10-15-2020 2:00 PM

Cognitive Resources Are Recruited Consistently Across People During Story Listening

Matthew T. Bain, *The University of Western Ontario*

Supervisor: Dr. Ingrid S. Johnsrude, *The University of Western Ontario*
Co-Supervisor: Dr. Björn Herrmann, *Rotman Research Institute*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Neuroscience

Abstract

Degraded speech encoding as a result of hearing loss increases cognitive load and makes listening effortful. Standard hearing assessment does not capture this cognitive impact of hearing impairment. Speech in noise testing measures intelligibility for isolated sentences that are typically not engaging and lack meaningful context. These materials may not capture the processes involved in everyday listening situations, in which people are often intrinsically motivated to comprehend the speech they are hearing. The current study explored a novel approach using natural, spoken stories. We first characterized time courses of executive load during story listening in young individuals with normal hearing using a reaction time (RT) task. We then computed correlations between executive load time courses (operationalized as reaction times) to quantify their reliability. Reaction-time time courses were significantly correlated across participants, suggesting consistent cognitive recruitment across individuals. Synchronization of RTs across participants was related to ratings of story enjoyment, but not absorption, suggesting that enjoyment, one key facet of engagement, predicts the degree to which a story's cognitive demands are experienced similarly by listeners. Correlated executive load time courses among healthy individuals may be sensitive to abnormal mental states: divergence from the canonical time courses characterized here could serve as a sensitive tool for characterizing listening effort.

Keywords

hearing loss, listening effort, cognitive load, dual task, intersubject correlation, cognitive control

Summary for Lay Audience

Many older people experience difficulty understanding speech in minimal background noise, and often report listening to be effortful. Increases in listening effort are associated with declines in quality of life and mental health, but clinical tests of speech perception are not sensitive to the effort patients report. Speech perception testing relies on standardized sentences that lack meaningful context. These tests may fail to capture the key cognitive processes that support listening in everyday environments, in which people are often motivated to comprehend the speech they listen to. In this study we explore a novel method of assessing listening using engaging spoken stories. Our findings suggest that the dynamics of cognitive processing during listening are consistent across individuals, and that the consistency of cognitive processing is related to story enjoyment. High consistency in cognitive processing among young individuals with normal hearing provides sensitivity to abnormal mental states, potentially enabling detection of people who find listening in background noise to be unusually effortful. Building on existing tests of speech perception, this research opens the door to methods of assessing listening effort that better capture the processes underlying listening in everyday environments.

Acknowledgments

I would like to thank everyone who assisted and supported me throughout this research project. I express my deepest gratitude to Dr. Ingrid Johnsrude and Dr. Björn Herrmann, whose supervision has been essential and has taught me so much. Thank you for your patience and unwavering support. I would like to thank my wonderful advisory committee members and lab members, whose guidance throughout this project has been invaluable. Many friends and family members have provided advice and encouragement along the way, and to each of you I am extremely grateful. This project has been a challenging and rewarding experience, and I could not have completed it without you all.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

Hearing loss affects more than 4 in 10 people over the age of 50 (Feder, 2015) and is often diagnosed long after hearing-related difficulties, such as understanding speech in noisy environments, are first experienced (Pichora-Fuller & Souza, 2003). For young people with normal hearing, difficulty understanding speech is typically restricted to very noisy environments. In contrast, 15-40% of people over 50 experience difficulty understanding speech even with minimal background noise (Feder, 2015; Helfer et al., 2017), and often report listening in such environments to be effortful and tiring (Gatehouse & Noble, 2004; Goderie et al., 2020; Pichora-Fuller et al., 2016). Degraded speech encoding as a result of hearing impairment or background noise increases cognitive load – the degree to which cognitive capacities such as working memory and knowledge-guided perception are taxed to compensate for an impoverished speech signal (Johnsrude & Rodd, 2016) – making listening effortful (Alhanbali et al., 2018; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). As background noise is ubiquitous in everyday social environments, listening effort poses major challenges to communication that can lead to declines in quality of life (Nachtegaal et al., 2009), social isolation (Ramage-Morin, 2016), and negative health outcomes (Nachtegaal et al., 2009), possibly including cognitive decline (Lin et al., 2011; Wayne & Johnsrude, 2015). Despite these consequences, existing measures and treatments of hearing impairment are not sensitive to listening effort (Löhler et al., 2019; Ruggles et al., 2011; Ruggles et al., 2012; Tremblay et al., 2015). A measure of hearing impairment that is sensitive to real-world listening challenges and can be used to inform optimal hearing aid fitting practices is critically needed.

1.1 Hearing loss

Hearing loss arises through a variety of pathologies in the ear and brain

Hearing impairment is associated with a variety of peripheral and neural pathologies that affect perception in different ways (Gratton & Vázquez, 2003; Plack et

al., 2014). At the periphery, dysfunction of the inner and outer hair cells (IHCs and OHCs) is a common cause of hearing loss (Moore, 2007). Outer hair cells amplify quiet sounds of a given frequency (corresponding to the preferred frequency of the basilar membrane at the location they are attached to) by modulating basilar membrane movement. Dysfunction of OHCs is either caused directly, by damage, or indirectly, by changes to the metabolism of the stria vascularis, which supplies their energy (Gratton & Vázquez, 2003; Moore, 2007). At the perceptual level, hair cell dysfunction results in a reduction in sensitivity, or an elevation in the sound pressure level required for sounds of a given frequency to be audible (i.e., threshold elevation). As OHCs are most concentrated at the base of the basilar membrane (sensitive to high frequencies), this loss of sensitivity affects high frequency hearing most dramatically. In addition to amplifying sounds of a given frequency, OHCs attenuate sounds of neighbouring frequencies, sharpening the frequency tuning of the basilar membrane (Moore, 2007). Loss of OHCs therefore is also associated with a reduction in frequency selectivity (Patterson et al., 1982). This is particularly detrimental for understanding speech in noisy environments, where the frequencies most important for speech perception are easily obscured (Lesica, 2018). In addition, loss of OHCs is associated with increased loudness recruitment, or more rapid growth in perceived loudness with sound level than is normal (Moore, 2007).

Occuring at the interface between peripheral and neural structures, cochlear synaptopathy is a form of hearing loss associated with degeneration of the synapses between inner hair cells (IHCs) and auditory nerve (AN) terminals (Kujawa & Liberman, 2015; Liberman et al., 2016). This impairment preferentially affects high threshold, low spontaneous rate AN fibers, which, although not essential for encoding sound in quiet environments, play a critical role in encoding speech in the presence of background noise (Lesica, 2018; Liberman et al., 2016). Beyond the periphery, hearing loss is associated with degeneration of spiral ganglion neurons (Bao & Ohlemiller, 2010) and changes in function of auditory circuits, including a loss of inhibitory tone (Salvi et al., 2017) and hyperresponsiveness to sound (Herrmann et al., 2018). Impairments in encoding the fine temporal structure of sounds (i.e., temporal fine structure) are common (Lorenzi et al., 2006), with significant detriment to speech perception (Smith et al., 2002), although the specific physiological mechanisms of this impairment are unknown (Parthasarathy et al.,

2020). As hearing impairment affects hearing sensitivity, frequency selectivity, the encoding of speech in the presence of background noise, and the encoding of speech temporal fine structure, an informative assessment of hearing impairment should evaluate each of these aspects of hearing.

How hearing loss is diagnosed and treated

Traditional methods of assessing hearing, including pure tone audiometry and speech in noise testing, are not sensitive to the experience of effort that people often have while trying to understand speech in the moderately noisy environments of everyday life (Pichora-Fuller et al., 2016). Pure tone audiometry, which measures hearing thresholds for pure tone frequencies played in quiet (Beck et al., 2018.; Walker, 2013), captures deficits in hearing sensitivity, but is not a good predictor of speech perception (Tremblay et al., 2015). Speech in noise testing – the standard method of assessing speech perception – measures intelligibility for standardized sentences played in different levels of background noise (Wilson et al., 2007), but also fails to account for real world communication difficulties (Ruggles et al., 2011). An individual who has clinically normal pure tone thresholds and speech in noise performance may be sent home from an audiological clinic with a clean bill of hearing health, despite reporting major hearing and communication difficulties in their everyday life (Lesica, 2018; Parthasarathy et al., 2020).

Hearing aids, the primary line of treatment for hearing impairment, are fitted according to a combination of audiometry and speech in noise perception, with a focus on frequencies that are important for the perception of speech (namely .5, 1, 2, and 4 kHz; Peelle & Wingfield, 2016). The main function of hearing aids is to amplify sounds of particular frequencies for which an individual has most pronounced loss of sensitivity (Lesica, 2018). Additional signal processing is sometimes employed in these devices, such as noise reduction algorithms to improve speech perception. Although amplification and noise reduction restore hearing sensitivity, they do not effectively compensate for loss of frequency selectivity or speech in noise challenges. This might explain why of the >40% of people over the age of 50 who have significant hearing loss (Feder, 2015), only

10-20% actually wear a hearing aid (Lopez-Poveda et al., 2017; Öberg et al., 2012). More research into the physiological underpinnings and cognitive neuroscience of listening effort is needed to better inform hearing aid fitting and ensure that patient needs are met.

1.2 Listening effort

Conflicting definitions of listening effort

Different studies and models often use different definitions of the term "listening effort" (Lemke & Besser, 2016). Widespread disagreement about the processes and mechanisms underlying listening effort makes it difficult to connect results across studies and may limit progress on measures and therapies. Some authors refer to listening effort as something that is deliberately *exerted* - through the allocation of cognitive resources or energy - in service of a listening task (McGarrigle et al., 2014; Pichora-Fuller et al., 2016). Others refer to listening effort as a subjective *experience* – a perceptual consequence of a challenging listening task (Herrmann & Johnsrude, 2020; Johnsrude & Rodd, 2016; Krueger et al., 2017; Lemke & Besser, 2016). As explored in more detail in section 1.6 below, in this study, we conceptualize listening effort as a subjective experience, resulting from an interaction between the demands associated with a listening task and the cognitive resources an individual possesses that can contribute to mitigating these demands, 'filling in the gaps' in an impoverished speech signal, and achieving intelligibility (Johnsrude & Rodd, 2016). The processing demands of a given listening task might include signal degradation, background noise, linguistic complexity, and other demands related to the acoustic and linguistic properties of the speech; the cognitive resources involved might include selective attention, working memory, knowledge, and any other cognitive proesses that can be recruited to aid speech perception. Hearing impairment, like many speech processing demands, reduces the fidelity of the speech signals that are sent to the brain, increasing the load on cognitive processes such as selective attention and perceptual closure (Johnsrude & Rodd, 2016). When one's cognitive resources are only just sufficient for coping with the demands of a listening task, listening effort is experienced (Herrmann & Johnsrude, 2020; Johnsrude & Rodd, 2016).

Varied approaches to measuring listening effort

Many behavioural and physiological measures of listening effort have been developed. Self-report measures are the most straightforward approach to ascertaining real-world communication difficulties. However, as the same self-reported rating of effort may represent different levels of effort for different raters, self-report measures are hard to standardize across listeners and can therefore be unreliable (Johnsrude & Rodd, 2016). Physiological measures include pupil dilation (Koelewijn et al., 2012, 2015; Zekveld et al., 2014) and galvanic skin response (Mackersie et al., 2015), where increased pupil diameter and increased skin conductance, respectively, are considered markers of listening effort. Although these measures correlate strongly with listening effort, they index general physiological arousal, and are susceptible to misinterpretation (Johnsrude & Rodd, 2016). Several electroencephalographic (EEG) measures of listening effort have also been identified, such as wavelet phase synchronization stability (WPSS) of the late auditory response (Bernarding et al., 2013, 2017) and parietal alpha power (Marsella et al., 2017).

Most behavioural measures of listening effort other than subjective ratings employ a dual-task protocol (Gagné et al., 2017; Wu et al., 2016). Dual tasks involve a primary and secondary task. For a dual task assessing listening effort, the primary task is usually to listen to and understand the speech materials used in the experiment. The key assumption of dual-task protocols is that cognitive resources are finite, so if the primary and secondary task are presented concurrently, then as the demands of the primary task increase, the cognitive resources available for performing the secondary task are depleted, and performance on the secondary task is hindered (Gagné et al., 2017). Under this assumption, researchers using a dual task operationalize listening effort as the amount to which performance is hindered on a secondary task when it is presented concurrently with the primary task, relative to when it is presented alone (Gagné et al., 2017). The degree to which performance is hindered is referred to as the ‘dual-task cost’. Common secondary tasks include pressing a button in response to a visual probe or recalling a sequence of digits (Gagné et al., 2017). The dual-task cost in these cases is the increase in response time (RT) to respond to the probe or the decrease in digit recall

performance when the secondary task is presented concurrently with the listening task. Dual tasks provide an indirect index of listening effort and depend, to some extent, on the type of secondary task used, as well as the modality in which the secondary task is presented (Johnsrude & Rodd, 2016). There is no standard method of assessing listening effort using a dual-task protocol. As such, different studies tend to use different secondary tasks and a different operationalization of listening effort (Gagné et al., 2017).

1.3 Listening effort is not a unitary construct

Different listening challenges recruit different brain networks

No single cognitive process fully accommodates the challenges encountered in everyday listening situations (Johnsrude & Rodd, 2016). Different speech materials and listening challenges impose loads on different cognitive systems, associated with different brain networks (Peelle, 2018; Peelle & Wingfield, 2016). However, listening effort is often conceptualized as a unitary construct. This may have contributed to the proliferation of one-dimensional measures, including subjective (self-report), behavioural, and physiological measures (Pichora-Fuller et al., 2016). Behavioural measures and physiological measures like pupil diameter provide a gross index of a complex and multidimensional construct – a single readout of the magnitude of challenges experienced while listening. In contrast, neural measures have the potential to reveal the spectrum of brain networks and activation patterns associated with a given listening task (Johnsrude & Rodd, 2016). Such measures provide the most informative and complete picture of the physiological processes underlying listening effort.

Activity in domain-general brain networks compensates for demands during speech perception

Some evidence indicates that a network involving lateral frontal cortex that is activated by a wide variety of cognitive tasks – the so called ‘multiple demand’ (MD) network – is also active during effortful listening. The MD network is involved in cognitive control (i.e., goal-directed processing; Cole et al., 2013; Duncan, 2010; Gratton et al., 2018; Paxton et al., 2008), and encompasses a range of prefrontal and parietal

regions that show elevated activity in response to acoustically degraded speech. Thus, the MD network may serve as a compensatory hub during effortful listening, mobilizing cognitive resources as required to mitigate the specific processing demands of a listening task (Peelle, 2018). The MD network can be divided into two subsystems, both involved in attention-based task monitoring: the frontoparietal network (FPN), consisting of regions along bilateral dorsolateral prefrontal cortex and intraparietal sulcus, and the cinguloopercular network (CON), consisting of dorsal anterior cingulate cortex and bilateral anterior insula/frontal operculum (Gratton et al., 2018; Peelle, 2018). These networks are associated with similar processes – task switching, error detection, response selection (Cole et al., 2013) – but may subserve effortful listening under different conditions (Peelle, 2018; Sridharan et al., 2008). In addition to the FPN, left premotor cortex appears to be involved in effortful listening when speech is degraded but still highly intelligible (Peelle, 2018; Peelle & Wingfield, 2016). Its role may have to do with meeting increased verbal working memory demands. In contrast, the CON is associated with effortful listening when speech intelligibility begins to suffer (Peelle & Wingfield, 2016). The CON's role in speech understanding is clearly demonstrated by past research in which increased CON activity, measured using functional magnetic resonance imaging (fMRI) was found to predict accuracy on a subsequent word recognition in noise task (Vaden et al., 2013; Vaden et al., 2016). A thorough characterization of the spectrum of neural responses associated with effortful listening would be invaluable for evaluating the specific hearing difficulties associated with different speech processing demands and cognitive abilities.

1.4 Limitations of existing measures of hearing and listening effort

Cognitive control is modulated by motivation

Importantly, cognitive control is modulated by motivational state (Yee & Braver, 2018). The extent to which the compensatory processes associated with efficient speech understanding are engaged depends on whether or not a person is actively attending to the speech (Johnsrude & Rodd, 2016). In one study, participants either attended to spoken

sentences or a distractor and brain responses were measured using fMRI (Wild et al., 2012). The researchers found that frontal brain regions were not active when participants did not attend to the target speech, but showed elevated responses to degraded (noise vocoded) speech relative to clear speech when participants did attend to the target speech. In contrast, regions along the superior temporal sulcus showed reduced responses to degraded speech under no attention, but elevated responses under attention (Wild et al., 2012). A recognition test administered after the experiment revealed that clear speech was processed whether or not it was attended, but degraded speech recognition was improved under directed attention. This pattern of behavioural results and neural activity suggests that attention is required for degraded speech to be processed in certain regions of the brain's language network. Frontal brain regions might enhance lower level processing of speech (in temporal regions) when the speech is attended and degraded but still potentially intelligible (Wild et al., 2012). Feedback projections throughout the entire auditory pathway, including between frontal regions and speech-sensitive cortex (Davis & Johnsrude, 2007; de la Mothe et al., 2012), provide a possible pathway through which this modulation occurs (Wild et al., 2012a; Wild et al., 2012b). The modulatory effect of attention on processing in speech networks reinforces the importance of attention during effortful listening – when listeners attend to degraded speech, their neural responses to it change. Furthermore, it is possible that the subjective experience of effort, as well as the brain networks involved in mitigating it, may differ when this attention is extrinsically motivated (as is the case here) and intrinsically motivated (which is more common in natural listening settings).

