition, one of the non-target circles was simultaneously presented in

475 red (i.e., a visual singleton). The visual singleton was always in the opposite half of the circle

20

as the target number, to maximise distraction from the target. In the "without-singleton" condition, there was no visual singleton present. There were 64 trials (32 per condition), and trial order was randomised across conditions. Each number was presented in each circle twice, once in each condition. For each number, there was a singleton in each of the eight circles once. Letters I and Z were removed due to visual similarity to numbers 1 and 2.

Reaction time (RT) was measured as the time between stimulus onset and the participant's keyboard response. The z-transformed difference in reaction times (RTs) between correctly answered with-singleton and without-singleton trials was used to score performance. This difference assessed the cost of inhibiting a distractor (i.e., a singleton).

Letter-Number Sequencing Task (Exec_LNS). Participants heard a sequence of numbers and letters spoken by a male talker. After hearing each sequence, they were instructed to type the numbers in ascending order, followed by the letters in alphabetical order (Wechsler, 1997). Sequences began with a length of two items and increased in length by one item every three trials, with a maximum length of eight. Each item was presented 1s after the onset of the previous item. If all three trials of a particular length were answered incorrectly, the task was terminated.

Each trial was scored as either correct or incorrect. A trial was scored as correct only if there were no discrepancies between the response and the sequence presented. Each participant received a score as the total number of trials correctly responded to (with a maximum of 21). The scores were then z-transformed.

Reading Span Task (Exec_RSpan). Participants were required to read sets of written sentences, which were presented one word at a time. Each word was presented for 500 ms, with no gaps between words. Following the presentation of each sentence set, participants were instructed to type the final words of each sentence in the set and respond to a question relating

21

500   to the semantic content of one sentence within that set. Sentences were taken from the original

501   Daneman and Carpenter (1980) reading span task and modified to remove any outdated or

502   potentially distressing content. There were eight trials in total. After every two trials, the length

503   of the set increased by one sentence, from a minimum of three sentences to a maximum of six

504   sentences per set.

505   Trials were scored as correct if the participant correctly transcribed all final words of each

506   sentence set. The scoring rules for misspelt words were the same as in the listening task.

507   Performance was calculated as the *z*-transformed proportion of correct trials. Responses to the

508   semantic questions were recorded but not used in further analysis.

509   *2.4.2 Procedure*

510   Participants were instructed to use the same computer and headphone set-up as in Session 1.

511   Session 2 began with a simple audio check. First, white noise was played at the same RMS

512   level as in Session 1, and participants adjusted their volume to an audible and comfortable

513   listening level. Following this, three trials of three spoken digits were played, and participants

514   were asked to type these digits in the order they were presented. Before being included in

515   analysis, participant performance was reviewed to ensure that at least two trials had been

516   answered correctly.

517   Participants then performed the six cognitive tasks in the following order: 1) Non-word

518   Repetition Task (Phon_NW); 2) Rhyme Judgement Task (Phon_RJ); 3) RSVP Task

519   (Att_RSVP); 4) Singleton Task (Att_Singleton); 5) Letter-Number Sequencing Task

520   (Exec_LNS); 6) Reading Span Task (Exec_RSpan). Session 2 took around one hour to

521   complete.

522   **3. Analyses**

523 All analyses were conducted using R (version 4.4.0, embedded in R Studio version 2024.04.2)

524 and the packages *dplyr* (1.1.4), *emmeans* (1.10.1), *lme4* (1.1-35.3) and *psych* (version 2.5.6).

525 To investigate the general effects of spectral and dichotic RM in the listening task, we fitted

526 two Generalised Linear Mixed-Effects Models (GLMMs) with trial-level transcription scores

527 from the listening task as the outcome variable. We modelled each trial as a data point, using a

528 binomial distribution and a logit link. Both GLMMs had spectral separation, dichotic

529 separation, and their interaction as fixed effects. The first GLMM also included the six

530 cognitive task scores in interaction with dichotic separation and spectral separation to assess

531 whether the influence of each individual cognitive ability on performance varied across

532 listening conditions. For the six cognitive tasks, a correlation matrix was generated to explore

533 the strength and direction of the relationship between pairs of tasks and the internal consistency

534 of each task was assessed using Cronbach's alpha. In an attempt to generate a measure of

535 general cognitive ability for each participant, we conducted an exploratory factor analysis on

536 the scores from the six cognitive tasks using the *stats* package (4.4.0) in R. In this analysis, we

537 specified a single underlying factor, which was then used to generate a single score (CogFact)

538 for each participant via the regression method (DiStefano et al., 2009). We then incorporated

539 the individual CogFact scores (*z*-transformed) into a second GLMM, in interaction with

540 dichotic separation and spectral separation, to assess the influence of general cognitive ability

541 on listening performance.

4. Results

543 Performance under each listening condition is shown in Figure 2. The first GLMM included

544 main effects and interactions of Dichotic Separation (collocated, dichotic) and Spectral

545 Separation (overlapped, interleaved), along with six cognitive test scores. The model included

546 random intercepts and slopes for Dichotic Separation and Spectral Separation at both the

23

547   participant and item levels[1]. Performance for each participant on each trial of the listening task

548   was expressed as a proportion between 0 and 1. The two levels of Dichotic Separation and

549   Spectral Separation were sum-coded as -1 and +1, respectively. The full model specification

550   was as follows:

551   *Proportion_Correct ~ Dichotic Separation * Spectral Separation l * (Phon_NW +*

552   *Phon_RJ + Att_RSVP + Att_Singleton + Exec_LNS + Exec_RSpan) + (1 + Dichotic*

553   *+ Spectral | participant) + (1 + Dichotic Separation + Spectral Separation | item)* [2]

554   Using Generalised Variance Inflation Factors (GVIF, Dancey & Reidy, 2007), we determined

555   that there were no concerns regarding multicollinearity amongst the predictors, including the

556   six cognitive scores: all GVIF values were low (range: 1.00–1.35). Significance testing was

557   conducted via likelihood Ratio Tests (LRTs) comparing the full model to reduced models. The

558   main effect of Dichotic Separation was significant, χ^2 (1) = 350.66, $p < .001$, revealing

