The role of individual variability in tests of functional hearing

by

Maury Courtland

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(Linguistics)

December 2019

Copyright 2019 Maury Courtland

Dedication

To my family, who have been there for me since the beginning.

Acknowledgments

It takes a village to raise a doctoral student. This acknowledgments list attempts to begin to repay the massive debt owed to the many people who have made my journey possible.

Mentors and Colleagues Firstly, an enormous amount of gratitude goes to my advisor, Jason Zevin. You have taught me so much both directly and by facilitating my own learning. You have been supportive, knowledgeable, nonjudgmental, and an amazing scientific and academic role model. Thank you as well for caring about me as a whole human being, and for checking in to make sure my sanity, happiness, and life outside of work were all intact throughout the long grad school process. Lastly, thank you for creating a warm, welcoming, and nurturing lab environment that created the type of community I thought I wouldn't find again after Haverford. On that note, thank you to all my fellow Zevlings – especially Andrés Benitez and Erin Ryan – you made my time in grad school far more tolerable through everything from venting sessions over lunch to feverish whiteboard brainstorms to celebratory lab outings. I owe a particularly large thanks to our lab manager, Melissa Reyes (and originally Brian Bauman), and the numerous RAs over the years for what I'm sure were many long hours of running my behavioral experiments. Without you, this dissertation would not be possible. Secondly, I owe so much to my original and co-advisor, Louis Goldstein, your wisdom and patience guided me through my first years in the program as my interests changed like the (non-LA) weather. You taught me to appreciate life even in grad school, to have a healthy perspective on research and academia, to not take things *too* seriously, to get to know how I work, and to love Los Angeles. Lastly, a huge thanks to Morteza Dehghani, you could not have come into my life at a better time. You have always been understanding, supportive, insightful, kind, motivating, and uplifting.

Also, thanks to my colleagues and collaborators in Morteza's lab – especially Brendan Kennedy, Aida Davani, and Leigh Yeh. Thank you as well to the members of the Linguistics department with whom I shared many fun times exploring and enjoying Los Angeles and Southern California. Principal among the linguists are Caitlin Smith, Ana Besserman, Brian Hsu, Johanna Klages, Bhamati Dash, Hayeun Jang, and Cynthia Lee. Thanks as well to Shaikat Hossain and Tanner Sorenson for being the best part of HCN events.

Family Thank you to my wife, Jamie Courtland, for *everything*. You are my role model, brain-storming partner, confidant, therapist, cheerleader, comedian, running partner, and so much more. Additionally, thank you for being infinitely patient with the woes of a long distance relationship. We made it. Thank you to my mother and father, Shira Lander and David Portnoy, your complementary and complimentary support carried me through this ordeal. You showed ceaseless interest in my work, enduring many long conversations and ramblings. Ima, you have always served as my academic role model: from you I inherited my research curiosity and persistence. I owe a special thanks to you for suggesting I take a Linguistics 101 course my Freshman year at Haverford, otherwise this might be a Physics dissertation. Aba, you always know how to calm me down in times of stress, putting things in perspective and providing me with invaluable advice on navigating life. From you, I inherited my love of the spotlight and my drive to enact practical change with my research. To my brother, Zachary Lander-Portnoy, and his wife Elana Lander-Portnoy, you both have always known how to help me not take myself too seriously and celebrated my accomplishments with me. To my paternal grandparents, Vivian and Gerald Portnoy, your unconditional love and support and peaceful San Diego getaways sustained my mental health throughout this trying program. Thank you both for teaching me what's really important in life, and for trying your very best to understand my dissertation. Additional thanks to my great Aunt Denise Friedman, for your San Diego getaways and constant encouragement that I have been on the right track all along. Thank you to my Aunt (Sharon Portnoy), Uncle (Mark Danzig), and cousins (Leo and Ben Danzig), for your frequent entertaining questions about Linguistics and for acknowledging at family events that I know a thing or two