Motivation is an important factor in models of listening effort

Existing models of listening effort generally identify motivation as a factor in modulating the relationship between listening and listening effort. The Framework for Understanding Effortful Listening (FUEL; Pichora-Fuller et al., 2016), for example, integrates Kahneman's (1973) capacity model of attention and Brehm & Self's (1989) motivational intensity theory into a unified concept of listening effort (Richter, 2013). The FUEL considers the amount of effort expended during listening to be dependent on a listener's motivation to achieve a goal and/or obtain rewards (Pichora-Fuller et al., 2016).

Similarly, Lemke & Besser (2016) note that a listener's motivation to succeed at a listening task influences their listening experience and allocation of cognitive resources in service of the task. Factors that increase motivation to achieve a task goal, such as the perceived importance or likelihood of success, therefore also affect effort (Richter, 2013).

Measures using isolated sentences might not capture processes underlying real-world listening

The isolated sentences used for speech in noise testing might fail to tap into the processes underlying real world communication. These sentences tend to lack personal relevance to listeners and as a result not be particularly interesting (e.g., "They are buying some bread."; Wilson et al., 2007). Additionally, they lack broader context, limiting the contextual cues that can be used by listeners to aid linguistic processing of speech. In contrast, the speech that listeners engage with in everyday life is often interesting and has meaningful context, such that listeners are intrinsically motivated to comprehend it (Picou et al., 2014). Accurate expression, and thus measurement, of listening effort might require the use of ecologically valid speech materials that intrinsically motivate hearers to listen.

1.5 Stories might motivate effortful listening more than isolated sentences

The use of narrative stimuli (e.g., spoken stories) in measuring listening effort might have several advantages over isolated sentences. Stories are often embedded in a rich context that intrinsically motivates listening. Stories have high ecological validity, evidenced by their ubiquity across human history and societies (Brown, 2004). They play an important role in everyday life, as they promote social connectedness (Smith et al., 2017), play a role in self-identity formation (Bamberg, 2011), and help us understand our relationship to the world and others (Dunlop & Walker, 2013). It has also been suggested that storytelling may have been evolutionarily advantageous to the hunter-gatherer ancestors of humans because it facilitates cooperation (Smith et al., 2017). In accordance with the importance of narratives in everyday life, there might be brain networks that uniquely subserve listening when speech is engaging and meaningful, and listeners are

therefore intrinsically motivated to comprehend it. Stories might provide a richer, more naturalistic window on how people experience listening challenges in the real world.

1.6 Listening effort arises from an interaction between processing demands and cognition

Listening effort depends on processing demands, cognitive abilities, and motivation

We theorize that listening effort arises due to an interaction between the processing demands imposed by a speech stimulus and the cognitive resources that an individual possesses for coping with these demands (Johnsrude & Rodd, 2016). Processing demands impose loads on cognitive resources (Wendt et al., 2016). Different individuals have unique profiles of cognitive abilities. (Akeroyd, 2008; Bharadwaj et al., 2015; O'Neill et al., 2019; Rudner et al., 2012; Sommers et al., 2015; Zekveld et al., 2012). Individuals differ in terms of their working memory capacity, IQ, fluid intelligence, and other cognitive abilities. As a result, different individuals are differentially equipped to cope with a given processing demand. The same processing demands may be met to different degrees in different individuals. When one's cognitive resources are only just sufficient for coping with processing demands, listening effort is experienced (Johnsrude & Rodd, 2016).

Motivation, attention, and engagement play interacting roles in influencing listening effort. Attention to a narrative can be motivated either extrinsically or intrinsically. Extrinsically motivated attention is based on externally imposed punishments or incentives; intrinsically motivated attention is generated internally based on personal interest or driven by stimulus characteristics (e.g., gunshots). We theorize that engagement can be thought of as intrinsically motivated attention. The decision to attend to a stimulus – which may have a subconscious and conscious component – is updated from moment-to-moment, as an individual listens to a narrative and receives feedback about their experience (in terms of their enjoyment, emotional engagement, comprehension, anticipation, and other dimensions of narrative listening). Motivation to listen modulates the amount of effort that an individual is willing to experience in service

of a listening task. This dictates at what point they will 'give up' and disengage from listening (Herrmann & Johnsrude, 2020).

Different processing demands tax different aspects of cognition

Different speech processing demands impose a load on different cognitive systems (Davis et al., 2011; Hervais-Adelman et al., 2012; Peelle & Wingfield, 2016; Rodd et al., 2012). Demands such as speech signal degradation and background noise, which physically occlude parts of the speech signal and render them unintelligible (i.e., energetically mask the speech), impose a load on cognitive processes such as knowledge-guided perception and verbal working memory (Johnsrude & Rodd, 2016; Wayne et al., 2016; Wild et al., 2012; Zekveld et al., 2013). The cognitive processes recruited are involved in 'filling in the gaps' in the impoverished speech signal, for example, by relying on context or background knowledge. Such 'acoustic demands' (which are related to the clarity of a speech signal) impose load on the cognitive processes involved in increasing a speech signal's intelligibility (Johnsrude & Rodd, 2016). Other acoustic demands such as signal distortion, reverberation, or competing speech, which do not physically occlude the speech signal but still reduce its intelligibility (i.e., informationally mask the speech) impose a load on cognitive processes such as selective attention, sound source segregation, and voice identity processing (Johnsrude & Rodd, 2016). The cognitive processes involved in perceptually separating the speech signal from the noise are also recruited. Acoustic demands also include demands related to a speaker's vocal characteristics, such as an unfamiliar accent or underarticulation (Johnsrude & Rodd, 2016). Such demands reduce intelligibility and are met by a combination of the aforementioned cognitive processes (Johnsrude & Rodd, 2016).

In contrast to acoustic demands, linguistic processing demands do not reduce the intelligibility of a speech signal but make it difficult to resolve its semantic meaning (Holmes et al., 2018). Such demands – which include the presence of homophones (words that have the same pronunciation but different meanings) and syntactic complexity – impose a load on cognitive systems that support linguistic processing (Johnsrude & Rodd, 2016; Rodd et al., 2012; Rodd et al., 2005). The processes recruited

to cope with linguistic demands – such as verbal working memory – probably overlap with those recruited to cope with acoustic demands.

In addition to acoustic and linguistic demands, speech in the form of narratives also presents demands related to the broader meaning of the narrative (Naci et al., 2014). These demands impose a load on higher-order functions that coordinate and plan more basic cognitive processes. Such functions include thinking about the meaning behind a character's actions and making predictions about the plot (Naci et al., 2014). This *executive load* is conceptually somewhat different from the *cognitive load* imposed by the acoustic and linguistic demands of words and sentences. A complete characterization of listening effort might require that the materials used to measure it present executive demands related to discourse processing, in addition to acoustic and linguistic demands common to words and single sentences.

1.7 Dual-task response times serve as a proxy of executive load

Dual-task paradigms have been successfully used to measure the cognitive load imposed during listening to sentences. In one study, participants listened to sentences that either contained homophones (e.g., "she filed her nails before she polished them") or no homophones (e.g., "there was beer and cider on the kitchen shelf"; Rodd et al., 2010). This task served as the 'primary task'. While listening, participants performed a letter case-judgement task (secondary task). A letter was presented on a computer screen at an unpredictable point in time and in response participants were required to categorize the letter as upper- or lowercase as quickly as possible with a corresponding keypress. Response times on the secondary task were longer for sentences containing homophones than for sentences containing no homophones, even though both types of sentence had a single, clear, sentence-level meaning. The researchers interpreted this finding as indicating that the semantic ambiguity created by the presence of homophones imposes a load on cognitive systems that are also required to perform the case judgement task, such as those involved in response selection on a visual judgement task. This overlap suggests that both tasks rely, to some extent, on the same domain-general cognitive systems. These cognitive systems may correspond to those that support language processing under

challenging listening conditions, including the left inferior frontal gyrus (a component of the FPN; Rodd et al., 2010). This experiment provides a means of behaviourally assessing the cognitive load associated with speech materials.

Dual-task paradigms have also been successfully used to measure the executive load imposed during listening to narratives. In one study, participants watched an engaging movie and answered comprehension questions when it was over (primary task; Naci et al., 2014). While listening, they performed a go/no-go task (secondary task). A number from one to nine was presented on a computer screen every two seconds and participants were required to press the corresponding numeric key on a keyboard, as quickly as possible ("go"), except when a specified digit was presented, in which case they were to withhold their response ("no-go). As attentively viewing a movie presents demands related to comprehending the executive aspects of the movie (e.g., those related to plot), the researchers interpreted RTs to the dual task as an index of executive load. This interpretation was then validated in a separate experiment, conducted on a new sample, in which participants listened to the story while their brain activity was recorded using fMRI. The averaged time course of dual-task RTs collected from the first experiment was used as a regressor in the model of brain activity for participants who listened to the story in the second experiment. This analysis was intended to reveal to what extent RTs in the dual task reflected executive load. Brain activity in higher order (e.g., FP) networks related to executive processing was strongly predicted by dual-task RTs, suggesting that dual-task RTs are in fact a reliable index of executive load (Naci et al., 2014). These experiments provide a foundation for behaviourally measuring executive load related to narrative comprehension and identifying brain networks sensitive to this load.

1.8 Intersubject correlations between neural responses index engagement with narratives

Cognitive neuroscientists are increasingly using naturalistic stimuli, such as movies and narratives, to study the brain in an ecologically valid manner (Nastase et al., 2019). Past research has shown that when different individuals are exposed to the same narrative stimulus (e.g., a story), their brains show synchronized patterns of activity,

throughout much of the brain, including auditory, visual and FP cortices (Hasson et al., 2004; Hasson et al., 2008). This synchronization is thought to reflect the degree to which a brain region is involved in the processing of the story (Hasson et al., 2004). Since stories cannot be easily modelled using traditional parametric approaches, researchers often use correlations between brain activity time courses of different individuals to analyze neural responses to narrative stimuli. This approach – referred to as intersubject correlation (ISC) – has been applied across neuroimaging domains, including fMRI (Nastase et al., 2019), EEG (Dmochowski et al., 2012), and MEG (Chang et al., 2015).

Different factors, including the level of coherence of the story and the listeners' engagement with the story, influence the magnitude and extent of ISC observed, particularly in higher order (often frontoparietal) brain regions. Temporally scrambling a story results in a reduction in FP ISC, where the magnitude of the reduction depends on the level of narrative organization (e.g., paragraph, sentence, or word) at which it is scrambled – i.e., an intact story drives greater FP ISC than does a story scrambled at the paragraph level than does a story scrambled at the word level (Hasson et al., 2008; Simony et al., 2016). Correlations in auditory and visual cortices are more stable (Naci et al., 2014; Simony et al., 2016). This hierarchical functional activity reinforces the importance of using narrative stimuli to study speech, as neural responses depend on the level of linguistic processing required to comprehend the speech (Peelle & Wingfield, 2016). Frontoparietal ISC also depends on the degree to which a story engages listeners – a more engaging story elicits greater ISC than does a boring story (Dmochowski et al., 2012, 2014; Ki et al., 2016; Schmälzle et al., 2015). This modulation is thought to reflect the extent to which a story drives consistent higher- order processing of its meaning across individuals (Naci et al., 2014). An engaging story captures listeners' attention, allowing the story to drive executive processing – reasoning about characters' actions, forming predictions, mentally solving problems related to the plot, retaining information about plot in working memory – in a manner that is consistent across individuals (Naci et al., 2014). As the higher-order neural responses observed therefore depend on how engaging the story is, it is important to select engaging materials for experiments using narratives.

1.9 Correlated time courses among healthy individuals are sensitive to abnormal mental states

Intersubject correlation can be used to detect neural indices of abnormal populations. Abnormalities cause consistent functional changes in the brain (Hasson et al., 2009). Individuals with Down syndrome (Anderson et al., 2013) or autism (Hasson et al., 2009; Salmi et al., 2013) show widespread reduction in neural ISC during movie viewing or story listening when compared with healthy controls, in particular in brain regions related to the processing of social information (e.g., default mode network; Salmi et al., 2013). Additionally, ISC patterns differ with personality traits (Finn et al., 2018). These findings suggest that brain activity that is synchronized between healthy individuals during narrative exposure can be used to reveal abnormalities within groups. In accordance with this, Naci et al. (2014) used an ISC analysis to determine if two behaviourally nonresponsive patients who were presented with an engaging movie showed signs of consciousness. Correlated time courses of executive load collected from healthy individuals were averaged and used to predict the patients' brain activity, recorded using fMRI, in regions associated with executive function (FPN). In one patient, the averaged executive load time course significantly predicted FPN activity, suggesting that processing of the movie's executive demands was similar in the patient compared to the healthy group. The researchers used this finding to infer that the patient was, in fact, conscious, despite being behaviourally nonresponsive. The second patient's FPN activity, however, was not significantly predicted by the averaged executive load time course, suggesting abnormality. The authors speculated that this patient did not have a similar conscious experience of the movie's executive demands and was likely not conscious. Importantly, the researchers showed that the averaged executive load time course significantly predicted FPN activity during movie viewing in every healthy individual (Naci et al., 2014). I intend to extend this approach to detect abnormalities in neural processing of a spoken story when listening is effortful.

1.10 Overview of present study

Objective, rationale, & hypothesis

In the present study, our overarching objective was to develop tools for characterizing spoken stories that could be used to investigate listening effort in a sensitive and ecologically valid manner. Unlike the highly controlled sentences typically used to assess listening effort, the demands of a natural narrative are not as tightly controlled. Narrative demands – talker, linguistic, executive, etc. – all vary dynamically over the time course of the story. Reasoning that it is important to understand the demands of such a stimulus before using it to measure listening effort, we employed a case-judgement task to index cognitive load during speech listening (Rodd et al., 2010). Cognitive load serves as a gross readout of speech-stimulus demands, to the extent that they act on a listener's brain activity. We adapted the dual task method used by Naci et al. (2014) to characterize the load imposed by an auditory, as opposed to audiovisual, narrative over its time course. We adopt the term 'executive' load to more specifically refer to the load imposed by the integrative demands of narrative comprehension, which we were most interested in characterizing.

To assess the reliability of the obtained time courses, we developed a modified ISC analysis for analyzing behavioural responses to a natural stimulus, such as a narrative. Building on the results of Wild et al. (2012), for which participants' attention to speech stimuli was extrinsically motivated by experimenter instructions, we encouraged intrinsically motivated attention by selecting stories that we considered engaging. Our reasoning was that engagement is critical to the experience of listening effort, and a key factor that existing clinical tests often fail to elicit. To ensure the suitability of the stories we selected for future clinical use and investigations of the neural basis of listening effort, we analyzed to what extent ISC of behavioural responses is modified by individuals' engagement with a narrative.

We conducted the study on a sample of young individuals with normal hearing, with the intention of obtaining a normative reference against which individuals with hearing impairment can be compared in future studies. To investigate how listeners

engage with stories in the absence of added processing demands such as background noise, we used acoustically clear stories. Following the lab experiment, we designed and conducted an online experiment on a sample of individuals with normal hearing, to assess the replicability of the results obtained in the lab. Age was not constrained for the online replication, allowing us to investigate to what extent our results generalized to a broader age group. We hypothesized that stories drive consistent executive processing – that is, consistent recruitment of cognitive resources in service of the listening task, and more specifically, in service of narrative comprehension – and that this consistency depends on engagement.

Aims & methods

Our specific aims were two-fold. First, we sought to measure the temporal dynamics of cognitive resource recruitment associated with the unique executive demands of two short spoken stories. To characterize executive load we used a dual-task paradigm similar to that employed by Rodd et al. (2010). Each participant listened to two spoken stories while performing a concurrent case-judgement task. Participants answered comprehension questions after each story. Our second aim was to test whether ISC depends on the degree of engagement in the story. To address this aim we measured engagement using a subjective questionnaire, and examined whether ISC of case-judgement RT time courses was positively correlated with engagement. We adapted the ISC analysis commonly used in neuroimaging experiments (Hasson et al., 2004) to compute correlations between time courses of behavioural responses (RTs) to the case-judgement task. We then computed the correlation between engagement ratings and similarity of a given participant's behavioural responses with the group-averaged response.

Chapter 2

2 Methods

2.1 Lab experiment

2.1.1 Participants

Seventy adults (mean: 21.1 years; range: 18-27 years; 42 female) from Western University (Ontario, Canada) took part. Six participants who reported that they were non-fluent English speakers were dropped from the analysis. Within each participant, the experiment was run in two blocks, each corresponding to one story. Two blocks were dropped due to violation of one or more of the following exclusion criteria: no response or an erroneous response on more than 15% of case-judgement trials ($N = 2$); a score on the comprehension questionnaire of less than 70% (three or more errors out of 10 comprehension questions; $N = 0$); or familiarity with the stimulus (i.e., the story the participant heard during the block; $N = 1$). The final sample consisted of 63 participants. Each version of the experiment (see below) was run in at least 10 participants.

This study was approved by the Western University Non-Medical Research Ethics Board. Participants were recruited through email or Western University's Psychology SONA pool and compensated with $5 CAD per hour of participation or 0.5 course credits per hour if they were enrolled in an applicable Western University course. Written informed consent was obtained from each participant prior to the experiment. In addition, participants completed a demographics questionnaire consisting of questions about their language background and hearing abilities. All participants reported having normal hearing, normal or corrected-to-normal vision, and no known neurological impairments.