559   significantly worse performance in the collocated condition ($M = 0.48$, $SD = 0.38$) compared

560   to the dichotic condition ($M = 0.72$, $SD = 0.29$). The main effect of Spectral Separation was

561   also significant, χ^2 (1) = 563.76, $p < .001$, revealing significantly worse performance in the

562   overlapped condition ($M = 0.44$, $SD = 0.36$) compared to the interleaved condition ($M = 0.77$,

563      $SD = 0.27$). The main effects of Phon_RJ (χ^2 (1) = 17.91, $p < .001$), Att_RSVP (χ^2 (1) = 5.79,

564      $p < .05$), and Exec_LNS (χ^2 (1) = 6.57, $p < .05$) were significant, indicating that, overall, higher

565   Phon_RJ, Att_RSVP and Exec_LNS scores were associated with higher accuracy scores on

566   the listening task. There was also a significant interaction between Dichotic Separation and

---

[1] More complex models with any additional slopes of Dichotic Separation:Spectral Separation for participant and item intercepts failed to converge.

[2] This model produced a warning message – "Model is nearly unidentifiable: large eigenvalue ratio". A simpler model setting the number of Gauss-Hermite quadrature points (nAGQ) to 0 converged without error and produced very similar results to the model reported here.

567   Spectral Separation, $\chi^2(1) = 351.79$, $p < .001$. This interaction is described in more detail in

568   the second GLMM.

569   **Figure 2**

570   *Average performance for each Dichotic Separation and Spectral Separation condition in the*
571   *listening task*

572

573
574
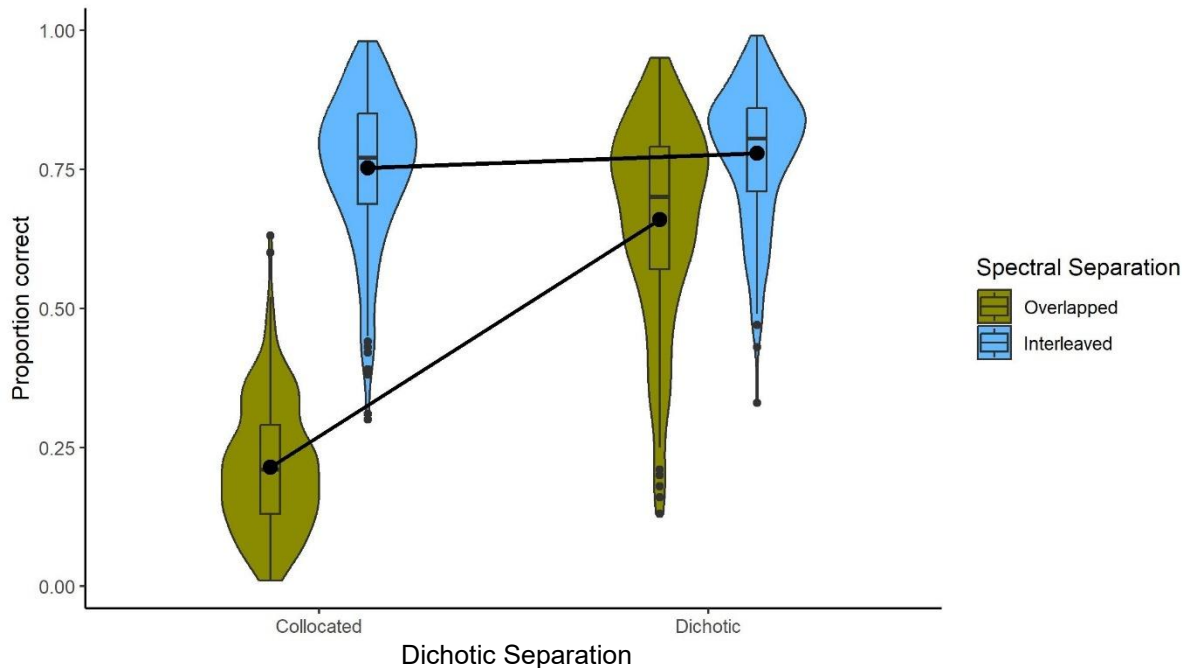
575   Bivariate Pearson correlation coefficients between cognitive tasks are reported in Table 2a.

576   Most of the cognitive tasks were positively correlated with each other. The phonological tasks

577   (Phon_NW and Phon_RJ) and the executive function tasks (Exec_LNS and Exec_RSpan)

578   correlated with the highest number of other cognitive tasks, while the two attention tasks

579   (Att_RSVP and Att_Singleton) did not correlate significantly with any other cognitive tasks.

580   Cronbach's alpha scores for each cognitive task are reported in Table 2b. For the Att_RSVP

581   task, two scores are reported: one for the selective attention condition and one for the divided

582   attention condition. Three tasks – one from each domain – showed poor internal consistency

583   (Phon_NW, Exec_LNS and Att_RSVP). However, the remaining three tasks produced good or

584   acceptable scores, with one (Phon_RJ) demonstrating excellent consistency.

585   **Table 2a**

586   *Correlation matrix (Pearson's r) of all six cognitive tasks*

	Phon_ NW	Phon_ RJ	Att_ RSVP	Att_ Singleton	Exec_ LNS	Exec_ RSpan
Phon_ NW	1.00	0.24	0.003	-0.03	0.32	0.24
Phon_ RJ	0.24	1.00	0.10	0.03	0.30	0.23
Att_ RSVP	0.003	0.10	1.00	-0.01	0.05	0.07
Att_ Singleton	-0.03	0.03	-0.01	1.00	0.11	0.06
Exec_ LNS	0.32	0.30	0.05	0.11	1.00	0.37
Exec_ RSpan	0.24	0.23	0.07	0.06	0.37	1.00

587

588   **Table 2b**

589   *Internal consistency (Cronbach's alpha scores) of all six cognitive tasks*

Phon_ NW	Phon_ RJ	Att_ RSVP (selective)	Att_ RSVP (divided)	Att_ Singleton	Exec_ LNS	Exec_ RSpan
0.38	0.95	0.37	0.45	0.81	0.55	0.73

590

591   For the second GLMM, the scores from the six cognitive tasks were entered into an exploratory

592   factor analysis to derive a single underlying cognitive factor. The uniqueness and loading

593   values of each cognitive task are reported in Table 3. Visual inspection of the scree plot and

594   associated eigenvalues suggested that one factor was sufficient to represent the data. This was

595   confirmed by a chi-square test ($p = .87$), indicating that the single-factor model represented the

596   data well.

597 **Table 3**

598 *Outputs of the factor analysis showing the uniqueness and loading values of each cognitive*
599 *task*

	Phon_NW	Phon_RJ	Att_RSVP	Att_Singleton	Exec_LNS	Exec_RSpan
Uniqueness	0.779	0.791	0.991	0.990	0.528	0.724
Loading	0.470	0.457	0.093	0.100	0.687	0.525

600

601 Individual factor scores were derived from this factor analysis to create a measure of overall

602 cognitive ability for each participant (labelled "CogFact"). The CogFact scores were then

603 incorporated into the second GLMM as an additional factor alongside Dichotic Separation and

604 Spectral Separation. This GLMM included all main effects and all two-way and three-way

605 interactions. As before, random intercepts and slopes for Dichotic Separation and Spectral

606 Separation were modelled at the participant and item levels[3], and the levels of Spectral

607 Separation and Dichotic Separation were sum-coded as -1 and +1. The full three-factor model

608 was specified as:

609 *Proportion_Correct ~ Dichotic Separation * Spectral Separation * CogFact + (1 +*

610 *Dichotic Separation + Spectral Separation | participant) + (1 + Dichotic Separation*

611 *+ Spectral Separation | item)*

612 The second GLMM had a slightly lower BIC (56846) compared with the first GLMM (57015),

613 suggesting that the use of the CogFact scores produced a better fit for the data than using the

614 six individual cognitive task scores.

615 The results of the second GLMM are presented in Table 4(a). This analysis produced similar

616 main effects and interaction of Dichotic Separation and Spectral Separation to the GLMM

---

[3] More complex models with any additional slopes of Dichotic:Spectral for participant and item intercepts failed to converge.

617   reported previously. In addition, there was a significant main effect of CogFact, χ^2 (1) = 35.59,

618    $p < .001$, suggesting that high cognitive capacity was associated with superior listening

619   performance. There were no significant two-way interactions involving CogFact. However, a

620   significant three-way interaction between CogFact, Dichotic Separation, and Spectral

621   Separation was obtained, χ^2 (1) = 11.32, $p < .001$. This interaction is illustrated in Figure 3.

622   Post-hoc tests using the *emmeans()* function provided the standard error values of each listening

623   condition   (Collocated_Overlapped   =   0.07,   Collocated_Interleaved   =   0.06,

624   Dichotic_Overlapped = 0.07, Dichotic_Interleaved = 0.07).   These values suggest that

625   variability was similar across the four listening conditions, indicating that there were neither

626   floor nor ceiling effects in any condition.

627   Post-hoc pairwise comparisons were conducted using emmeans() to unpack the significant

628   interaction between Dichotic Separation and Spectral Separation. The results are presented in

629   Table 4(b). All pairwise comparisons were significant (p < .05), but the estimates suggest that

630   performance   was   poorest   when   there   was   neither   dichotic   nor   spectral   RM

631   (Collocated_Overlapped) and best when there was both dichotic and spectral RM

632   (Dichotic_Interleaved). Of the two interim conditions, performance was better when only

633   spectral RM was present compared to when only dichotic RM was present.

634   To assess how the relationship between CogFact and task performance varied across the four

635   main listening conditions (i.e. the significant three-way interaction), we tested estimated

636   marginal trends (simple slopes) using the *emtrends()* function for all conditions and all possible

637   pairwise comparisons. The results are presented in Table 4(c). Inspection of the 95%

638   confidence intervals for the estimated CogFact slopes indicated that all intervals excluded zero,

639   suggesting that CogFact significantly predicted performance in each condition. The pairwise

640   comparisons showed no significant differences between conditions, but the estimated slope of

641 CogFact under Collocated_Overlapped was the smallest, reflecting a weaker positive

642 relationship between CogFact and listening performance in the most challenging condition.

643 **Table 4**

644 *(a) Results from Likelihood Ratio Tests (LTRs) conducted for the overall GLMM with*
645 *CogFact, Dichotic* Separation *and Spectral* Separation

	CogFact	Dichotic Separation	Spectral Separation	CogFact: Dichotic Separation	CogFact: Spectral Separation	Dichotic Separation: Spectral Separation	CogFact: Dichotic Separation: Spectral Separation
χ^2	35.59	349.95	562.22	1.60	0.55	348.85	11.32
p value	< .001 ***	< .001 ***	< .001 ***	.21	.46	< .001 ***	< .001 ***

646

647 *(b) Post-hoc comparisons of listening performance across pairs of listening conditions*
648 *(Tukey corrected)*

|	C_I – D_I	C_I – C_O	D_I – D_O	C_O – D_O	C_I – D_O	D_I – C_O
Estimate	-0.18	2.82	0.69	-2.31	0.51	3.00
Standard error	0.07	0.07	0.07	0.08	0.08	0.08
p value	< .05 *	< .001 ***	< .001 ***	< .001 ***	< .001 ***	< .001 ***

649

(c) Post-hoc tests of the relationship between CogFact and listening performance across
651 *listening conditions (Tukey corrected)*

	C_I	C_O	D_I	D_O		
Estimate	0.31	0.23	0.30	0.34		
Standard error	0.06	0.06	0.06	0.06		

	C_I – D_I	C_I – C_O	D_I – D_O	C_O – D_O	C_I – D_O	D_I – C_O
Estimate	0.01	0.08	-0.03	-0.11	-0.02	0.07
Standard error	0.04	0.04	0.04	0.04	0.06	0.05
p **value**	.99	.14	.82	.05	.98	.39