2.1.2 Story and questionnaire materials

Two stories were selected to be used in this experiment, both recorded at NPR's live storytelling event *The Moth*: "Alone Across the Arctic" (which we will refer to as "Arctic"), told by Pam Flowers (13.3 minutes; 16 bits/sample; 44.1 kHz sampling rate), and "Swimming with Astronauts" (which we will refer to as "Space"), told by Michael Massimino (13.5 minutes; 16 bits/sample; 44.1 kHz sampling rate). The amplitudes of

both stories were root-mean-square normalized to the same sound level. Following each story, participants viewed two questionnaires in sequence: a comprehension questionnaire and a questionnaire assessing their experience of the story. Each comprehension questionnaire (Appendix A) consisted of 10 multiple choice questions, each with four options (e.g., "How did the narrator train for his swim test?": A: "Took his kids to the pool every day"; B: "Practiced martial arts"; C: "Signed up for a swim class"; D: "Swam at the lake every day").

The second questionnaire assessed listeners' experience of the story in terms of four dimensions of narrative engagement (Busselle & Bilandzic, 2009): enjoyment, attention, emotional engagement (i.e., the experience of different affective states), and mental simulation (i.e., imagination of the world of the story). This questionnaire contained 18 items that participants rated on a Likert scale from 1 ("strongly disagree") to 7 ("strongly agree"). Each item targeted one dimension of narrative engagement (e.g., enjoyment item 1: " I thought it was an exciting story."; attention item 1: "When I finished listening I was surprised to see that time had gone by so fast."). Three items targeted mental simulation; five items targeted each of the remaining dimensions. Prior to the analysis, we chose to focus on two subscales of engagement: enjoyment and absorption. Absorption (i.e., immersion in the world of the story) is a subscale that combines attention, emotional engagement, and mental simulation. An independent investigation revealed that these subscales explain the majority of variance in responses. Item presentation order was randomized for each questionnaire and participant.

2.1.3 Method

Experimental design

Participants listened to both stories, each constituting one block of the experiment. Story order was counterbalanced, such that half of the participants listened to "Arctic" first and the other half listened to "Space" first. Participants were instructed to listen carefully to the story and understand it as best they could, while simultaneously performing a letter case-judgement task. For this task, they were instructed to press a key on a keyboard as quickly as possible whenever they saw a letter appear on the computer

screen in front of them, to indicate whether that letter was lower- or upper-case. Letters appeared at pseudorandom time points during the story (every 6-14 sec). Responses and response times (RTs) were recorded for each trial.

To enable correlation between the case-judgement RT time courses of different participants (refer to *Analysis* section), we developed a sampling protocol that allows pseudorandom trial presentation for each individual but yields evenly spaced trials when groups of participant time courses are combined. Six different versions of the case-judgement task, each with a different temporal distribution of case-judgement trials, were generated for each story, with constraints set so that within each version, trials occurred at pseudorandom time points between 6 and 14 seconds apart. Across the six versions, a trial occurred every two seconds. This permitted RT time courses with a sampling resolution of 0.5 Hz to be extracted from subsets of six participants, each assigned to a different version, and subsequently correlated. We refer to a single configuration of participants sampled in this manner as a ‘supersubject’ (Fig. 1; Appendix B). The six timing versions were counterbalanced across participants, for each story.

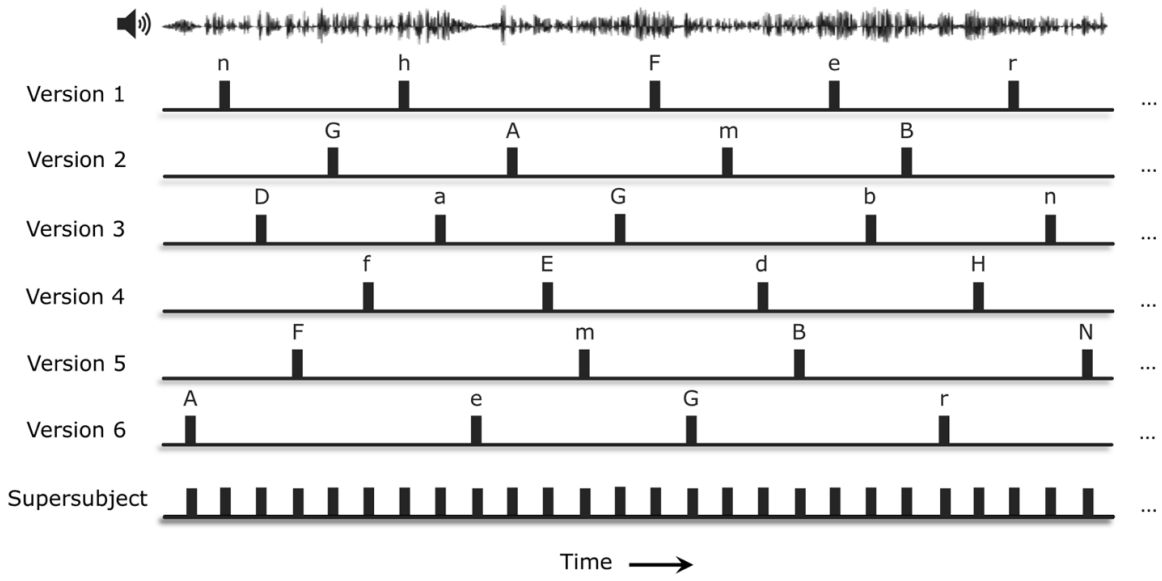

Figure 1. Supersubject sampling. Supersubject time courses are obtained by collapsing across groups of time courses. Each line in this visualization represents the unique distribution of times relative to the story at which letters appear for a given timing version over a 52 s window. Each subject is assigned to one version. They would see letters (random in identity and case) appear at time points indicated by the ticks. When the time points across six participants (one per version) are collated into one "supersubject" (bottom line), the result is a time course with regularly spaced samples: in our experiment, the supersubject sampling resolution was 0.5 Hz.

Participants heard each of the two stories within a single-walled sound-attenuating chamber (Eckel Industries). Sound was played over closed-ear headphones (Sennheiser HD 280 PRO), from a Steinberg UR22MKII audio interface (sampling at 16 bits; 44.1 kHz) connected to a Lenovo ThinkPad X270 laptop. Participants viewed a fixation cross (40 pixels) in the center of a computer screen. Letters appeared on the screen in place of the fixation cross at time points determined by the timing version to which they were assigned. For each participant, the presented letters were randomly drawn from a subset of letters (A/a, B/b, D/d, E/e, F/f, G/g, H/h, N/n, Q/q, R/r, T/t), excluding those with a similar lower-/upper-case appearance (e.g., O/o; Rodd et al., 2010). Letter randomization was performed under two constraints: the same letter (irrespective of lower-/upper-case form) was never presented multiple times in sequence, and no more than three lower- or

upper-case letters were ever presented in sequence. For lower case letters, participants were instructed to press the ‘x’ key on a keyboard using the index finger of their left hand; for upper case letters, they were instructed to press the ‘m’ key on the keyboard using the index finger of their right hand. Each letter remained on the screen for two seconds or until participants hit a valid key.

2.1.4 Analysis

Processing

For each story and participant, case-judgement RT was standardized to the mean and standard deviation. Standardization was carried out separately for lower- and upper-case letters, as previous research using a similar paradigm showed that participants tend to respond slower to lower case letters (Rodd et al., 2010). Individual case-judgement responses were dropped if they were incorrect (e.g., if a participant pressed the ‘m’ key in response to a lower case letter) or if RT was greater than 1.96 (z_{crit} for $\alpha = .025$) times the standard deviation above the mean RT for that participant, story, and letter case (in which case the trial was deemed an outlier). A median of 1 erroneous trials and 2 outlier trials were dropped from the analysis for each participant. Response-time time courses were then collapsed over groups of participants, each tested using a different timing version, into single supersubject time courses. The resulting time courses reflected the concurrent executive demand of the story, sampled at 0.5 Hz. Exhaustive sampling of the dataset without replacement yielded 10 supersubjects. We refer to a single configuration of supersubjects obtained by sampling in this manner as a ‘set’.

To ensure that the correlation analysis is not biased by the specific configuration of supersubjects chosen, we used bootstrapping to randomly generate 50 sets of supersubjects and performed the analysis within each set. As there are 2.28e39 possible unique sets (refer to Appendix B for proof), it was not computationally feasible to perform an exhaustive analysis. To ensure that the distribution of correlations was representative while not sacrificing computational efficiency, we performed the analysis within 50 randomly chosen sets. We found that correlations were highly reliable across

sets (refer to range statistics under *Inter-supersubject time course correlations* section, below).

Inter-supersubject correlation

Correlations between supersubject time courses and permutations for assessing significance were performed within each set. To correlate supersubject time courses, we used an exhaustive split-half approach. Each set was split into two equal-sized halves of supersubjects (their configuration constituting a single 'split'). Each half was then averaged into a single time course and the two resulting time courses were correlated. Averaging across supersubjects within split-halves (as opposed to, for example, performing pairwise correlations between supersubjects) ensures that each RT time course is representative of several participants' performance.

Within each set, we performed split-half inter-supersubject correlations exhaustively, such that every unique configuration of two equal-sized groups of supersubjects was generated, averaged, and correlated. This results in $\frac{1}{2}\binom{n}{k}$ unique splits and a distribution of corresponding 'observed' correlations, where *n* is the number of participants per timing version and $k = \frac{n}{2}$. For the lab data, this resulted in 126 unique splits (i.e., 126 correlation values within each set).

To test the hypothesis that the time courses are correlated, against the null hypothesis that they are not, we temporally permuted the time courses. If changing the temporal alignment between two time courses has no effect on the correlation, then they were not correlated to begin with. Permutations were performed by circularly shifting supersubject time courses by a random number of time points ranging from the selected temporal smoothing window size to the number of samples in the time course. One averaged supersubject time course within each split was shifted and correlated with the other, unshifted, time course. We refer to the resulting Pearson correlation coefficient between average supersubjects as the 'ISC' (in keeping with the commonly used term 'intersubject correlation', which typically refers to correlations between brain activity of individual subjects).

Circularly shifting a time series only alters phase, preserving both amplitude and frequency spectra (Lancaster et al., 2018). However, temporal smoothing, in addition to enhancing the signal-to-noise ratio of noisy data, 'smears' data across the span of the smoothing window. As a result, circularly shifting time courses might yield inflated permutation correlations for any shift smaller than the window size. Under the null hypothesis that the time courses are not correlated, these inflated correlations are artifactual. To ensure that we did not inflate permutation correlations, we limited the range of possible shifts to only those above the window size (i.e., all whole numbers between the window size and the total number of samples in the time course). This ensures that all permutation correlations are equally likely under the null hypothesis. Five hundred permutations and correlations were performed, resulting in a null distribution of 500 permutation correlations (i.e., a 'permutation distribution') for each observed correlation within a set. This was repeated 126 times, for each of the 126 correlation values.

To determine the significance of the observed (true) correlations, we compared each true correlation to those obtained through permutation. Prior to computing statistics, we normalized all correlations using Fisher's *z*-transformation. We computed the significance of each true correlation as the proportion of values in the corresponding null (permutation) distribution that exceeded the true correlation value. This proportion provides a measure of the likelihood of observing the true correlation under the null hypothesis. We obtained summary *r* and *p* statistics by computing the median, first across splits, and then across sets. We inverse Fisher transformed the resulting *r* value.

To investigate further whether each story elicited consistent behavioural performance, we performed an analysis that accounts for the complete distribution of observed correlations, as opposed to their average. Within each set, we collapsed all observed correlations into an 'observed distribution', and all correlations obtained by permutation into a single aggregate permutation distribution. We computed the median range and variance of each observed and permutation distribution as a measure of their dispersion. We then compared each observed and aggregate permutation distribution by computing a receiver operating characteristic (ROC) curve. We computed *d'* and area

under the curve (AUC) as a measure of the overlap between the distributions (i.e., how discriminable they are). We interpreted the average of these metrics across sets as a measure of how strongly a story elicited consistent behavioural performance, indicative of consistency in executive load over listeners and the degree to which the story consistently 'drives' cognition over listeners. A high average *d'* and AUC suggests that the story consistently drives cognition across listeners. A numerical summary of the correlation and resampling procedure is provided in Appendix B.

Executive load time course variance analysis

Next, we investigated the temporal dynamics of the supersubject time courses. Our aim was to identify consistent changes over time in the average executive load time course for each story, which would help to explain any observed inter-supersubject correlations. We refer to the average executive load time course for a given story as its 'canonical executive load time course'. We performed this analysis within each set to account for any variations in average RT between different configurations of supersubjects. This was necessary because subtle variations may have been introduced by temporally smoothing supersubject time courses within sets. Therefore, within each set, we computed the mean supersubject RT time course, as well as the variance at each time point. We obtained the canonical executive load time course for each story by averaging supersubject RT time courses across sets. Performing this analysis set-wise allowed us to account for subtle variations in average RT between sets. A very similar result can be obtained by averaging RT time courses across individual participants.

To determine the significance of each point in the canonical executive load time course, we found the average time course of variance across sets. The resulting time course was converted to a 95% confidence interval on the canonical time course, $\bar{x}$, based on, $\bar{x} \pm z_{.025} \cdot \frac{\sigma}{\sqrt{n}}$, where *n* is the number of supersubjects within a single set. The significance of each point in the canonical time course was tested against zero (since RT was standardized): significant RTs were those whose 95% confidence interval did not intersect with zero.

We then independently analyzed the contribution of significant and nonsignificant RTs to the observed inter-supersubject correlations. We extracted the significant sections of each supersubject time course – as determined by comparing the canonical time course to zero – within each set and concatenated them. This was repeated for all nonsignificant sections. We then ran the inter-supersubject correlation analysis independently on the 'significant time courses' and 'nonsignificant time courses' and determined the significance of the resulting median ISC values. We also compared these values using a one-tailed *z* test (significant time course ISC > nonsignificant time course ISC) of the difference between independent correlations. All correlation values were Fisher *z*-transformed prior to comparison.

Next, we conducted a qualitative analysis to relate significant sections of the executive load time courses to story characteristics. This analysis was based on the 'reverse correlation' analysis described in the seminal intersubject correlation paper (Hasson et al., 2004).We first identified each segment of the story with eight or more consecutive significant RT samples (corresponding to 16 sec of the story). This focused our analysis on significant segments that were long enough to carry meaningful information. To ensure that the segments we identified were meaningful with respect to both the lab experiment and the online replication, we reduced the segments of interest to regions of overlap between the identified segments for both experiments. To account for any slight misalignments between the experiments due to smearing as a result of temporal smoothing, we widened each segment of interest by the smoothing window size. We then extracted the speech spoken by the narrator during each identified overlapping section. This analysis allowed us to make qualitative observations about the characteristics of the stories that might give rise to RTs that consistently differ from zero across participants. More generally, these observations shed light on the story characteristics that might drive behavioural consistency, reflecting consistent narrative-driven cognition.

Assessing relationship between inter-supersubject correlation and engagement

Our next aim was to examine whether objectively measured correlations related to subjectively rated enjoyment and absorption. Since we perform correlations at the supersubject level, we rank ordered the participants within each timing version according to their ratings on the narrative engagement questionnaire and assembled supersubjects in that order. This allowed us to obtain a range of ratings corresponding to different supersubjects. The result is a set of supersubjects, the first composed of the least engaged participants and the final composed of the most engaged participants. This was done separately for the two narrative engagement subscales of interest (enjoyment and absorption). We then calculated each supersubject's average rating of enjoyment or absorption and computed the correlation between each corresponding supersubject's average executive load time course and the canonical executive load time course for the same story. To ensure the independence of the time courses being correlated, we computed a unique canonical executive load time course for each comparison, leaving out the supersubject being correlated. In other words, for a given supersubject and story, the canonical executive load time course was computed as the average RT time course across all individual participants – except those making up that supersubject – and smoothing the result. We then calculated the Spearman correlation between each supersubject's average rating and correlation. To maximize our sample size and resolution for this analysis, we pooled the data for each condition – "Arctic"/"Space" and lab/online (see below for details of online experiment).

2.2 Online experiment

2.2.1 Participants

An online version of the experiment was subsequently run in an independent group of individuals to assess the replicability of the lab results. One hundred thirty-six fluent English speakers (mean: 40.6 years; range: 22-69 years; 42 female) took part in the online experiment. Fifty blocks were dropped due to violation of one or more of the following exclusion criteria: no response or an erroneous response on more than 15% of

case-judgement trials ($N = 30$); a score on the comprehension questionnaire of less than 70% ($N = 34$); or familiarity with the stimulus ($N = 0$). The final sample consisted of 116 participants. Each version of the experiment was run in at least 14 participants.

The online experiment was programmed using jsPsych, a JavaScript library for running behavioural experiments in a web browser, and hosted on Pavlovia, an online platform for running behavioural experiments. Participants were asked to wear headphones and perform the experiment in a silent environment. The experiment that these participants viewed in their web browser looked very similar to the in-person experiment in terms of questionnaire presentation, trial presentation, and randomization. Minor differences were due to technical limitations with jsPsych or Pavlovia or participants' failure to comply with the experiment instructions. The online data were processed and analyzed using the exact same pipeline as the lab data.