652 *Note.* C = Collocated, D = Dichotic, O = Overlapped, I = Interleaved

653

654 **Figure 3**

655 *Listening performance (Proportion Correct) as a function of individual CogFact scores for*
656 *each of the four conditions of the design. Note that the weakest relationship is seen in the*
657 *Collocated_Overlapped condition.*

658

659
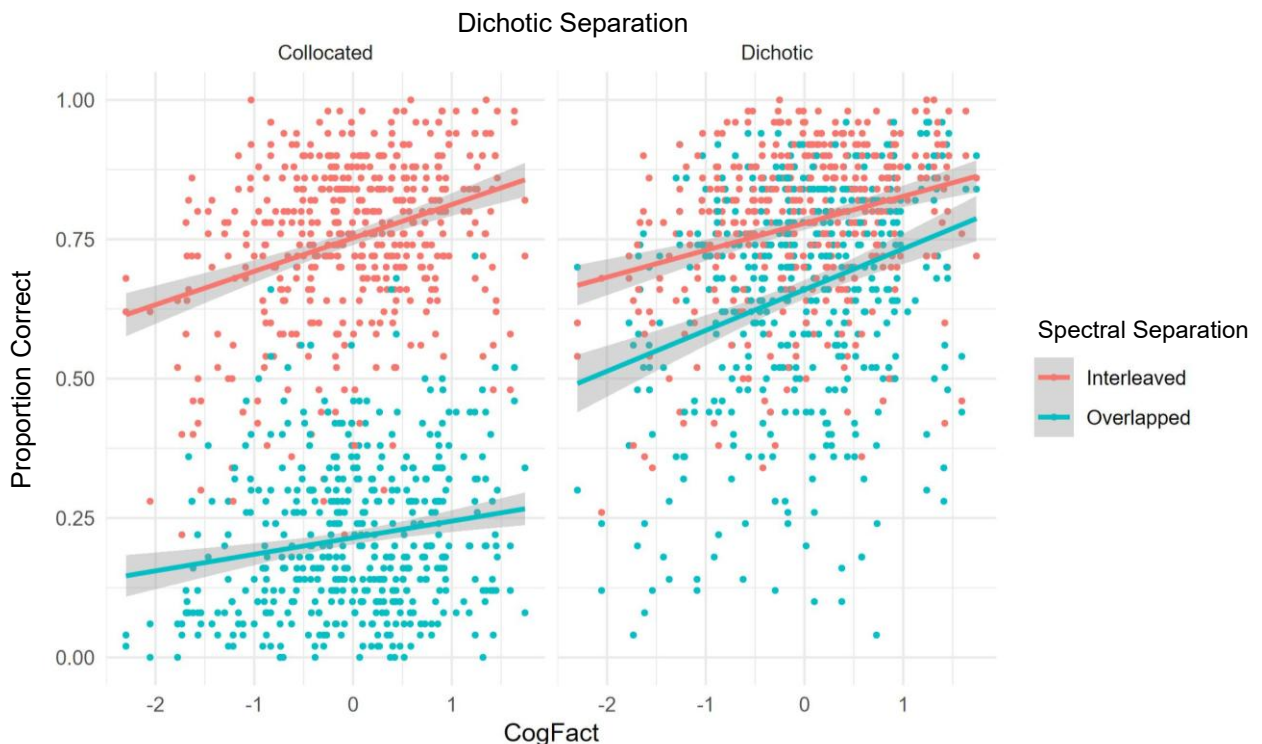
660 **5. Discussion**

661 Understanding speech in the presence of a competing talker can be challenging due to both

662 energetic masking (EM; spectro-temporal overlap between target and masker) and the need to

663 stream, and selectively attend to, the voices. Speech-on-speech listening can be facilitated

664 through separating the two talkers either spatially (via manipulating perceived location or ear

665 of presentation) or spectrally (via filtering the voices into non-overlapping frequency bands).

666 Both types of separation reduce EM as well as providing cues to streaming. The performance

667 benefits that this creates – known as spatial (dichotic) and spectral release from masking

668 (RM) – are well documented; however, certain key questions remain unanswered. In

669 particular, the relative benefit of these two types of RM, the effects of combining them, and

670 the relationship of these benefits to individual differences in cognitive abilities are unclear. In

671 the current study, we addressed these questions using a selective listening task, in which

672 participants heard meaningful sentences spoken by two talkers simultaneously and were cued

673 to report the sentence spoken by one of the talkers. The talkers were presented as either

674 collocated (no dichotic RM) or dichotic (dichotic RM), and in spectrally overlapping (no

675 spectral RM) or spectrally interleaved (spectral RM) frequency bands. We also administered

676 a battery of cognitive tasks and used them to derive a single overall cognitive score for each

677 participant. The results from the specific cognitive tests and the combined cognitive score

678 were then used to assess the relationships between cognition, listening performance, and RM

679 effects.

680 **5.1 Release from masking**

681 Performance on the listening task was generally high (> 60% correct), with the exception of

682 the most challenging listening condition (Collocated_Overlapped) in which neither dichotic

683 nor spectral RM was present, where performance dropped to 21%. As this pattern would

684 suggest, we found main effects of both dichotic and spectral separation: performance was

685 higher in the dichotic than the collocated conditions, and in the spectrally interleaved than the

686 spectrally overlapping conditions. These results support a range of previous studies

687 demonstrating that spatial (dichotic) and spectral separation can improve intelligibility of a

688 target during speech-on-speech listening (e.g., Knight et al., 2023; Kidd et al., 2010;

689 Litovsky, 2012).

690 Performance improvements were similar when only one (of either dichotic or spectral) RM

691 was present: performance improved relative to the no-RM condition

692 (Collocated_Overlapped) by 45% when stimuli were dichotically (but not spectrally)

693 separated and by 54% when stimuli were spectrally (but not dichotically) separated. This

694 suggests that, at least in the context of the dichotic and spectral separation used in this study,

695 listeners are able to exploit both cues to a similar extent – a finding in line with Best et al.

696 (2006). However, we also found a significant interaction between dichotic and spectral

697 separation, with post-hoc tests revealing differences between all four listening conditions.

698 Thus, despite the generally similar magnitude of RM effects, there was nevertheless a

699 significant difference between the conditions with spectral RM only (Collocated_Interleaved)

700 and dichotic RM only (Dichotic_Overlapped), with better performance seen in the spectral

701 RM condition. This seems likely to reflect differences in processing at higher levels of the

702 auditory pathway. Specifically, information from our two ears is integrated in the midbrain

703 (e.g., Palmer et al., 2002), potentially allowing for some cross-ear interference higher in the

704 auditory pathway despite dichotic release from EM at the cochlear level. However,

705 information in different frequency bands is not integrated in this way, with tonotopy (i.e., the

706 activation of distinct neural locations by different frequencies) being maintained through to

707 the level of the auditory cortex (e.g., Humphries et al., 2010). As a result, there is less

708 potential for cross-frequency interference at these higher processing levels. From this

709 perspective, then, it makes sense that a larger performance benefit might arise from spectral,

710 rather than dichotic, RM.

711   This finding should not, however, be taken to indicate that spectral RM always provides the

712   largest perceptual benefit; instead, it seems likely to be a product of the current experiment's

713   somewhat artificial listening situation. In real-world contexts, discrete spectral separation (as

714   implemented here) is rare: spectra typically overlap, with spectral cues to streaming arising

715   from moment-by-moment changes in the degree of overlap.[4] In other words, data availability

716   arising from real-world spectral differences tends to be less extensive and more dynamic than

717   in our experiment. Furthermore, in the context of common real-world spectral differences,

718   such as those related to voice differences (e.g., male vs. female voices), it may be that

719   listeners rely more heavily on spatial than spectral cues (Allen et al., 2008). That is, listeners

720   in real-world situations may weight spectral cues less heavily for a variety of reasons. Finally,

721   any comparisons of the relative benefit of dichotic and spectral RM need to be contextualised

722   within a range of possible performance benefits. In other words, although the overall

723   magnitude of dichotic vs. spectral RM appeared to be similar in the current study, dichotic

724   presentation represents the upper limit of RM (no further separation is possible), whereas

725   spectral RM may afford much larger benefits than observed here if implemented in a different

726   spectro-temporal configuration. Such questions remain to be investigated.

727   Despite the slightly larger performance benefit derived from spectral RM relative to dichotic

728   RM, it was not the case that spectral RM rendered dichotic RM ineffective: we observed a

729   significant additional improvement in performance when both types of RM were present

730   compared to either type of RM alone. The reasons for this are unclear. However, one

731   possibility is that the addition of dichotic RM on top of spectral RM caused a perceptual

732   reweighting of the existing spectral cues in a manner that was beneficial to performance.

733   Regardless of the underlying mechanism, however, it is clear that the highest performance

---

[4] Complete dichotic isolation of sound sources, as implemented here, is also rare; however, it seems likely that listeners are more familiar with the process of attending to different spatial locations (even when signalled entirely by ear of presentation) than with attending to separate spectral regions.

734 occurred when both types of RM were present. This is in line with Best et al. (2013). It is also

735 supported by the work of Ihlefeld and Shinn-Cunningham (2008), who found that both timbre

736 and location cues contributed independently to listeners' ability to accurately report a target

737 sentence during speech-on-speech listening, with performance highest when both were

738 present as opposed to either feature alone.

739 The observed interaction between dichotic and spectral separation reflects, in part, the fact

740 that only a relatively small (albeit significant) additional benefit arises when both types of

741 RM are present compared to either spectral or dichotic RM alone (see Figure 2). One

742 possibility is that spectral and dichotic RM in fact combine additively, and that the observed

743 interaction results from ceiling effects. However, it should be noted that mean accuracy in the

744 easiest condition was only around 75% correct, leaving room for improvement. Nevertheless,

745 future studies should attempt to address this question by targeting lower overall performance

746 levels.

747 Finally, it must be remembered that dichotic listening was used in the current study as a

748 proxy for true spatial hearing. The generalisability of the current findings to situations

749 involving cues such as interaural time differences (ITDs) and free-field presentation of sound

750 sources remains to be established.

751 **5.2 Contribution of cognitive abilities to listening performance**

752 Scores on our combined cognitive factor showed a positive relationship with overall listening

753 performance. This is in line with the prevailing view that good cognition supports successful

754 speech-in-noise listening (e.g., Akeroyd, 2008; Rönnberg et al., 2022). However, there was

755 also a highly significant 3-way interaction between dichotic separation, spectral separation,

756 and cognition. Post-hoc analyses revealed that, although still positive, the weakest

757 relationship between performance and cognition was observed for the most challenging

758   listening condition, i.e., the one in which neither dichotic nor spectral RM was present

759   (Collocated_Overlapped). Although overall performance was poorest in this condition,

760   variability was comparable to that in the other conditions – in other words, the weaker

761   relationship seems unlikely to be due to a floor effect. Furthermore, this correlation was

762   weaker than the correlation between cognition and listening performance in either of the

763   conditions featuring only one type of RM. Although these post-hoc comparisons did not

764   reach significance after correction for multiple tests, their broad pattern supports an account

765   of speech-on-speech listening in which the relative balance between data and resource limits

766   shifts according to the nature of the stimuli. As described above, when EM is strong and/or

767   cues to streaming are absent, a data limit is reached, and there is no longer enough of the

768   signal available for the application of cognitive resources to have a strong impact on

769   performance (c.f. Norman & Bobrow, 1975). This interpretation is in line with Mattys et al.'s

770   (2025) DRL framework and adds nuance to models such as the Ease of Language

771   Understanding model (ELU, Rönnberg et al., 2022b), which suggest that cognitive resources

772   are always positively linked to performance during listening in challenging conditions,

773   regardless of the nature or extent of the challenge. A data- and resource-limit account would

774   impose a boundary on the extent of this positive relationship.

775   The main effects of Phon_RJ, Att_RSVP and Exec_LNS in our first GLMM showed, that, of

776   our six cognitive tasks, one measure of phonological processing (Rhyme Judgement), one

777   measure of attention (RSVP), and one measure of executive function (LNS) were associated

778   with performance across all conditions of the listening task. This finding is in line with

779   existing research highlighting the importance of phonological skills, executive function and

780   domain-general selective attention when listening to speech in noisy environments (Clayton

781   et al., 2016; Millman & Mattys, 2017; Smith & Pichora-Fuller, 2015). However, it must be

782   noted that the internal consistency of both the Att_RSVP and Exec_LNS measures was low.