This study was approved by the Western University Non-Medical Research Ethics Board. Participants were recruited through Amazon Mechanical Turk (MTurk) using CloudResearch (a participant-sourcing platform that interfaces with MTurk) and compensated with $5 CAD per hour of participation. Electronic informed consent was obtained from each participant prior to the experiment. All participants completed the Demographics Questionnaire and reported having normal hearing, normal or corrected-to-normal vision, and no known neurological impairments.

2.2.2 Story and questionnaire materials

2.2.3 Method

Experimental design

Participants heard each of the two stories through a pair of their own headphones, connected to their personal computer. Presentation of the online experiment was closely modelled after the lab experiment.

2.2.5 Analysis

Processing

A median of 1 erroneous and 2 outlier case-judgement trials were dropped from the analysis for each non-rejected participant. Exhaustive sampling of the online dataset without replacement yielded 14 supersubjects. There were 4.39e65 possible unique sets for the online experiment (refer to Appendix B for proof). Processing of the online dataset was identical to that of the lab dataset.

Inter-supersubject correlation

There were 1716 unique splits for a given set (i.e., 1716 correlation values within each set). The permutation analysis of shifting one average supersubject time course in a split and correlating both resulting time courses was repeated 1716 times, for each of the 1716 correlation values. Inter-supersubject correlations were computed using the same analysis applied to the lab dataset.

Time course variance analysis

This analysis was identical to that run on the lab dataset.

Assessing relationship between inter-supersubject correlation & engagement

This analysis was identical to that run on the lab dataset.

2.3 Combined analysis

The following analyses were conducted by combining the lab and online datasets.

2.3.1 Smoothing window size determination

Following the matched-filter theorem, we reasoned that the temporal smoothing window that maximizes the correlation (i.e., best captures the story-driven ‘signal’) is the one best matched to the dynamics of executive load during story listening (Jacobson,

1989). To identify this window, we temporally smoothed each supersubject time course using the lowess (locally weighted scatterplot smoothing) method with a window size ranging from 0 to 100 samples (0-200 s). The lowess method is more robust to outliers than moving average and kernel-weighted methods. We then calculated the median Pearson correlation between supersubject time courses for every window size (see *Inter-supersubject correlations*) and plotted ISC as a function of window size.. Ideally, a global maximum of this curve, corresponding to the window size that maximizes the agreement between supersubject time courses, would be identifiable. This would provide the optimal temporal smoothing window size, which captures information about the intrinsic temporal dynamics of the signal. However, as the window size of temporal smoothing increases, two smoothed time series tend to approach horizontal lines, with the correlation between them approaching one. In this case, the global maximum of the 'ISC by window size' curve corresponds to an artifactual correlation introduced by the analysis, rather than a true correlation captured by enhancing the intrinsic dynamics of the signal. Therefore, a global maximum corresponding to this intrinsic window cannot always be identified. However, it may still be possible to identify an inflection – or 'knee' – point, after which the curve begins to flatten out, indicating that increasing the temporal window size yields little additional benefit to the reliability of supersubject time courses.

To increase our confidence in the location of this knee point, we performed a follow-up analysis, this time using correlations between the average executive load time courses from independent datasets (lab and online) as our dependent measure. This analysis extended the 'replication analysis' (described below) over a range of window sizes. We obtained the canonical executive load time course for each story and experiment (lab and online) by averaging RT time courses across all individual participants and temporally smoothing the result using the lowess method for a window sizes ranging from 0 to 100 samples. In contrast to the time course analysis (described above), for which the executive load time course was calculated by averaging supersubjects set-wise, this method of obtaining the canonical time course gives a very similar result but is less computationally demanding. We considered the optimal smoothing window to be the one that maximizes the correlation between the canonical

executive load time course for the lab and online experiments. If a global maximum could not be identified, we identified the knee point instead. We reasoned that this between-dataset reliability analysis is less susceptible to the artifactual correlations observed in the previous (within-dataset) analysis because it only involves a single correlation between distinct, representative canonical time courses, rather than an average correlation over many sets of less representative supersubject time courses.

2.3.2 Replication analysis

The similarity of the results obtained in the lab and online was quantified as the agreement between each experiment's canonical executive load time course for the same story when temporally smoothed with the window size identified in the analysis described above. We computed the Pearson correlation between each lab/online ('IL'/'OL') pair of the four executive load time courses (IL-Arctic; OL-Arctic; IL-Space; OL-Space), to quantify their similarity. This resulted in two 'congruent' correlations (IL-Arctic, OL-Arctic; IL-Space, OL-Space) and two 'incongruent' correlations (IL-Arctic, OL-Space; OL-Arctic, IL-Space).

Using a one-tailed *z* test of the difference between independent correlations, we then tested the difference between the average correlation for the congruent comparisons and the average correlation for the incongruent comparisons ($r_{congruent} > r_{incongruent}$). Correlation coefficients were Fisher *z*-transformed prior to comparison. We reasoned that if the lab and online canonical time courses for "Arctic" and "Space" are not correlated (null hypothesis), the congruent correlations would be equivalent to (i.e., not greater than) the incongruent correlations. The outcome of this test served as a summary statistic of the similarity of the lab and online canonical executive load time course for a given temporal smoothing window size.

Chapter 3

3 Results

3.1 Behavioural performance

The modal comprehension score was 100% (range: 70-100%) for both stories. This was true for both the lab ($N = 63$) and online ($N = 112$) datasets. Fig. 2a and 2b show the distribution of comprehension scores for the lab and online experiment respectively. Ratings of enjoyment and absorption on the narrative engagement questionnaire are summarized in Table 1. Fig. 3a (lab dataset) and Fig. 3b (online dataset) show the distribution of ratings for each subscale and dataset. The ratings – from 1 ("strongly disagree") to 7 ("strongly agree") – for "Arctic" and "Space" were compared for each subscale and dataset using a two-tailed, paired *t*-test. For the online experiment, ratings of enjoyment were higher for "Arctic" than "Space", $t_{(105)} = 2.22$, $p = .028$. This was the only significant difference.

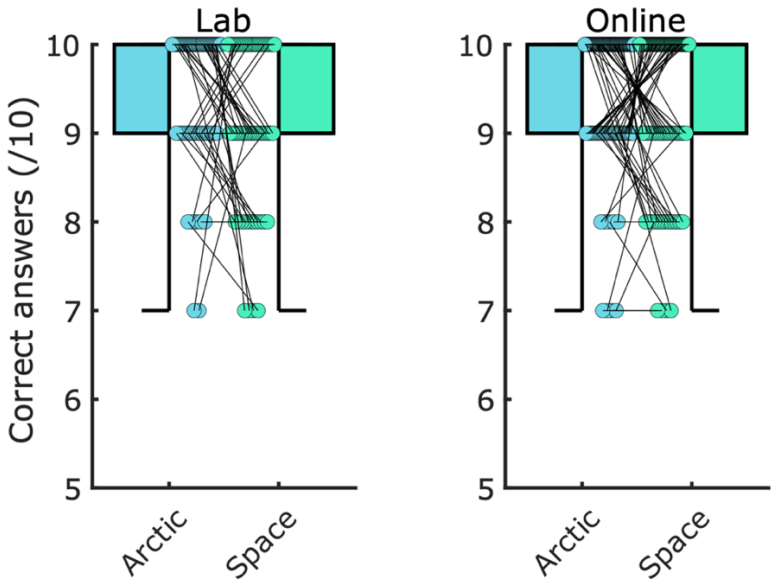

Figure 2. Distributions of comprehension scores. Average comprehension scores for "Arctic" (blue) and "Space (green), for the lab experiment (left) and online experiment (right). Comprehension was evaluated using a comprehension questionnaire with 10 multiple choice questions. Upper edge of the boxplot corresponds to the modal score; lower edge corresponds to the first quartile; upper and lower whiskers correspond to maximum and minimum; individual participant scores for "Arctic" and "Space" are connected with a line.

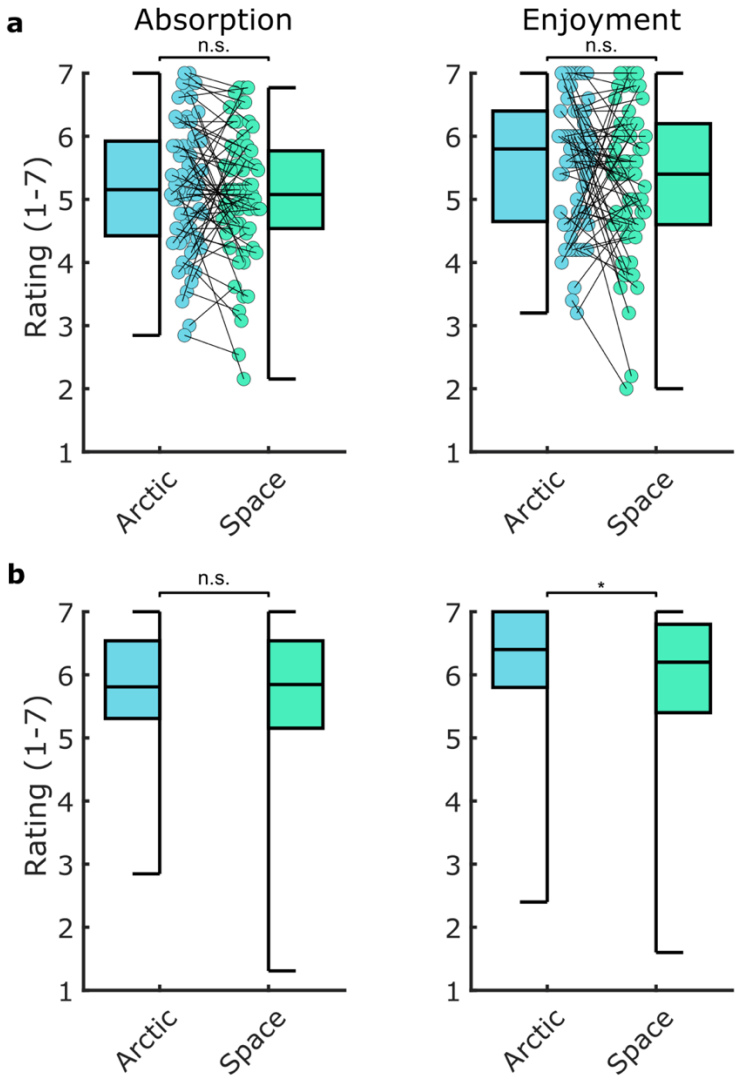

Figure 3. Distributions of engagement ratings: absorption and enjoyment. Average ratings of absorption (column 1) and enjoyment (column 2) for "Arctic" (blue) and "Space (green). *a* (top row): lab experiment; *b* (bottom row): online experiment. Upper and lower edges of the boxplots correspond to the third and first quartile, respectively; midpoint lines to median; upper and lower whiskers to maximum and minimum. Individual participant scores (not shown for the larger online dataset) for "Arctic" and "Space", shown by the coloured circles, are connected with a line. Braces show the significance of two-tailed, paired *t*-tests comparing ratings for "Arctic" and "Space". Asterisks indicate significant differences (n.s., not significant; *$p < .05$; **$p < 0.01$; ***$p < 0.001$).

Table 1: *Summary of engagement ratings: absorption and enjoyment.*

Condition	Rating (1-7)	
	Mean	Range
	Absorption	
Arctic/lab	5.2	2.8-7
Arctic/online	5.8	2.8-7
Space/lab	5.1	2.2-6.8
Space/online	5.7	1.3-7
	Enjoyment	
Arctic/lab	5.6	3.2-7
Arctic/online	6.1	2.4-7
Space/lab	5.3	2-7
Space/online	5.9	1.6-7

Mean RT on the case-judgement task before normalization was, for the lab dataset: 715.3 ms (σ = 147.5) for "Arctic" and 691.7 ms (σ = 128.7) for "Space"; for the online dataset: 731.6 ms (σ = 154.6) for "Arctic" and 736.3 ms (σ = 176.1) for "Space". Case-judgement RTs, separated by case, are summarized in Table 2. All participants included in the analysis comprehended the stories and performed the case-judgement task adequately. Mean RTs for lower- and upper-case letters were compared for each story and dataset using a one-tailed (lower > upper), paired *t*-test. All four comparisons yielded significantly longer RTs for lower than upper case letters as expected based on previous research (Rodd et al., 2010): "Arctic"/lab, $t_{(62)} = 1.87$, $p = .033$; "Arctic"/online, $t_{(111)} = 4.01$, $p < .01$; "Space"/lab, $t_{(62)} = 3.49$, $p < .01$; "Space"/online, $t_{(109)} = 5.14$, $p < .01$. Fig. 4a (lab dataset) and Fig. 4b (online dataset) show the distribution of average RTs for lower- and upper-case letters, for each story.

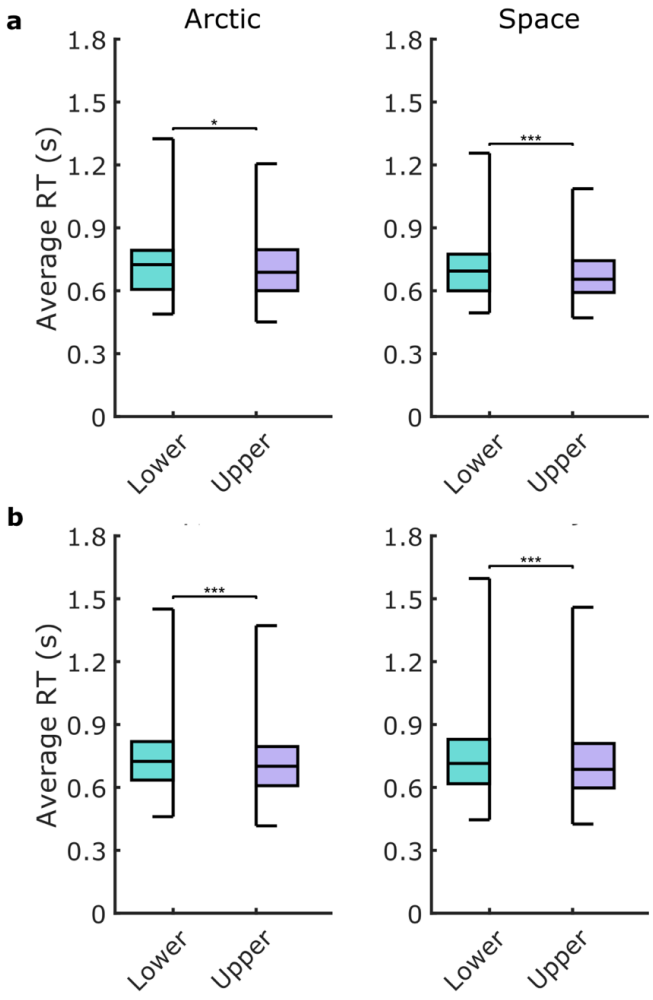

Figure 4. Distributions of case-judgement response time. Average non-standardized case-judgement response time for lower- and upper-case letters. *a* (top row): lab experiment; *b* (bottom row): online experiment. Column 1: "Arctic"; Column 2: "Space". Upper and lower edges of the boxplots correspond to the third and first quartile, respectively; midpoint lines to median; upper and lower whiskers to maximum and minimum. Braces show the significance of two-tailed, paired *t*-tests comparing mean lower case (turquoise) and mean upper case (purple) RT for "Arctic" and "Space". Asterisks indicate significant differences (n.s., not significant; *$p < .05$; **$p < 0.01$; ***$p < 0.001$).

Table 2: *Summary of case-judgement RTs.*

Condition	Case	Response time (ms)	
		Mean	Standard deviation
Arctic/lab	Upper	708.4	146.5
	Lower	722.5	154.9
Arctic/online	Upper	718.6	158.3
	Lower	744.4	157.8
Space/lab	Upper	679.6	125.5
	Lower	703.4	137.5
Space/online	Upper	720.8	172.5
	Lower	751.4	185.7

3.2 Inter-supersubject time course correlations

Pearson correlations between supersubject time courses were computed to quantify the degree to which the stories reliably drive cognition. Supersubject time courses were significantly correlated across participants, suggesting that cognitive resources are recruited reliably across participants during story listening. Inter-supersubject correlation statistics are summarized in Table 5. The median ISC value for each story and dataset was positive and significant ($p < .05$), indicating that the stories elicited reliable behavioural performance across participants. These results indicate that RT time courses are consistent across participants and capture meaningful information about the demands associated with the narrative stimulus. Fig. 5a shows the frequency distribution of observed inter-supersubject correlations and correlations obtained through permutation (null); Fig. 5b shows the median inter-supersubject correlation for each distribution. Fig. 5c shows the receiver operating characteristic (ROC) curves for the comparison between the observed and permutation correlation distributions, as well as the corresponding *d'* AUC, which serve as measures of the distributions' discriminability

(these statistics also summarized in Table 3). The area under the curve for each condition was above chance (.5), indicating again that the stories significantly elicited consistent behavioural responses.