783   On the one hand, low consistency may lead to an underestimation of the true relationship

784   between variables; on the other hand, it may mean that any identified relationships are

785   potentially unreliable. By contrast, the Phon_RJ measure had excellent internal consistency.

786   As a result, only a relationship between phonological processing and listening performance

787   can be asserted with confidence.

788   Interestingly, the other measure of executive function – the Reading Span Task (RSpan) –

789   was not associated with listening performance, despite its acceptable internal consistency.

790   This is surprising given that span tasks of this nature are often considered a "gold standard"

791   measure of cognition in the speech-in-noise literature, and are thought to capture a range of

792   complex working memory skills relevant for successful speech perception in challenging

793   environments (Rönnberg et al., 2022b). Although a relationship is not always observed

794   between speech-in-noise listening and RSpan scores, particularly in young normal-hearing

795   adults, this is suggested to be because of a lack of a detectable relationship between speech

796   perception and cognition in general for this group (Fullgräbe & Rosen, 2016), and not – as is

797   the case here – because other cognitive measures provide greater predictive power. One

798   factor to consider is that Reading Span Tasks, and indeed span tasks in general, exist in

799   diverse formats and can be implemented in a variety of ways. Although the underlying

800   mechanism is not yet well understood, different implementations of the RSpan not only

801   produce different patterns of performance but also appear to have different relationships to

802   speech-in-noise perception (Heinrich, 2025). The absence of a relationship between the

803   RSpan and listening performance in the current study may therefore reflect the specific

804   details of our implementation in addition to the role of complex working memory and

805   executive function in speech-on-speech listening in general.

806   BIC values indicated that the GLMM fitted with the combined cognitive factor (CogFact)

807   was a slightly better fit to the data than the GLMM that used all six cognitive tasks

36

808 separately. This suggests that combining the cognitive task scores into a single factor

809 captured a meaningful underlying general ability that could be related to variation in

810 performance on the listening task. Nevertheless, it is worth noting that the two tests of

811 attention (Singleton and Rapid Serial Visual Presentation tasks) did not correlate with any of

812 the other cognitive measures and produced low loadings on our combined cognitive factor. It

813 is possible that, although our tasks were developed from existing paradigms (e.g., Adamo et

814 al., 2008; Theeuwes, 1992), task-specific demands rendered the data too noisy to reliably

815 detect individual differences in attentional control. The fact that we observed a significant

816 relationship between individual RSVP scores and listening task performance may at first

817 suggest that this explanation is unlikely. Instead, it may be tempting to conclude that attention

818 is separable from other cognitive domains such as executive function and phonological

819 processing. However, given the substantial body of work linking attention to other facets of

820 cognition, this also seems unlikely (e.g., Baddeley, 2000; Diamond, 2013). Furthermore, the

821 poor internal consistency for the RSVP task makes any relationships between it and other

822 measures questionable. An alternative possibility is that shared variability in the attention

823 tasks and listening tasks may have been introduced by the online nature of the experiment

824 and the resulting difficulty in controlling for potentially important hardware-related factors

825 such as screen size, visual angle and contrast (Chen et al., 2011; Yazgan et al, 2023) or

826 participant-related factors such as attention and focus during time-sensitive tasks. Future

827 studies could investigate the degree of variability in online task performance due to these

828 factors and the possibility of measuring and/or controlling them remotely.

829 Finally, we observed a high correlation between dichotic and spectral RM effects. In other

830 words, participants who accrued a large performance benefit from one type of RM tended to

831 also benefit from the other type. This suggests that broadly similar underlying skills are

832 required to exploit the additional information made available through release from masking.

833   This is despite the fact that the cues involved are of a very different nature, with dichotic

834   release involving manipulation of the locus of sound presentation and spectral release

835   involving manipulation of the sound's physical properties. Furthermore, as mentioned above,

836   there is a considerable range of individual differences in cue use and strategy during speech-

837   in-noise listening. With these factors in mind, it is plausible that the ability to exploit dichotic

838   vs. spectral release from masking might vary within an individual. Nevertheless, the results

839   reported here are in line with existing work indicating that an individual's speech-on-speech

840   listening ability is correlated not just across different types of masking but for similar

841   masking tasks across different modalities (Byrne et al., 2023).

842   **6. Conclusion**

843   In this study, we explored the effects of spatial (specifically dichotic) and spectral release

844   from masking (RM) during speech-on-speech listening using a selective attention paradigm

845   in which participants had to attend to a target voice which was separated from the masker

846   voice either dichotically, spectrally, both, or not at all. We observed significant performance

847   benefits for both types of RM, with the best performance shown when both were present

848   simultaneously. We also administered a battery of cognitive tasks in order to assess the

849   relationship of these RM benefits to individual differences in cognitive abilities. The results

850   revealed that the relationship between cognitive ability and listening performance was

851   generally positive, but weakest for the most challenging listening condition (i.e., the

852   condition in which no RM was present). Short-term phonological memory, executive

853   function and attention seemed to be the strongest predictors of absolute listening

854   performance. We additionally observed a strong correlation between dichotic and spectral

855   RM benefits, suggesting that broadly similar underlying skills are required to exploit RM

856   cues. We interpret these results in terms of a data- and resource-limit framework, which we

857   suggest may be a useful tool to account for speech perception performance across a range of

858   different tasks and listening environments. Future use of the framework may allow us to add

859   nuance to existing theories of the relationship between speech perception and cognitive

860   abilities.

861   **References**

862   Adamo, M., Pun, C., Pratt, J., & Ferber, S. (2008). Your divided attention, please! The

863   maintenance of multiple attentional control sets over distinct regions in space. *Cognition,*

864   *107*(1), 295-303.

865   Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual

866   differences in cognitive ability? A survey of twenty experimental studies with normal and

867   hearing-impaired adults. *International Journal of Audiology, 47*(sup2), S53-S71.

868   Allen, K., Carlile, S., & Alais, D. (2008). Contributions of talker characteristics and spatial

869   location to auditory streaming. *The Journal of the Acoustical Society of America, 123*(3), 1562-

870   1570.

871   Allen, K., Alais, D., Shinn-Cunningham, B., & Carlile, S. (2011). Masker location uncertainty

872   reveals evidence for suppression of maskers in two-talker contexts. *The Journal of the*

873   *Acoustical Society of America, 130*(4), 2043-2053.

874   Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). A dynamic auditory-

875   cognitive system supports speech-in-noise perception in older adults. *Hearing Research, 300*,

876   18-32.

877   Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla

878   in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1),

879   388–407.

880   Apoux, F., & Healy, E. W. (2010). Relative contribution of off-and on-frequency spectral

881   components of background noise to the masking of unprocessed and vocoded speech. *The*

882   *Journal of the Acoustical Society of America, 128*(4), 2075-2084.

883   Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in*

884   *Cognitive Sciences, 4*(11), 417–423.

885   Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H.,

886   Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior*

887   *Research Methods, 39*(3), 445-459. DOI: 10.3758/BF03193014.

888   Best, V., Thompson, E. R., Mason, C. R., & Kidd, G. (2013). Spatial release from masking as

889   a function of the spectral overlap of competing talkers. *The Journal of the Acoustical Society*

890   *of America, 133*(6), 3677-3680.

891   Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two

892   simultaneous talkers. *The Journal of the Acoustical Society of America, 109*(3), 1101–1109.

893   Byrne, A. J., Conroy, C., & Kidd, G. (2023). Individual differences in speech-on-speech

894   masking are correlated with cognitive and visual task performance. *The Journal of the*

895   *Acoustical Society of America*, *154*(4), 2137-2153.

896   Calcus, A., Agus, T., Kolinsky, R., Colin, C., & Deltenre, P. (2015). Isolating informational

897   masking in both pure and complex tone sequences. *Ear and Hearing, 36*(3), 330-337.

898   Chen, I. P., Liao, C. N., & Yeh, S. H. (2011). Effect of display size on visual

899   attention. *Perceptual and Motor Skills*, *112*(3), 959-974.

900   Clayton, K. K., Swaminathan, J., Yazdanbakhsh, A., Zuk, J., Patel, A. D., & Kidd Jr, G. (2016).

901   Executive function, visual attention and the cocktail party problem in musicians and non-

902   musicians. *PLOS One*, *11*(7), e0157638.

903   Coffey, E. B., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in

904   musicians: A review. *Hearing Research*, *352*, 49-69.

905   Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited:

906   The importance of working memory capacity. *Psychonomic Bulletin & Review, 8*(2), 331–335.

907   Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, *31*(3-4), 579-584.

908   Culling, J.F., & Stone, M.A. (2017). Energetic masking and masking release. In Middlebrooks,

909   J., Simon, J., Popper, A., Fay, R. (Eds.) *The Auditory System at the Cocktail Party*. Springer

910   Handbook of Auditory Research, vol 60. Springer, Cham.

911   Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

912   Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and

913   reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450-466.

914   Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and

915   vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the*

916   *Acoustical Society of America*, *114*(5), 2913-2922.

917   Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various

918   types of masker noises. *Ear and Hearing*, *34*(3), 261-272.

919   Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135-168.

920   DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores:

921   Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation,*

922   *14*(1): 20. DOI: 10.7275/da8t-4g52.

923   Dryden, A. (2018). Multiple perspectives on the association between cognition and speech-in-

924   noise perception performance. *Unpublished PhD thesis*.

925 Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The association between

926 cognitive performance and speech-in-noise perception for adult listeners: A systematic

927 literature review and meta-analysis. *Trends in Hearing*, *21*, 2331216517744675.

928 Edmonds, B. A., & Culling, J. F. (2006). The spatial unmasking of speech: Evidence for better-

929 ear listening. *The Journal of the Acoustical Society of America, 120*(3), 1539–1545.

930 Engle, R. W., & Kane, M. J. (2004). Executive Attention, Working Memory Capacity, and a

931 Two-Factor Theory of Cognitive Control. In B. H. Ross (Ed.), *The Psychology of Learning and*

932 *Motivation: Advances in Research and Theory* (Vol. 44, pp. 145–199). Elsevier Science.

933 Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived

934 spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of*

935 *America, 106*(6), 3578–3588.

936 Füllgrabe, C., & Rosen, S. (2016). On the (un)importance of working memory in speech-in-

937 noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology, 7*, 1268.

938 https://doi.org/10.3389/fpsyg.2016.01268

939 Gatehouse, S., Naylor, G., & Elberling, C. (2003). Benefits from hearing aids in relation to the

940 interaction between the user and the environment. *International Journal of Audiology,*

941 *42*(sup1), 77-85.

942 Gifford, R. H., Dorman, M. F., Brown, C., & Spahr, A. J. (2012). Hearing, psychophysics, and

943 cochlear implantation: experiences of older individuals with mild sloping to profound sensory

944 hearing loss. *Journal of Hearing Science, 2*(4), 9-17.

945 Heinrich, A. (2021). The role of cognition for speech-in-noise perception: Considering

946 individual listening strategies related to aging and hearing loss. *International Journal of*

947 *Behavioral Development*, *45*(5), 382-388.

948   Heinrich, A. (2025). The predictive ability of various aspects of memory to speech in noise

949   perception. *Speech in Noise Workshop 2025* (Lancaster, UK).

950   Heinrich, A., & Knight, S. (2016). The contribution of auditory and cognitive factors to

951   intelligibility of words and sentences in noise. In P. van Dijk et al. (Eds.), *Physiology,*

952   *Psychoacoustics and Cognition in Normal and Impaired Hearing*, Advances in Experimental

953   Medicine and Biology 894.

954   Heinrich, A., Ferguson, M. A., & Mattys, S. L. (2020). Effects of cognitive load on pure-tone

955   audiometry thresholds in younger and older adults. *Ear and Hearing, 41*(4), 907-917.