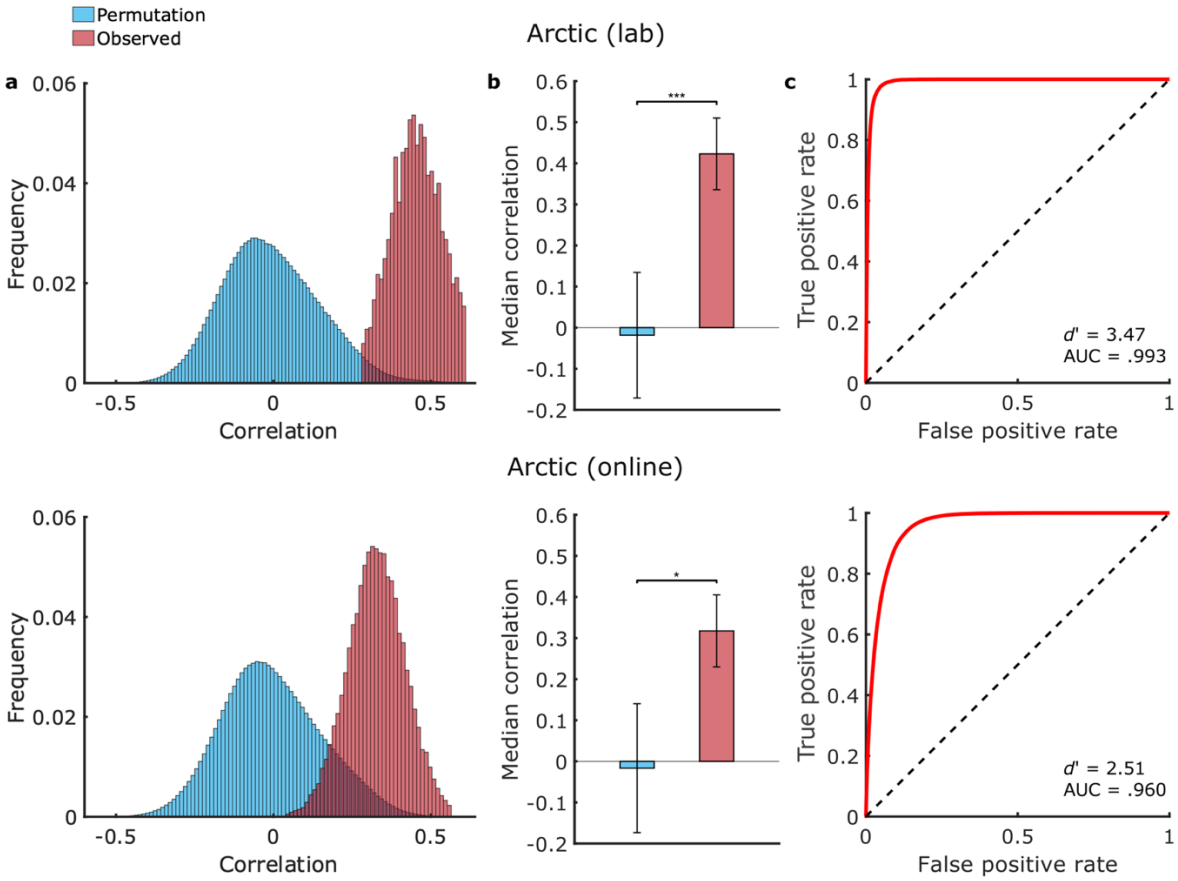

Figure 5. Cognitive recruitment is consistent across individuals during story listening. Results are shown for "Arctic". Top row: lab experiment; Bottom row: online experiment. *a*, Frequency distributions of observed (case judgement; red) and permutation (blue) correlations, averaged across sets. *b*, Median Pearson's correlation coefficient for each distribution. Error bars represent the standard deviation of the median correlation, computed as the square root of the median frequency distribution variance across all sets. Braces over bars show the significance of correlation values based on one-tailed permutation testing. Asterisks indicate significant differences (n.s., not significant; $*p < .05$; $**p < 0.01$; $***p < 0.001$). *c*, ROC curves comparing the observed and aggregate permutation correlation distributions. The dotted line indicates chance discriminability. Area under the curve (AUC) and d' quantify discriminability.

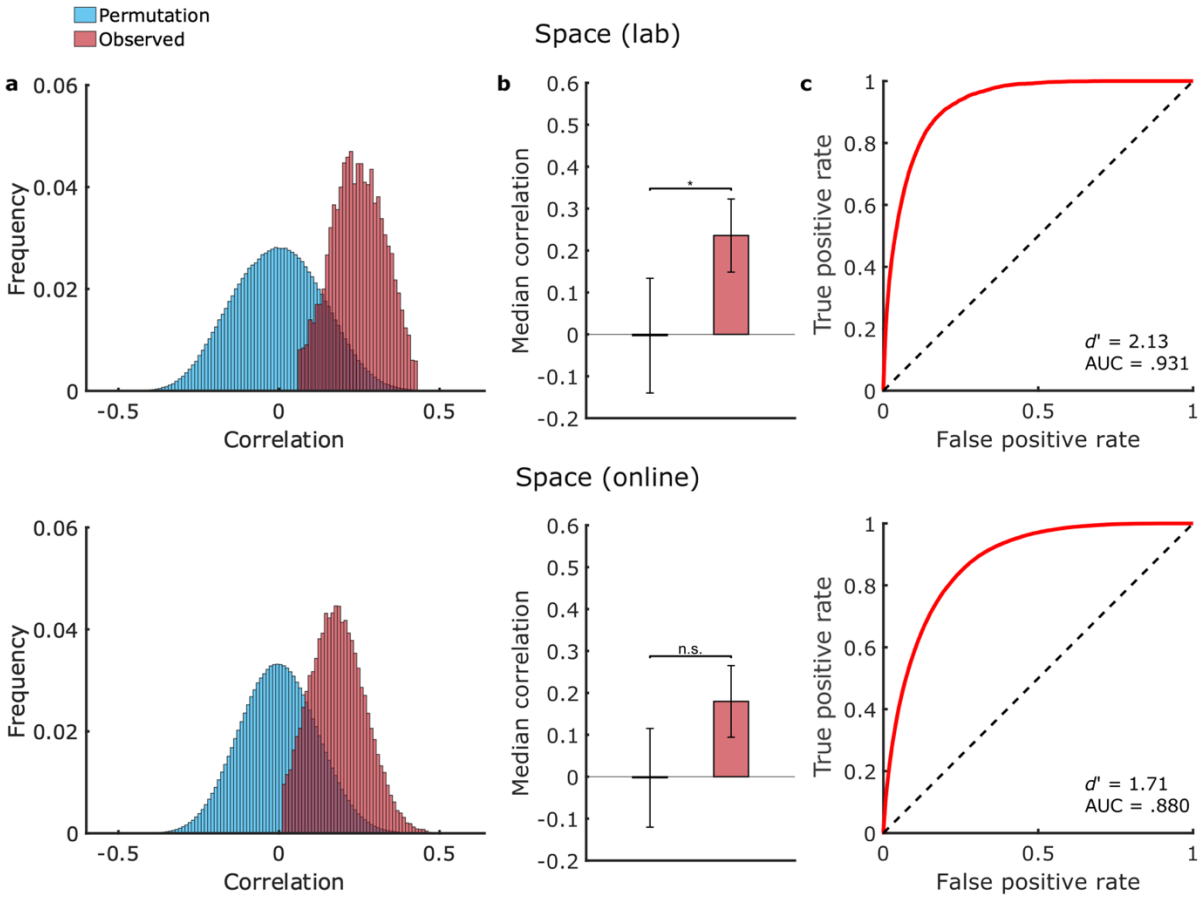

Figure 5 (continued). The same results are shown for "Space".

Table 3

Results of inter-supersubject correlation analysis.

Condition	Statistic	Range
	r	
Arctic/lab	$r = .423$, $p < .01$	.245-.596
Arctic/online	$r = .317$, $p = .017$	.071-.536
Space/lab	$r = .236$, $p = .031$	.032-.429
Space/online	$r = .180$, $p = 0.06$	.064-.422
	d’	
Arctic/lab	3.47	2.94-3.76
Arctic/online	2.51	2.10-2.99
Space/lab	2.13	1.60-2.76
Space/online	1.71	0.71-2.25
	AUC	
Arctic/lab	.99	.98-1.00
Arctic/online	.96	.93-.98
Space/lab	.93	.87-.97
Space/online	.88	.69-.94

Note. “lab”, lab dataset; “online”, online replication dataset.

3.3 Executive load time course variance analysis

The significance of each canonical executive load time course was tested against the null time course (zero). The executive load time course for each story (red line), along with the null time course (blue line) and labels of all sections of the story where observed RT differed from zero, is shown in Fig. 6a (lab dataset) and Fig. 6b (online dataset). Fig. 7 shows the median inter-supersubject correlation for the significant and nonsignificant time courses. These were obtained by extracting and concatenating only the significant or nonsignificant sections of each supersubject time course, and independently computing the median correlation for both new sets of time courses. For both stories, the median correlation for the significant time courses was significant ("Arctic": $r = .862$, $p < .01$; "Space": $r = .774$, $p < .01$); the median correlation for the nonsignificant sections was not. This indicates that the significant sections of the executive load time courses drive the observed inter-supersubject correlations, and the nonsignificant sections do not meaningfully contribute to this correlation. A one-tailed z test revealed that the median ISC for significant time courses was greater than that for nonsignificant time courses, both for "Arctic", $z = 6.79$, $p < .001$, and "Space", $z = 5.25$, $p < .001$. A qualitative analysis was conducted to relate story characteristics to the significant sections of the executive load time courses. Collected transcripts for this analysis are summarized in Table 3.

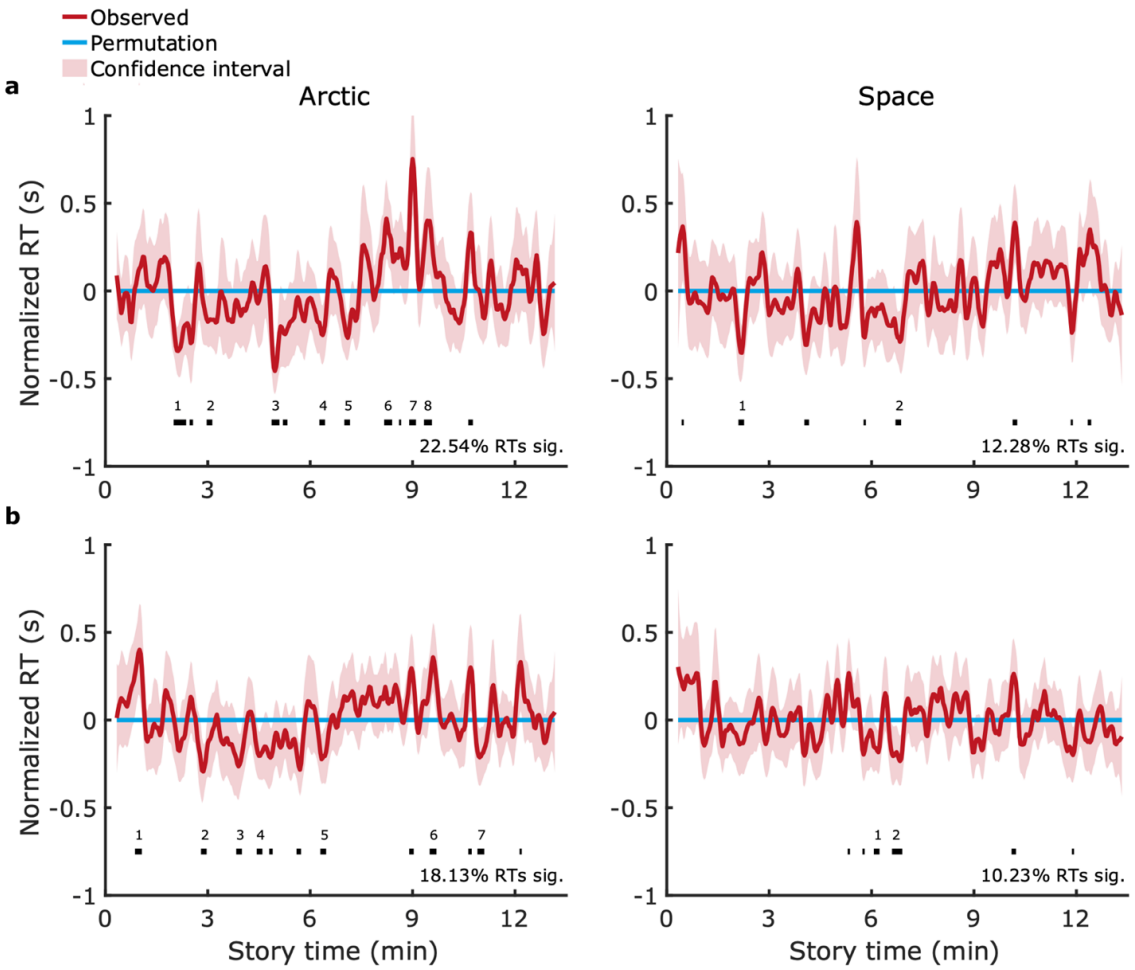

Figure 6. Executive load time courses differ from null. Standardized canonical executive load time courses differ from null (zero) time courses. *a*: lab experiment; *b*: online experiment. Column 1: "Arctic"; Column 2: "Space". The null time course (permutation; blue) is a line at zero. The pink shaded regions surrounding the observed (case judgement; red) time courses outline the 95% confidence interval on the mean. Wherever the confidence intervals do not overlap with the null time course (indicating that RT is either greater than or less than zero), RTs significantly differed ($p < .05$) from the null. Percentage of significant RTs is indicated in the bottom right corner of each graph. Black lines below the time courses underline all significant RTs. Significant sections exceeding 18 seconds are numbered and correspond to sections shown in column 1 of Table 3.

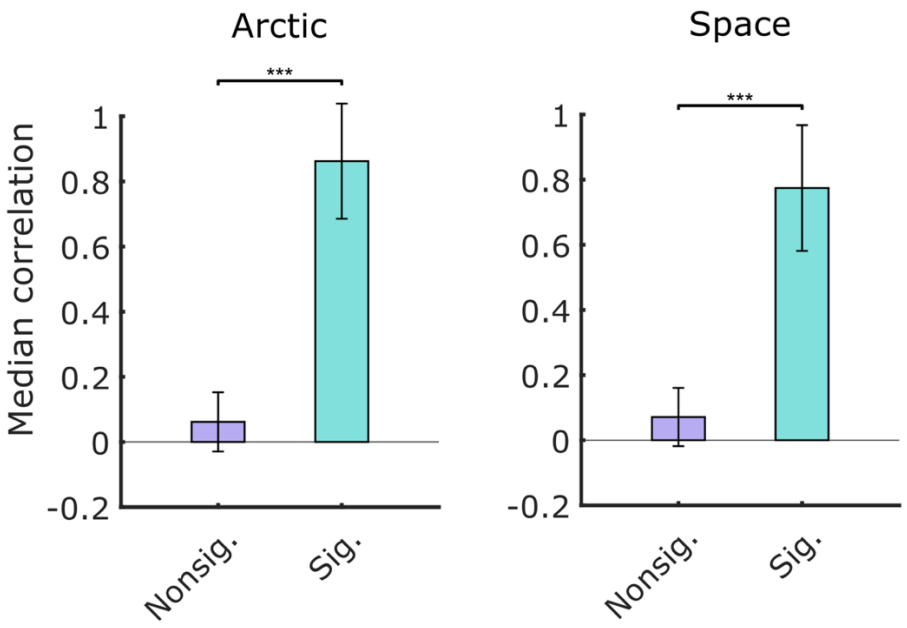

Figure 7. Nonsignificant sections of time courses do not contribute to ISC. Magnitude of inter-supersubject correlation between executive load time courses depends on the significance of the time courses. Median Pearson's correlation coefficient is shown for time courses consisting of only nonsignificant RTs (purple) and time courses consisting of only significant RTs (turquoise). Results are shown for the lab dataset, for "Arctic" (left) and "Space" (right). Error bars represent the standard deviation of the median correlation, computed as the square root of the median observed correlation distribution variance across all sets. Braces over bars show the significance of the median correlation based on one-tailed permutation testing. Asterisks indicate significant differences (n.s., not significant; *$p < .05$; **$p < 0.01$; ***$p < 0.001$).

Table 3: *Story transcripts for periods of significant case-judgement RTs.*

Section (lab; online)	Sign of RTs over section (lab; online)	Story time (mm:ss)	Transcription of speech from section
		Arctic	
2; 2	- ; -	02:53-03:03	"…and to be fair to Robert, he did work pretty hard. So off we went on our first day, and our first day was absolutely spectacular…"
4; 5	- ; -	06:13-06:23	"…I did that on my hands and knees. Well, of course it wasn't all bad weather. We had, just, so many spectacular days…"
8; 6	+ ; +	09:25-09:39	"…she banged right into him, and sent him tumbling to the bottom of the ravine, taking the whole team with him. She turned around and hightailed it back to her cub and when the dogs saw her running away, they tried to chase after her. But they couldn't quite get to her…"
		Space	
2; 2	- ; -	06:37-06:47	"…and I found out how the Navy was gonna get us ready to do this. You don't jump out of the plane the first day – what you do is you take it step-by-step, and they build you up, inch-by-inch…"

Note. Sections are based on the numbered black bars shown in fig. 6, which underline sections of each executive load time course during which RT was significant for longer than 18 seconds. Only segments of these sections that occurred at the same time (relative to the story) for the lab and online dataset are shown. Story time is given in mm:ss (minute:second) format.