956   Holmes, E., Zeidman, P., Friston, K. J., & Griffiths, T. D. (2021). Difficulties with speech-in-

957   noise perception related to fundamental grouping processes in auditory cortex. *Cerebral*

958   *Cortex*, *31*(3), 1582-1596.

959   Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit

960   provided by amplified speech to older adults. *Journal of the American Academy of Audiology,*

961   *18*(07), 590–603.

962   Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human

963   auditory cortex. *Neuroimage*, *50*(3), 1202-1211.

964   Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational

965   masking in a selective speech identification task. *The Journal of the Acoustical Society of*

966   *America, 123*(6), 4369–4379.

967   Janse, E., & Andringa, S. J. (2021). The roles of cognitive abilities and hearing acuity in older

968   adults' recognition of words taken from fast and spectrally reduced speech. *Applied*

969   *Psycholinguistics, 42*(3), 763–790.

970 Johns, M. A., Calloway, R. C., Phillips, I., Karuzis, V. P., Dutta, K., Smith, E., ... & Kuchinsky,

971 S. E. (2023). Performance on stochastic figure-ground perception varies with individual

972 differences in speech-in-noise recognition and working memory capacity. *The Journal of the*

973 *Acoustical Society of America*, *153*(1), 286-303.

974 Johnston, R. S., & McDermott, E. A. (1986). Suppression effects in rhyme judgement

975 tasks. *The Quarterly Journal of Experimental Psychology Section A*, *38*(1), 111-124.

976 Kidd, G., Mason, C. R., & Gallun, F. J. (2005). Combining energetic and informational

977 masking for speech identification. *The Journal of the Acoustical Society of America*, *118*(2),

978 982-992.

979 Kidd, G., Mason, C. R., Best, V., & Marrone, N. (2010). Stimulus factors influencing spatial

980 release from speech-on-speech masking. *The Journal of the Acoustical Society of*

981 *America*, *128*(4), 1965-1978.

982 Knight, S., Rakusen, L., & Mattys, S. (2023). Conceptualising acoustic and cognitive

983 contributions to divided-attention listening within a data-limit versus resource-limit

984 framework. *Journal of Memory and Language*, *131*, 104427.

985 Knight, S., & Heinrich, A. (2017). Different measures of auditory and visual stroop interference

986 and their relationship to speech intelligibility in noise. *Frontiers in Psychology*, *8*, 230.

987 Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-

988 accuracy trade-offs (?). *Behavior Research Methods*, *51*, 40-60.

989 Litovsky, R. Y. (2012). Spatial release from masking. *Acoustics Today, 8*(2), 18–25.

990 Mattys, S. L., O'Leary, R. M., McGarrigle, R. A., & Wingfield, A. (2025). Reconceptualizing

991 cognitive listening. *Trends in Cognitive Sciences*.

992 Middlebrooks, J. C. & Onsan, Z. A. (2012). Stream segregation with high spatial acuity.

993 *Journal of the Acoustical Society of America, 132*(6), 3896-3911.

994 Middlebrooks, J. C. & Green, D. M. (1991). Sound localization by human listeners. *Annual*

995 *Review of Psychology, 42*, 135-159.

996 Millman, R. E., & Mattys, S. L. (2017). Auditory verbal working memory as a predictor of

997 speech perception in modulated maskers in listeners with normal hearing. *Journal of Speech,*

998 *Language, and Hearing Research*, *60*(5), 1236-1245.

999 Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes.

1000 *Cognitive Psychology, 7*(1), 44–64.

1001 O'Leary, R. M., Neukam, J., Hansen, T. A., Kinney, A. J., Capach, N., Svirsky, M. A., &

1002 Wingfield, A. (2023). Strategic pauses relieve listeners from the effort of listening to fast

1003 speech: data limited and resource limited processes in narrative recall by adult users of cochlear

1004 implants. *Trends in Hearing*, *27*, 23312165231203514.

1005 O'Neill, E. R., Kreft, H. A., & Oxenham, A. J. (2019). Speech perception with spectrally non-

1006 overlapping maskers as measure of spectral resolution in cochlear implant users. *Journal of the*

1007 *Association for Research in Otolaryngology*, *20*(2), 151-167.

1008 Palmer, A. R., Shackleton, T. M., & McAlpine, D. (2002). Neural mechanisms of binaural

1009 hearing. *Acoustical Science and Technology*, *23*(2), 61-68.

1010 Rönnberg, J., Signoret, C., Andin, J., & Holmer, E. (2022). The Cognitive Hearing Science

1011 perspective on perceiving, understanding, and remembering language: The ELU model.

1012 *Frontiers in Psychology, 13*.

1013 Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. *IEEE*

1014 *Transactions on Audio and Electroacoustics, 17*, 225–246.

1015 Schneider, B. A., Avivi-Reich, M., & Daneman, M. (2016). How spoken language

1016 comprehension is achieved by older listeners in difficult listening situations. *Experimental*

1017 *Aging Research*, *42*(1), 31-49.

1018 Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in*

1019 *Cognitive Sciences, 12*(5), 182–186.

1020 Smith, S. L., & Pichora-Fuller, M. K. (2015). Associations between speech understanding and

1021 auditory and visual tests of verbal working memory: effects of linguistic complexity, task, age,

1022 and hearing loss. *Frontiers in Psychology*, *6*, 1394.

1023 Stubbs, R. J., & Summerfield, Q. (1990). Algorithms for separating the speech of interfering

1024 talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired

1025 listeners. *The Journal of the Acoustical Society of America*, *87*(1), 359-372.

1026 Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception &*

1027 *Psychophysics*, *51*(6), 599-606.

1028 Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control,

1029 secondary memory, or both? A direct test of the dual-component model. *Journal of Memory*

1030 *and Language, 62*(4), 392–406.

1031 Wechsler, D. (1997). *Wechsler Adult Intelligence Scale III*. New York: The Psychological

1032 Corporation.

1033 Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to

1034 facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79* (7),

1035 2064–2072.

1036 Yazgan, M. T., Yağımlı, M., & Ozubko, J. (2024). Evaluating and modeling the effects of

1037 brightness on visual attention using multiple object tracking method. *Journal of Information*

1038 *Display*, *25*(3), 271-293.