3.4 Assessing relationship between ISC of supersubjects & engagement

To examine whether objectively measured correlations related to subjectively rated absorption and enjoyment, each supersubject was assembled according to their ratings on each of these scales. In other words, we identified the participant who, in each version, had the highest absorption rating – these six individuals were then assembled into a supersubject. Data from the six subjects (one per version) with the second highest absorption ratings was then combined, and so on. We then analyzed the relationship between each supersubject's average absorption rating and the correlation of that supersubject time course with the canonical executive load time course. This analysis was repeated for enjoyment ratings. The one-tailed ($r > 0$) Spearman correlation between enjoyment and time-course correlation was $r_{(94)} = .304$, $p = .018$, and for absorption was $r_{(94)} = .048$, p = .372; see Fig. 8. The significant relation between enjoyment and correlation indicates that the higher self-reported absorption in the story, the more reliably the story drove the executive load time course.

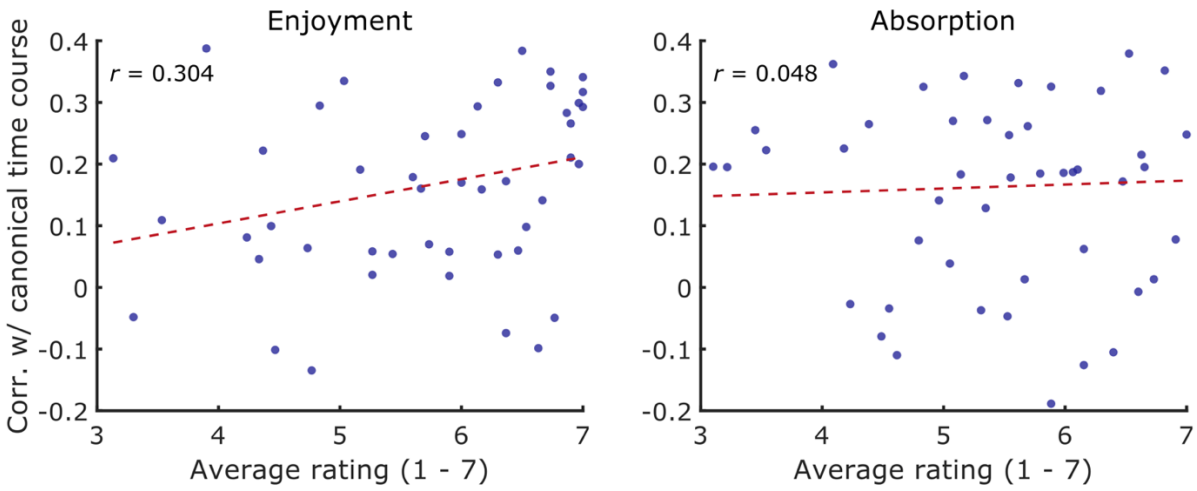

Figure 8. Inter-supersubject correlation is related to story enjoyment, but not absorption. Left: enjoyment; Right: absorption. The Pearson's correlation coefficient between each supersubject time course and the corresponding canonical executive load time course is shown as a function of each supersubject's average rating of enjoyment or absorption. Individual supersubject values are shown as blue circles; the line of best fit is shown as a red dashed line. The Spearman's rank correlation coefficient between both variables is shown in the top left corner of the graphs (a significant positive correlation, $p = .018$, was found for enjoyment). Data are pooled for both stories ("Arctic" and "Space") and both datasets (lab and online).

3.5 Temporal smoothing window size determination and replication analysis

To identify the smoothing window size best matched to the intrinsic dynamics of executive load during story listening, we temporally smoothed supersubject time courses with window sizes ranging from 0 to 100 samples. Fig. 9 shows correlation as a function of window size for both stories and datasets. As the Pearson correlation between time courses tends to rise as the temporal smoothing window applied to the time courses increases, we were not able to identify a clear knee point from this analysis (i.e., global maximum) corresponding to the temporal dynamic window of executive load.

A follow-up analysis was conducted to identify this window based on correlations between the canonical executive load time course for the lab and online datasets, across a

range of temporal smoothing window sizes. The results of this analysis are shown in Fig. 10. Fig. 10a shows congruent correlations (same story time course correlation between datasets (IL-Space, OL-Space; turquoise) as a function of window size for both stories, as well as the average incongruent correlation (across both stories: IL-Arctic/Space, OL-Space/Arctic; OL-Arctic/Space, IL-Space/Arctic; purple). The difference between the congruent and average incongruent correlation was computed using a *z* test for every window size. Fig. 10b shows the *z* score for these tests as a function of window size. This value served as a measure of the agreement between the lab and online datasets. When the results for both stories were combined, the window size corresponding to the maximum *z* value (i.e., the window size that maximized the agreement between the datasets) was 15 samples (30 sec of story time). We interpreted this window size as the one best matched to the story-driven signal present in the executive load time courses. However, reasoning that the temporal window of executive load fluctuations may differ for different stories, we analyzed each story separately. Through qualitative assessment, we concluded that for each individual story, the entire region of the *z* score curve between ~10 and 20 samples could be reasonably considered part of the 'knee point' that we sought to identify through this analysis. We therefore opted to use a more conservative window size (10 samples) to smooth supersubject time courses for all analyses.

We then computed the correlation between canonical executive load time courses, temporally smoothed using a window size of 10 samples, to quantify the agreement between the lab and online datasets. The Pearson correlation coefficient between the canonical "Arctic" time course for the lab (IL-Arctic) and online (OL-Arctic) dataset was $r = .525$; the same correlation for "Space" (IL-Space, OL-Space) was $r = .385$. The average incongruent correlation (IL-Arctic, OL-Space; OL-Arctic, IL-Space) was $r = .042$. One-tailed *z* tests comparing each story's congruent correlation to the average incongruent correlation between both stories (which served as an approximation of the null hypothesis that the lab and online results for a given story are not correlated) were both significant: "Arctic", $z = 2.96$, $p < .01$; "Space", $z = 1.99$, $p = .023$. This indicates that the executive time courses obtained online were similar to those obtained in person.

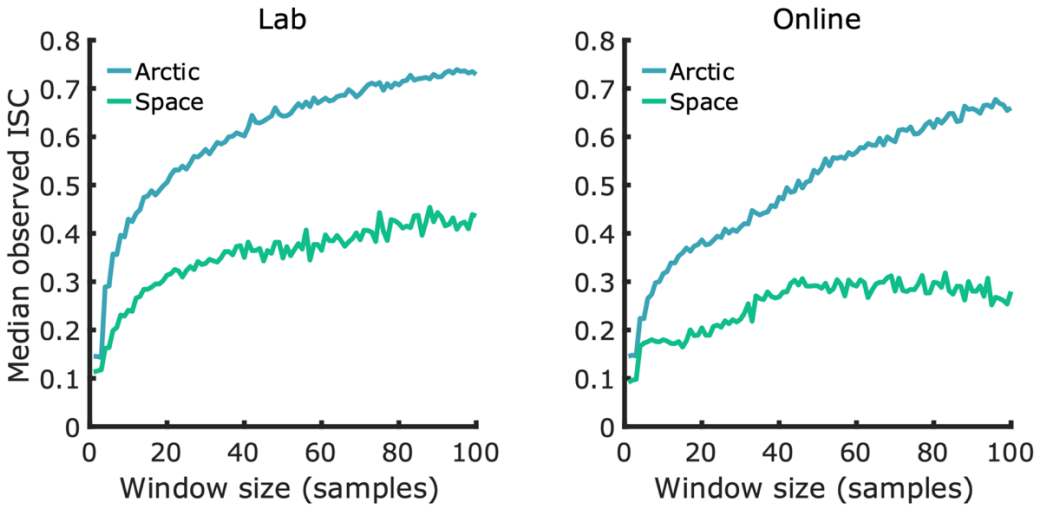

Figure 9. Identifying the dynamic window of executive load: within-experiment. Identifying the temporal window of executive load dynamics during story listening: within-dataset ISC analysis. Left: lab dataset; right: online dataset. Both figures show median inter-supersubject correlation for “Arctic” (blue) and “Space” (green) as a function of temporal smoothing window size (2 sec/sample).

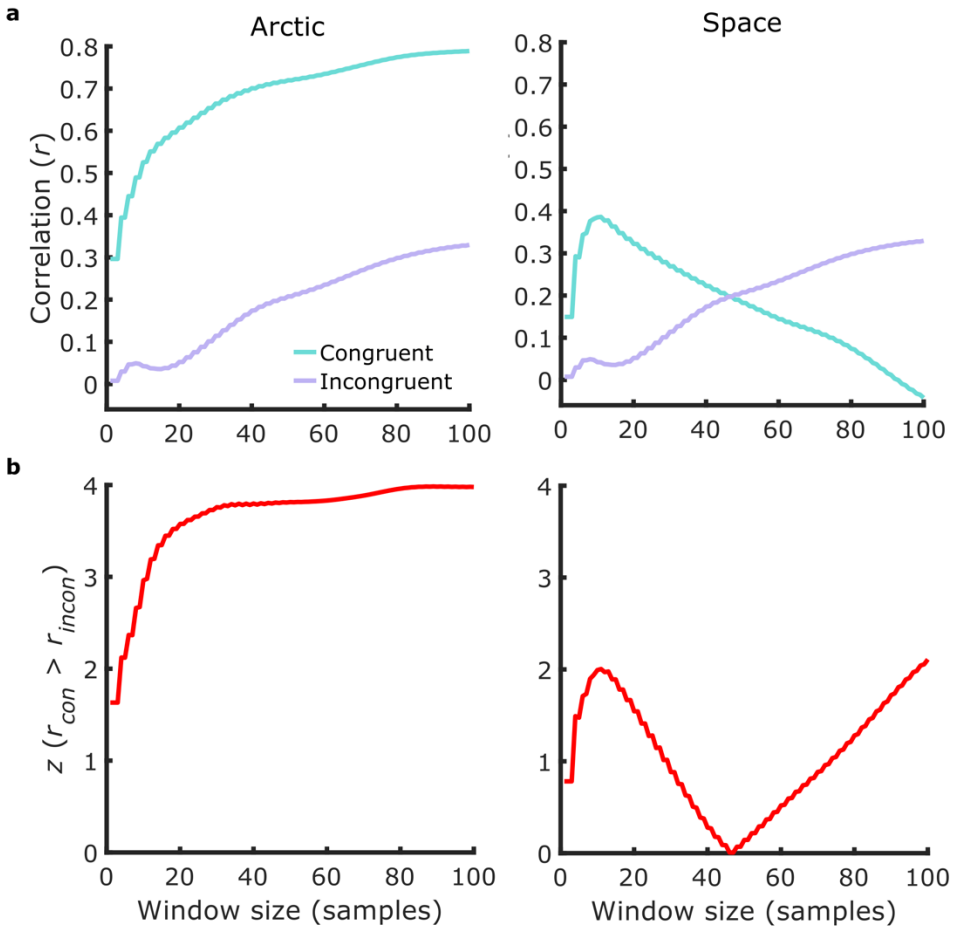

Figure 10. Identifying the dynamic window of executive load: between-experiment. Identifying the temporal window of executive load dynamics during story listening: between-experiment replication analysis. Column 1: “Arctic”; column 2: “Space”. *a*, Congruent correlation (turquoise) between canonical executive load time courses for indicated story(ies) and average incongruent correlation (purple) as a function of smoothing window size. *b*, *z* score for each one-tailed *z* test of the difference between congruent and average incongruent correlation (r_{con}, congruent > r_{incon}, incongruent) across the range of smoothing window sizes.

Chapter 4

4 Discussion

In this study we investigated how listeners engage with acoustically clear, spoken stories. Our overarching objective was to develop methodological tools that could be used to evaluate such natural, ecologically valid stimuli in future investigations of listening effort. Addressing our first aim, we used a dual task to characterize the dynamics of cognitive recruitment associated with the executive demands of two stories. To determine whether cognitive recruitment during story listening is consistent across individuals, we computed the correlation between executive load time courses across supersubjects. For both stories, this analysis revealed consistent dynamics of executive load during story listening. We then performed an exploratory analysis to investigate the factors that drove the observed consistency in executive load, identifying significant peaks throughout each story that explain the observed consistency and may be related to aspects of the narrative. Under the assumption that engagement is a prerequisite for listening effort and the corollary that any materials used to measure listening effort should elicit engagement, we quantified the degree to which listeners engaged with the stories. Addressing our second aim, to investigate whether ISC of behavioural responses depends on engagement, we used participant ratings of engagement to examine whether they correlated with ISC. This analysis revealed that story enjoyment, but not absorption, was significantly related to consistency of executive load time courses.

4.1 Case-judgement results are consistent with past research

The results of our case-judgement task reflect past research. Case-judgement RT was consistently greater, for both story stimuli and both the lab and online experiments, for lower case than upper case letters. Rodd et al. (2010) used a similar dual-task paradigm to the one used here, in which participants listened to sentences instead of stories, and found the same result. This finding is probably a result of participants, most of whom were right-handed, using the index finger of their dominant hand to respond to upper case letters (Rodd et al., 2010).

The case-judgement RTs we collected were, for both the lab and online experiments, longer on average than those found by Rodd et al. (2010). This can be explained by the long average time interval (~6-14 sec) between case-judgement trials. Unlike during sentence listening (for which participants were able to anticipate a case-judgement trial within every, on average, five second long sentence; Rodd et al., 2010), participants listening to the stories in the present study were less able to anticipate the appearance of trials, and as a result less prepared to respond, resulting in longer RTs.

4.2 Dynamics of cognitive recruitment during story listening are consistent across individuals

Our first aim was to measure the cognitive demands associated with each story. We did so using a dual task, in which listeners attended to stories while simultaneously performing an intermittent case-judgement task (Gagné et al., 2017). To characterize the reliability of case-judgement performance, we computed the correlation between RT time courses across individuals. Using past ISC analyses of brain activity as a foundation (Dmochowski et al., 2012; Hasson et al., 2004), we developed an analogous method for computing ISC between time courses of behavioural responses. For both story stimuli, we observed a significant correlation between case-judgement RT time courses across supersubjects. As the stories were the only factor common to all participants that also varied dynamically, we attribute the effect of the observed consistency to the stories. Furthermore, assuming that ISC of behavioural performance (i.e., 'behavioural ISC'), like that of neural activity (i.e., 'neural ISC'), depends on engagement, we predicted that consistent case-judgement performance across individuals depends on consistent engagement during story listening. We carried out further analyses, described below, to investigate this prediction.

As the demands of case judgement do not vary to a large degree depending on the specific letter presented, fluctuations in RT on the case-judgement task can be accounted for by fluctuations in the demands associated with listening to and comprehending the story. This implies that both tasks impose a load on overlapping brain networks and 'compete' for cognitive resources. We therefore interpret RTs as indexing the recruitment of cognitive processes required for both the case-judgement task (e.g., response selection;

Rodd et al., 2010) and the integrative demands of narrative comprehension. The specific networks taxed might include the – predominantly frontal – cognitive control networks involved in meeting the demands of a wide range of tasks spanning several domains (Duncan, 2010; Duncan & Owen, 2000). These networks overlap with key language regions (e.g., LIFG; Fedorenko et al., 2012; Rodd et al., 2010) and frontoparietal regions involved in meeting the executive demands of narrative comprehension (Naci et al., 2014).

This interpretation is in line with previous studies using dual-task paradigms to index speech processing demands (Naci et al., 2014; Rodd et al., 2010). Rodd et al. (2010) found that participants took longer to perform a case-judgement task while listening to sentences that contained homophones, relative to matched sentences that did not contain homophones. The researchers interpreted the observed increase in case-judgement RT as reflecting the recruitment of cognitive processes involved in resolving semantic ambiguities. In another related study, Naci et al. (2014) found that RTs on a go/no-go task during movie watching significantly predicted brain activity in frontoparietal (FP) regions associated with higher-order cognitive functions. The researchers interpreted this as evidence that case-judgement RTs during movie watching capture the cognitive load imposed by *executive* aspects of the narrative (e.g., those related to character motivations, plot, etc.) – that is, RTs in this case captured *executive* load.

In the present study, the cognitive resources recruited for performing the case-judgement task are probably involved in meeting the executive demands of narrative comprehension (e.g., making predictions about the events in the story or thinking about its meaning), and to a lesser extent, in compensating for linguistic and talker demands (e.g., resolving semantic ambiguities or perceptual uncertainty related to an unfamiliar accent; Johnsrude & Rodd, 2016). Therefore, the term “cognitive load” best encompasses the host of cognitive processes involved in listening to and comprehending a narrative, but we use the term “executive load” to refer to those cognitive processes specifically involved in narrative comprehension. Furthermore, assuming that case-judgement RT corresponds to the recruitment and depletion of cognitive resources in response to a load

(Kahneman, 1973; Gagné et al., 2017; Rodd et al., 2010), we consider "cognitive recruitment" the most empirical term for the processes indexed by case-judgement RTs.

4.3 Significant case-judgement response times drive inter-supersubject correlation

To investigate the factors that drove consistent executive load during story listening, we performed a model-free analysis that has been employed in past ISC analyses to relate brain activity to stimulus properties (Finn et al., 2018; U. Hasson et al., 2004; Kauttonen et al., 2015). In this analysis, the time points of significant peaks and troughs in brain activity are identified and used to guide an investigation of the stimulus elements that drove them. Our analysis of executive load time courses revealed sections of both stories during which case-judgement RTs significantly differed from the null time course. Analyzing ISC separately for supersubject time courses that consisted of only the significant or nonsignificant sections revealed that only significant sections contribute meaningfully to the observed ISC. This finding is in line with expectations and extends past neuroimaging research (U. Hasson et al., 2004) by demonstrating definitively that nonsignificant sections of an averaged time course – of, e.g., case-judgement RT or brain activity – do not contribute to ISC between them.

We examined transcripts of the stories to relate the identified significant sections of the time courses to the story elements that drove them. To increase our confidence in the importance of the identified sections, we limited this investigation to only those sections that were significant for both the lab and online experiment. A qualitative analysis of the transcripts revealed suggestive patterns (e.g., each positive section of the Arctic story includes mention of an encounter with a polar bear, whereas the negative sections do not), but without further quantitative analysis it is not possible to draw conclusions about which factors drive executive load. More research is necessary to develop principled methods for applying the reverse correlation approach for auditory stimuli. Future research should explore quantitative approaches to relating story elements to consistent fluctuations in behavioural responses (Kauttonen et al., 2015).

4.4 Inter-supersubject correlation is related to story enjoyment

Our second aim was to determine whether behavioural ISC, like analogous neuroimaging analyses, depends on degree of engagement during story listening (Dmochowski et al., 2012, 2014; Poulsen et al., 2017; Schmälzle et al., 2015). Our analysis revealed that a supersubject's correlation with the canonical executive load time course is significantly correlated with their ratings of enjoyment, but not their ratings of absorption. In other words, the listeners who enjoyed the story had more consistent case-judgement task performance than listeners who did not enjoy the story. This finding suggests that story enjoyment – a facet of engagement, as defined by the narrative engagement questionnaire (Busselle & Bilandzic, 2009) – modulates the consistency with which a story drives executive load.

Despite the significance of the correlation between enjoyment ratings and time course correlation, our resolution for observing this effect may have been limited. We selected stories for this experiment on the basis that we considered them entertaining, well-told, and appealing to a broad group of individuals. Although participants typically preferred one story over the other, most reported that they found both stories engaging. In fact, the minimum rating of both enjoyment and absorption across supersubjects was greater than three, although ratings were collected using a Likert scale from 1 to 7. This result indicates that even the participants who gave the lowest ratings of absorption and enjoyment only somewhat disagreed with the statements used to assess them. If these participants were, in fact, somewhat engaged in the stories, our sensitivity to the relationship between engagement and ISC would be limited. It is possible that comparing ISC for the stories used here to less engaging control materials (e.g., randomly ordered sentences) would reveal significant differences across all subscales of narrative engagement. Additional investigation into the nature of engagement and how best to assess it will also help inform future research.

According to the conceptual framework we established at the outset of this paper, engagement with a narrative is updated moment-to-moment, and in a challenging listening environment it determines the magnitude and duration of effortful listening that

a listener will endure before disengaging from the narrative. I further theorize that engagement with a story modulates the consistency with which the demands of the story impinge on brain activity across listeners. If a group of individuals who each hear a story are engaged, the story will impose a consistent load on their brain activity and the demands associated with the story will be reflected in their brain activity and behaviour, resulting in high ISC measurements.

This theory rests on two assumptions: first, we must assume that each story is associated with a unique time course of demands – acoustic, linguistic, executive, etc. Some demands are subjective (e.g., accent familiarity, word familiarity); others are objective (e.g., signal quality). Individual processing of all demands varies depending on the individual's unique cognitive abilities or neural organization. Despite this subjective aspect of the demands of a narrative, for a given sample of participants, the assumption holds: there is a unique average time course of demands associated with every narrative. The second assumption is that in a group of highly engaged individuals listening to a narrative, the demands of the narrative are represented in the activity of each of their brains with high fidelity. In other words, all information about the demands of a stimulus is encoded in the neural activity of all individuals. In reality, an individual's attention to a narrative waxes and wanes over the time course of the narrative. Their neural representation of the narrative is, as a result, mixed with individual noise. Intersubject correlation therefore decreases as the consistency of individuals' attention to a narrative over its time course decreases. There may be additional differences in neural responses to a narrative when listeners are intrinsically motivated to listen to it, as opposed to when they are extrinsically motivated to listen.

4.5 Dynamic window of executive load inconclusive

To compensate for experimental noise and obtain representative time courses of executive load, we applied temporal smoothing at the supersubject level. However, since we did not know the time scale over which executive load fluctuates, we analyzed the effect of a range of temporal smoothing window sizes on time-course consistency – both within and between independently collected datasets – to determine the optimal window for capturing such fluctuations. This exploratory analysis yielded inconclusive results for

both stories. Our analysis of how ISC is affected by temporal smoothing window size revealed no clear knee point where ISC is maximized. As expected, the correlation between temporally smoothed time courses tends to increase with the window size, approaching an asymptote at a value of one as the time courses become increasingly flat. Our analysis of the agreement between the canonical executive load time course for the lab and online datasets was also inconclusive. The difference between the average congruent correlation and the average incongruent correlation revealed a clear knee point (in the form of a global maximum) at a window size of fifteen samples. However, this maximum was revealed to be the result of averaging over the congruent correlation for both stories, each of which showed different trends. The congruent correlation for "Arctic" increased with increasing window size, approaching an asymptote at a value of one, but beginning to plateau around a window size of fifteen samples. In contrast, the congruent correlation for "Space" increased up to a point and then decreased until the lab and online time courses started to become anticorrelated. A follow-up analysis revealed that this was due to temporal smearing as a result of smoothing, which began to shift the executive load time course for the experiments out of phase beyond the observed turning point. The result is that when the congruent correlations for both stories are combined, a maximum is seen where the correlation for "Space" begins declining.

The best method for selecting the optimal temporal smoothing window size is still unclear. However, the results of this exploratory analysis are informative. Although the individual congruent correlation for each story did not show a global maximum, both correlations did show a clear inflection point at a window size of about fifteen samples. Correlation may not be an ideal dependent measure for identifying the temporal smoothing window best matched to the intrinsic dynamics of executive load, but it does seem to produce meaningful results. Future research should explore other approaches to optimizing the temporal smoothing window size for continuous stimuli.

4.6 Results of in-lab and online experiments are consistent

The results of the experiments conducted in the lab and online were highly similar. Dual-task performance, engagement ratings, and inter-supersubject correlations for each story followed similar patterns. The canonical executive load time course for

each story was highly consistent between the experiments. The consistency of online results with those collected in the lab suggests that online experimentation may be viable for carrying out behavioural experiments in which temporal precision is critical. Given the large sample sizes the supersubject method requires, until it is refined to be more efficient, online experimentation (for which many participants can often be run with little effort or time required) may be a good alternative to in-lab experimentation for future research using behavioural ISC.

4.7 Limitations

Additional sources of noise

As RT differs systematically between lower- and upper-case letters, failing to filter any analysis of case-judgement responses by case may add noise to the data. However, the supersubject sampling method makes it challenging to filter the data in this manner. Because each participant received a pseudorandom distribution of upper- and lower-case letters (half each), the lower- and upper-case letters across participants in a given story/version condition do not necessarily align in time. Moreover, even if trials for each case were aligned in time, filtering subsequent analyses by case would result in supersubject time courses with RTs unevenly distributed in time (as opposed to evenly distributed, at a 0.5 Hz resolution). However, as letter case did not vary systematically with narrative demands, not filtering the analysis by case does not confound the results – it is merely a source of noise. Future experiments using the supersubject method could overcome this limitation by extending the supersubject method to constrain lower- and upper-case letters to distributions of pseudorandomly distributed time points within groups of participants.

Interpreting inter-supersubject correlation distributions

After smoothing supersubject time courses, we computed their consistency using a split-half method intended to further increase their SNR. It should be noted that this approach – which consists of splitting the to-be-correlated time courses into every unique combination of two halves and correlating the averaged halves – yields a distribution of

correlation values that might underestimate the true variance (Chen et al., 2016). Because many splits have similar configurations of supersubjects, the values that make up the correlation distribution are not all distinct. For example, one split might differ from another by the position of only two supersubjects that were swapped between split halves. This property of correlation distributions is inherent to most correlation methods, including pairwise and leave-one-out methods. As a result, differences between observed and permuted correlation distributions may be inflated. However, measures of central tendency, effect size, and range are unaffected by this limitation (Chen et al., 2016). Researchers computing ISC should be aware that the variance of observed correlations may underestimate the true variance of the correlation distribution. Instead, they can rely on the range as a statistically valid measure of dispersion.

Supersubject sampling: trade-off between trial predictability and resolution

In this study we developed a novel sampling method that enabled us to obtain time courses of executive load with a constant sampling rate while also ensuring that trials were presented to participants at unpredictable points in time. This method consists of generating a set of timing versions that dictate when during a story case-judgement trials will be presented (and corresponding RTs sampled). For each version, a unique distribution of time points is generated within constraints, such that collapsing these distributions across versions yields a single (supersubject) time course with a constant sampling rate. Supersubject time courses are then smoothed and correlated. Two parameters of this sampling method can be varied: the average time between samples and the number of experiment versions. We selected parameters that minimized individual trial predictability and maximized sampling resolution (i.e., high average duration between samples and high number of experiment versions). This ensures the representativeness and resolution of executive load measurements but requires a large sample size to obtain complete time courses. As a first step toward developing an ISC analysis for behavioural data, this trade-off was warranted.

Research should be conducted to analyze the trade-off between supersubject resolution and trial predictability. Behavioural methods are more cost effective than

imaging methods; optimizing supersubject sampling will make this advantage more appreciable. Single-subject measures of executive load are ideal. These could be obtained by identifying neural correlates of executive load and measuring them during story listening to obtain a high-resolution readout of load over time. One such method is described in the '*Parcellating brain networks that support listening under different conditions*' section below.

4.8 Future directions

Using behavioural ISC to assess narrative engagement at the group level

Insofar as behavioural ISC is shown to depend on story engagement, it will provide a useful tool for objectively evaluating the temporal dynamics of engagement during story listening. At the group level, behavioural ISC has applications both within and between groups. For example, in a group of people with normal hearing it could be used to measure narrative engagement, as a means of quantifying different materials' capacity to elicit cognitive processes thought to require intrinsic motivation, such as listening effort (Herrmann & Johnsrude, 2020; Peelle, 2018; Pichora-Fuller et al., 2016; Richter, 2013). Using the methods we outline here for comparing inter-supersubject correlations or supersubject time courses, future research could also investigate how the dynamics of narrative engagement differ between special and neurotypical populations. Past research has found that individuals with impaired social cognition (e.g., due to Down syndrome (Anderson et al., 2013) or autism (Hasson et al., 2009; Salmi et al., 2013) show widespread reduction in neural ISC relative to neurotypical controls while viewing a movie, especially in brain regions involved in processing social information (e.g., default mode network; Salmi et al., 2013). These individuals may also show less consistent performance on behavioural measures that index social cognition. Additional populations of interest include individuals with aphantasia (i.e., the inability to produce visual mental imagery) or other disorders that might impair or alter narrative processing. Relative to neurotypical controls, these groups might show reduced ISC in response to a narrative. Alternatively, within-group ISC might be high for both groups, but the temporal

dynamics of narrative processing, evidenced by executive load time courses or frontal brain activity, might differ between them.

Using behavioural ISC to identify abnormal mental states within individuals

If the supersubject sampling method can be refined to work at the single-subject level, ISC between individuals and healthy controls might be used to detect abnormal mental states within individuals. In a previous experiment, an absence of consistency in FP brain activity during movie watching between a behaviourally nonresponsive individual and a healthy group provided evidence that the behaviourally nonresponsive individual was not conscious (Naci et al., 2014). A similar method using behavioural ISC could be developed to detect abnormalities related to narrative engagement. For example, an older individual who finds listening in background noise to be unusually challenging might show divergent executive load dynamics when compared with a group with normal hearing. This divergence might reflect abnormal narrative processing as a consequence of abnormal speech processing. In addition, time resolved behavioural ISC could be developed to detect the point at which an individual disengages from a narrative (Simony et al., 2016). This might shed light on how listening effort unfolds over time in real-world listening situations.

Parcellating brain networks that support listening under different conditions

I plan to use the materials and methods developed in this study to investigate how executive load dynamics relate to neural activity. Neural measures provide an objective window into the complex and heterogeneous underpinnings of speech comprehension. One method of connecting behaviour to brain activity is to use the executive load time course associated with a given story to model the brain activity of individuals who listen to that story. In a previous experiment, brain activity in the FP network during movie watching was strongly predicted by go/no-go RTs during movie watching, which were interpreted as a readout of the executive demands of the movie (Naci et al., 2014). Regressing executive load time courses onto brain activity might uncover the networks sensitive to this load. Future investigations could then obtain executive load time courses

at the single-subject level by indexing brain activity within these networks during story listening. Additionally, networks sensitive to the integrative demands of narrative comprehension could then be differentiated from other networks involved in narrative comprehension under challenging listening conditions, such as when listening is made effortful due to added background noise (Duncan, 2010; Peelle, 2018; Peelle & Wingfield, 2016). In doing so, we might begin to dissociate the multiple, concurrent processes that are active during listening and rely on different brain networks. This method presents a promising avenue for connecting behaviour and brain activity and enriching our understanding of the neural architecture of narrative comprehension and listening effort.

4.9 Conclusion: Advancing natural assessment of listening effort

In this study we have developed methodological tools that will push more natural assessment of listening forward. Existing measures of listening effort do not satisfactorily capture the real-world listening challenges that patients – even those with normal audiometric thresholds – often report (Lesica, 2018; Parthasarathy et al., 2020; Pichora-Fuller et al., 2016; Ruggles et al., 2011). The use of narrative stimuli has several potential advantages over the isolated sentences currently used in standard speech in noise testing. Namely, stories are ecologically valid and have the potential to intrinsically motivate listening (Bamberg, 2011; Brown, 2004; Dunlop & Walker, 2013; Picou et al., 2014; Smith et al., 2017; Wilson et al., 2007). In this study I describe a method for computing ISC between behavioural response time courses, opening a new avenue for assessing behavioural responses to natural, continuous stimulation. Inter-supersubject consistency in case-judgement responses indicates that these measurements capture meaningful information about the demands of a narrative and might reflect the extent to which listeners engage with the narrative. This information is vital for understanding how people listen to and engage with stories. Behavioural ISC has many applications, including, for example, quantifying the utility of different stories for assessing listening effort. Ecologically valid approaches to assessing listening effort, to which the methods developed here open the door, are critical.

References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(sup2), S53–S71.

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2018). Hearing handicap and speech recognition correlate with self-reported listening effort and fatigue: *Ear and Hearing*, *39*(3), 470–474.

Anderson, J. S., Nielsen, J. A., Ferguson, M. A., Burback, M. C., Cox, E. T., Dai, L., Gerig, G., Edgin, J. O., & Korenberg, J. R. (2013). Abnormal brain synchrony in Down Syndrome. *NeuroImage: Clinical*, *2*, 703–715.

Bamberg, M. (2011). Who am I? Narration and its contribution to self and identity. *Theory & Psychology*, *21*(1), 3–24.

Bao, J., & Ohlemiller, K. K. (2010). Age-related loss of spiral ganglion neurons. *Hearing Research*, *264*(1–2), 93–97.

Beck, D. L., Danhauer, J. L., Abrams, H. B., Atcherson, S. R., Brown, K., Chasin, M., Clark, J. G., Placido, C. D., Edwards, B., Fabry, D. A., Flexer, C., Fligor, B., Frazer, G., Galster, J. A., Gifford, L., Johnson, C. E., Madell, J., Moore, D. R., Roeser, R. J., … Wolfe, J. (2018). Audiologic considerations for people with normal hearing sensitivity yet hearing difficulty and/or speech-in-noise problems. *Hearing Review, 25*(10), 28–38.

Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., & Corona-Strauss, F. I. (2013). Neural correlates of listening effort related factors: Influence of age and hearing impairment. *Brain Research Bulletin*, *91*, 21–30.

Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., & Corona-Strauss, F. I. (2017). Neurodynamic evaluation of hearing aid features using EEG correlates of listening effort. *Cognitive Neurodynamics*, *11*(3), 203–215.

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015). Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience*, *35*(5), 2161–2172.

Brown, D. E. (2004). Human universals, human nature & human culture. *Daedalus*, *133*(4), 47–54.

Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology*, *12*(4), 321–347.

Chang, W.-T., Jääskeläinen, I. P., Belliveau, J. W., Huang, S., Hung, A.-Y., Rossi, S., & Ahveninen, J. (2015). Combined MEG and EEG show reliable patterns of electromagnetic brain activity during natural viewing. *NeuroImage*, *114*, 49–56.

Chen, G., Shin, Y.-W., Taylor, P. A., Glen, D. R., Reynolds, R. C., Israel, R. B., & Cox, R. W. (2016). Untangling the relatedness among correlations, part I: Nonparametric approaches to inter-subject correlation analysis at the group level. *NeuroImage*, *142*, 248–259.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, *16*(9), 1348–1355.

Davis, M. H., Ford, M. A., Kherif, F., & Johnsrude, I. S. (2011). Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI. *Journal of Cognitive Neuroscience*, *23*(12), 3914–3932.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147.

de la Mothe, L. A., Blumell, S., Kajikawa, Y., & Hackett, T. A. (2012). Cortical connections of auditory cortex in marmoset monkeys: Lateral belt and parabelt regions. *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, *295*(5), 800–821.

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications*, *5*(1), 4567.

Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention – A possible marker of engagement? *Frontiers in Human Neuroscience*, *6*.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483.

Dunlop, W. L., & Walker, L. J. (2013). The life story: Its development and relation to narration and personal identity. *International Journal of Behavioral Development*, *37*(3), 235–247.

Feder, K. (2015). Prevalence of hearing loss among Canadians aged 20 to 79: Audiometric results from the 2012/2013 Canadian Health Measures Survey. *Health Reports*, *26*(7), 10.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within Broca's Area. *Current Biology*, *22*(21), 2059–2062.

Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A., & Constable, R. T. (2018). Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nature Communications*, *9*(1), 2043.

Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, *21*,

Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *International Journal of Audiology*, *43*(2), 85–99.

Goderie, T. P. M., Stam, M., Lissenberg-Witte, B. I., Merkus, P., Lemke, U., Smits, C., & Kramer, S. E. (2020). 10-Year follow-up results of the Netherlands Longitudinal Study on Hearing: Trends of longitudinal change in speech recognition in noise. *Ear and Hearing, 41*(3), 9.

Gratton, C., Sun, H., & Petersen, S. E. (2018). Control networks and hubs. *Psychophysiology*, *55*(3), e13032.

Gratton, M. A., & Vázquez, A. E. (2003). Age-related hearing loss: Current research. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *11*(5), 367–371.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539–2550.

Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., & Behrmann, M. (2009). Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. *Autism Research*, *2*(4), 220–231.

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. *Projections*, *2*(1), 1–26.

Helfer, K. S., Merchant, G. R., & Wasiuk, P. A. (2017). Age-related changes in objective and subjective speech perception in complex listening environments. *Journal of Speech, Language, and Hearing Research*, *60*(10), 3009–3018.

Herrmann, B., & Johnsrude, I. S. (2020). A Model of Listening Engagement (MoLE). 25.

Herrmann, B., Maess, B., & Johnsrude, I. S. (2018). Aging affects adaptation to sound-level statistics in human auditory cortex. *The Journal of Neuroscience*, *38*(8), 1989–1999.

Hervais-Adelman, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes*, *27*(7–8), 1145–1166.

Holmes, E., Folkeard, P., Johnsrude, I. S., & Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *International Journal of Audiology*, *57*(7), 483–492.

Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, *34*(5), 523–534.

Jacobson, L. A. (1989). A matched filter data smoothing algorithm. *IEEE Transactions on Nuclear Science*, *36*(1), 1227–1231.

Johnsrude, I. S., & Rodd, J. M. (2016). Factors that increase processing demands when listening to speech. In *Neurobiology of Language* (pp. 491–502). Elsevier.

Kauttonen, J., Hlushchuk, Y., & Tikka, P. (2015). Optimizing methods for linking cinematic features to fMRI data. *NeuroImage*, *110*, 136–148.

Ki, J. J., Kelly, S. P., & Parra, L. C. (2016). Attention Strongly modulates reliability of neural responses to naturalistic narrative stimuli. *Journal of Neuroscience*, *36*(10), 3092–3101.

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*(2), 291–300.

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. *The Journal of the Acoustical Society of America*, *141*(6), 4680–4693.

Kujawa, S. G., & Liberman, M. C. (2015). Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss. *Hearing Research*, *330*, 191–199.

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., & Stefanovska, A. (2018). Surrogate data for hypothesis testing of physical systems. *Physics Reports*, *748*, 1–60.

Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: concepts and age-related considerations. *Ear and Hearing, 37*, 8.

Lesica, N. A. (2018). Why do hearing aids fail to restore normal auditory perception? *Trends in Neurosciences*, *41*(4), 174–185.

Liberman, M. C., Epstein, M. J., Cleveland, S. S., Wang, H., & Maison, S. F. (2016). Toward a differential diagnosis of hidden hearing loss in humans. *PLOS ONE*, *11*(9), e0162726.

Lin, F. R., Metter, E. J., O'Brien, R. J., Resnick, S. M., Zonderman, A. B., & Ferrucci, L. (2011). Hearing loss and incident dementia. *Archives of Neurology*, *68*(2).

Löhler, J., Cebulla, M., Shehata-Dieler, W., Volkenstein, S., Völter, C., & Walther, L. E. (2019). Hearing impairment in old age. *Deutsches Aerzteblatt Online*.

Lopez-Poveda, E. A., Johannesen, P. T., Pérez-González, P., Blanco, J. L., Kalluri, S., & Edwards, B. (2017). Predictors of hearing-aid outcomes. *Trends in Hearing*, *21*, 233121651773052.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, *103*(49), 18866–18869.

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear and Hearing*, *36*(1), 145–154.

Marsella, P., Scorpecci, A., Cartocci, G., Giannantonio, S., Maglione, A. G., Venuti, I., Brizi, A., & Babiloni, F. (2017). EEG activity as an objective measure of cognitive load during effortful listening: A study on pediatric subjects with bilateral, asymmetric sensorineural hearing loss. *International Journal of Pediatric Otorhinolaryngology*, *99*, 1–7.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper.' *International Journal of Audiology*, *53*(7), 433–445.

Moore, B. C. J. (2007). *Cochlear hearing loss: Physiological, psychological and technical issues* (2. ed). Wiley.

Nachtegaal, J., Smit, J. H., Smits, C., Bezemer, P. D., van Beek, J. H. M., Festen, J. M., & Kramer, S. E. (2009). The association between hearing status and psychosocial health before the age of 70 years: Results from an internet-based national survey on hearing. *Ear and Hearing*, *30*(3), 302–312.

Naci, L., Cusack, R., Anello, M., & Owen, A. M. (2014). A common neural code for similar conscious experiences in different individuals. *Proceedings of the National Academy of Sciences*, *111*(39), 14277–14282.

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, nsz037.

Öberg, M., Marcusson, J., Nägga, K., & Wressle, E. (2012). Hearing difficulties, uptake, and outcomes of hearing aids in people 85 years of age. *International Journal of Audiology*, *51*(2), 108–115.

O'Neill, E. R., Kreft, H. A., & Oxenham, A. J. (2019). Cognitive factors contribute to speech perception in cochlear-implant users and age-matched normal-hearing listeners under vocoded conditions. *The Journal of the Acoustical Society of America*, *146*(1), 195–210.

Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V., & Polley, D. B. (2020). Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing. *ELife*, *9*, e51419.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., & Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *The Journal of the Acoustical Society of America*, *72*(6), 1788–1803.

Paxton, J. L., Barch, D. M., Racine, C. A., & Braver, T. S. (2008). Cognitive control, goal maintenance, and prefrontal function in healthy aging. *Cerebral Cortex*, *18*(5), 1010–1028.

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214.

Peelle, J. E., & Wingfield, A. (2016). The Neural Consequences of Age-Related Hearing Loss. *Trends in Neurosciences*, *39*(7), 486–497.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing, 37*, 23.

Pichora-Fuller, M. K., & Souza, P. E. (2003). Effects of aging on auditory processing of speech. *International Journal of Audiology*, *42*(sup2), 11–16.

Picou, E. M., Aspell, E., & Ricketts, T. A. (2014). Potential benefits and limitations of three types of directional processing in hearing aids. *Ear and Hearing*, *35*(3), 339–352.

Plack, C. J., Barker, D., & Prendergast, G. (2014). Perceptual consequences of "hidden" hearing loss. *Trends in Hearing*, *18*, 233121651455062.

Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., & Hansen, L. K. (2017). EEG in the classroom: Synchronised neural recordings during video presentation. *Scientific Reports*, *7*(1), 43916.

Ramage-Morin, P. L. (2016). Hearing difficulties and feelings of social isolation among Canadians aged 45 or older. *Health Reports*, *27*(82), 12.

Richter, M. (2013). A closer look into the multi-layer structure of motivational intensity theory: The multi-layer structure of motivational intensity theory. *Social and Personality Psychology Compass*, *7*(1), 1–12.

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2012). Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. *Cerebral Cortex*, *22*(8), 1761–1773.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fmri studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269.

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2010). The role of domain-general frontal systems in language comprehension: Evidence from dual-task interference and semantic ambiguity. *Brain and Language*, *115*(3), 182–188.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology*, *23*(8), 577–589.

Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences*, *108*(37), 15516–15521.

Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. G. (2012). Why middle-aged listeners have trouble hearing in everyday settings. *Current Biology*, *22*(15), 1417–1422.

Salmi, J., Roine, U., Glerean, E., Lahnakoski, J., Nieminen-von Wendt, T., Tani, P., Leppämäki, S., Nummenmaa, L., Jääskeläinen, I. P., Carlson, S., Rintahaka, P., & Sams, M. (2013). The brains of high functioning autistic individuals do not synchronize with those of others. *NeuroImage: Clinical*, *3*, 489–497.

Salvi, R., Sun, W., Ding, D., Chen, G.-D., Lobarinas, E., Wang, J., Radziwon, K., & Auerbach, B. D. (2017). Inner hair cell loss disrupts hearing and cochlear function leading to sensory deprivation and enhanced central auditory gain. *Frontiers in Neuroscience*, *10*.

Schmälzle, R., Häcker, F. E. K., Honey, C. J., & Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience*, *10*(8), 1137–1143.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*(1), 12141.

Smith, D., Schlaepfer, P., Major, K., Dyble, M., Page, A. E., Thompson, J., Chaudhary, N., Salali, G. D., Mace, R., Astete, L., Ngales, M., Vinicius, L., & Migliano, A. B. (2017). Cooperation and the evolution of hunter-gatherer storytelling. *Nature Communications*, *8*(1), 1853.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Sommers, M., Tye-Murray, N., Barcroft, J., & Spehar, B. (2015). The effects of meaning-based auditory training on behavioral measures of perceptual effort in individuals with impaired hearing. *Seminars in Hearing*, *36*(04), 263–272.

Sridharan, D., Levitin, D. J., & Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, *105*(34), 12569–12574.

Tremblay, K. L., Pinto, A., Fischer, M. E., Klein, B. E. K., Klein, R., Levy, S., Tweed, T. S., & Cruickshanks, K. J. (2015). Self-reported hearing difficulties among adults with normal audiograms: The Beaver Dam Offspring Study. *Ear and Hearing*, *36*(6), e290–e299.

Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2013). The cingulo-opercular network provides word-recognition benefit. *Journal of Neuroscience*, *33*(48), 18979–18986.

Vaden, Kenneth I., Kuchinsky, S. E., Ahlstrom, J. B., Teubner-Rhodes, S. E., Dubno, J. R., & Eckert, M. A. (2016). Cingulo-opercular function during word recognition

in noise for older adults with hearing loss. *Experimental Aging Research*, *42*(1), 67–82.

Walker, J. J. (2013). *Audiometry Screening and Interpretation*. *87*(1), 8.

Wayne, R. V., Hamilton, C., Jones Huyck, J., & Johnsrude, I. S. (2016). Working memory training and speech in noise comprehension in older adults. *Frontiers in Aging Neuroscience*, *8*.

Wayne, R. V., & Johnsrude, I. S. (2015). A review of causal mechanisms underlying the link between age-related hearing loss and cognitive decline. *Ageing Research Reviews*, *23*, 154–166.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, *7*.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *Journal of Neuroscience*, *32*(40), 14010–14021.

Wild, C. J., Davis, M. H., & Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *NeuroImage*, *60*(2), 1490–1502.

Wilson, R. H., McArdle, R. A., & Smith, S. L. (2007). An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, *50*(4), 844–856.

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing*, *37*(6), 660–670.

Yee, D. M., & Braver, T. S. (2018). Interactions of motivation and cognitive control. *Current Opinion in Behavioral Sciences*, *19*, 83–90.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Heslenfeld, D. J., & Rönnberg, J. (2012). Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. *Brain and Language*, *122*(2), 103–113.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., & Rönnberg, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*, *134*(3), 2225–2234.

Appendices

Appendix A: Comprehension questionnaires

Arctic

1) What was the objective of the narrator's expedition?
    a. Dogsled across Arctic America
    b. Dogsled across Greenland
    c. Climb Mount Kanchenjunga
    d. Climb Mount Everest
2) How did the narrator prepare for her expedition?
    a. Trained in Alaska learning how to dogsled
    b. Exercised several hours each day
    c. Practiced meditation
    d. Practiced living on rations and taking cold showers
3) Which dog on the narrator's dog team never followed instructions?
    a. Robert
    b. Douggie-dog
    c. Billie
    d. Max
4) What did the narrator do for leisure while in her tent?
    a. Read
    b. Play games on her phone
    c. Paint
    d. Take photos
5) How many sponsorships did the narrator receive to support her expedition?
    a. None
    b. One
    c. Ten
    d. Fifty
6) What is the name of the narrator’s neighbour?
    a. Dave
    b. Andrew
    c. Alan
    d. Doug
7) How did the narrator’s neighbour respond when she told him about her expedition?
    a. He said she was going to fail.
    b. He threw her a party.

c. He offered to cover all expenses.
d. He said he was worried about her.

8) What were the weather conditions in the Arctic like the winter of the narrator's expedition?
   a. There were more storms than any other winter in recorded history.
   b. It was the warmest winter in recorded history.
   c. There were fewer storms than any other winter in recorded history.
   d. It was the coldest winter in recorded history.
9) How did the dogs respond when they saw the bears?
   a. Approached them
   b. Ran away in shock
   c. Barked aggressively
   d. Whimpered and submitted
10) How did the narrator's encounter with the polar bears end?
   a. The bears walked back to their den.
   b. She shot the bear with her shotgun.
   c. The bear killed one of her dogs.
   d. Her dogs killed the bear.

Space

1) How did the narrator train for his swim test?
   a. Took his kids to the pool every day
   b. Practiced martial arts
   c. Signed up for a swim class
   d. Swam at the lake every day
2) How many astronaut candidates reported for duty at the Johnson Space Center?
   a. Forty-four
   b. Two
   c. One hundred
   d. None
3) What was the final step of the first swim test in the pool?
   a. Tread water with hands above the water
   b. Hold breath for two minutes
   c. Swim three laps of the pool underwater
   d. Perform a water rescue
4) How many astronaut candidates passed the first swim test?
   a. All of them
   b. Two
   c. All but one

   d. None
5) Where is the home of naval aviation?
   a. Pensacola
   b. Washington
   c. Los Angeles
   d. Dallas
6) What were the astronaut candidates trained to do at the naval air station?
   a. Eject out of aircraft and survive in water long enough to be rescued
   b. Fly fighter jets
   c. Withstand the g-force of a rocket launch
   d. Live in space
7) How many times did the astronaut candidates have to complete the final exercise to pass?
   a. Two
   b. Ten
   c. One
   d. Five
8) What is the narrator afraid of, aside from the water?
   a. Heights
   b. Snakes
   c. Dogs
   d. Loud noises
9) What gave the narrator his greatest feeling of accomplishment?
   a. Passing the water survival course
   b. Completing his PhD at MIT
   c. Raising kids
   d. Seeing Earth from space
10) How often does the narrator swim, now that he is a good swimmer?
   a. Never swam again in his life
   b. Every weekend with his kids
   c. Occasionally at his cottage
   d. Every day at the local pool

Appendix B: Supersubject resampling and correlation reference

Parameters

- *N* participants
- *v* timing versions (∴ *v* participants form one supersubject)
- $n = \frac{N}{v}$ supersubjects per set (assuming exactly *n* participants per timing version)

Resampling & correlation

Step 1: Resampling sets of supersubjects

- With *v* groups of *n* participants, there are n^v ways to choose the first supersubject.
- Sampling participants without replacement there are $(n-1)^v$ ways to choose the second,
- $(n-2)^v$ ways to choose the third, etc.
- The number of possible set permutations, *s*, is therefore given by

$$\begin{aligned} s &= \prod_{i=0}^{n-1}(n-i)^v \\ &= (n-0)^v \cdot (n-1)^v \cdot (n-2)^v \cdot \cdots \cdot 1^v \\ &= (n!)^v \end{aligned}$$

Step 2: Computing splits within each set

- The number of unique splits, *h*, is given by

$$h = \frac{1}{2} n\mathrm{C}k \ ,$$

- where $k = \frac{n}{2}$

Step 3: Permuting time courses within each split

- *p* time course shifts, where number of possible shifts = number of samples in time course

---

Size of correlation distributions obtained

Observed

- $s \cdot h$ observed correlations (r_{obs})
- s observed correlation distributions (o_{dist})

Permutation

- $s \cdot h \cdot p$ permutation correlations (r_{perm})
- $s \cdot h$ permutation correlation distributions
- s aggregate permutation distributions (p_{dist})

Number of distance metrics obtained

- $s \cdot h$ p values (from comparing each r_{obs} to corresponding permutation p_{dist})
- s ROC curves (from comparing each o_{dist} to corresponding p_{dist})

Summary of actual values used for resampling & correlation

	Lab	Online
N	63	136
v	6	6
n	10	14
s	50	50
h	126	1716
p	500	500

Note. s was set manually due to computational limitations

Appendix C: Glossary of methodological terms

supersubject	An aggregate of participants (with size equal to versions of pseudorandomly distributed trial timings) with a complete time course of evenly-spaced responses (in this case with a resolution of 0.5 Hz)
set	One possible configuration of exhaustively sampled supersubjects, consisting, for each supersubject, of one randomly sampled participant from each timing version
split	One possible configuration of two equal-sized groups of supersubjects within a set, between which correlations are computed
inter-supersubject correlation	Correlation between supersubject time courses