about linguistics and science, and for supporting me in my arguments against prescriptivism. Thanks to my maternal grandparents, Rose and Yechiael Lander, for beginning the scholarly tradition that I am a part of today, and for (mostly) forgiving me for not choosing UMass. I owe an *enormous* amount to the Zarrow family – particularly to Josh, Marlene, and Melanie as well as Stan and Sheila – for adopting me and truly welcoming me into the family. For all the family dinners in Encino and Calabasas, for always sending me home with the leftovers that doctoral students subsist on, and for helping me de-stress on a regular basis with a healthy dose of fun and family. I am eternally grateful to you all. Thank you to my cousin Kayla Shore, your activism and conscientiousness have always been an inspiration for me. Thank you for helping me see how I can do good in the world through my work and support the values I hold dear. Thank you to my in-laws, Doug and Kerrie Croucher, for your enthusiastic support of and interest in my work, for always believing in me, and for your steadfast encouragement to continue my program.

Friends I owe much of my success to Kevin Li. Kev, you have been endlessly supportive, repeatedly convinced me to stay in grad school, been my professional guru and consultant on industry, been a role model for work ethic, dragged me out of my research hole on amazing adventures, kept me woke, and been the other half of some of the most insightful and life-changing conversations I've ever had. Thank you for sticking with me all these years. I owe much of the maintenance of my happiness throughout the PhD program to Jake Bakovsky (and more recently also Kristine Santos). Achi, you have always enthusiastically supported my research, have sympathized with my struggles and supported me through them, always put me in a joyous mood, offered wisdom from a different yet familiar life experience, and never shied away from a deep conversation. Thank you for all the amazing weekends down in OC. Thank you to Anna Schall, your continued friendship, chats, and visits were always highlights to look forward to. Lastly, thank you to Dalton Hughes and Megan Kelly, you made my visits to Durham feel like coming home and have made my transition to Durham better than I could have ever asked for.

Table of Contents

Abstract

This dissertation explores the role of individual variability in real-world (functional) hearing ability and how this ability is assessed in the clinic. It principally concerns two domains that play a role in overcoming the noisy environments in which speech regularly occurs: the ability to use acoustic structure inherent in noise to overcome its deleterious effects and the ability to use one's learned linguistic regularities to aid in reception of the target speech. The properties of each ability and the cognitive mechanisms that underlie them are probed and subsequently discussed in the chapters contained herein. Chapter 2 probes the use of temporal regularity present in noises to overcome their masking effects. Chapter 3 further delineates the properties of the noises found in chapter 2 and explores their relationship to canonical maskers. Chapter 4 explores the ability to use one's internal language model to improve in the reception of masked speech. Their individual abstracts are as follows:

Chapter 2 In the real world, speech perception frequently occurs under adverse listening conditions. Laboratory studies have identified distinct phenomena associated with more or less constant sources of noise (energetic masking), and competing verbal information (informational masking). One issue complicating direct comparisons between them is that common paradigms for studying energetic and informational masking differ along many dimensions. We have developed a paradigm that uses temporally patterned noise, with the goal of comparing energetic and informational masking under more comparable conditions. We hypothesized that listeners would be able to take advantage of the structure in the masking noise, providing a processing advantage over energetic masking. The initial experiment provides strong evidence for this hypothesis, but conceptual replications did not produce the same pattern of results – at

least with respect to measures of central tendency. A direct replication of the first experiment did not replicate the large differences in the means. Interestingly, however, analyses across all four experiments reveal robust evidence that patterned noise conditions produce increased individual variability. Further, we observed strong correlations, specifically between the patterned conditions. We attribute these findings to as yet unidentified cognitive ability differences allowing some participants to benefit from the use of additional temporal information while others are hurt by the addition of unusable distracting information. Hypothesized predictive measures of task performance, such as working memory, inhibitory control, and musical experience did not correlate with performance, however.

Chapter 3 This work serves as a companion to our previous work in Courtland et al. (2019). Here, we undertake several follow-ups to delineate the properties of the temporally regular masking patterns introduced in that work – particularly where the new maskers are situated vis-à-vis canonical energetic and informational masking. In this paper, we perform experiments using a dynamic spectral profile for the maskers, manipulating their global SNR, and increasing the number of patterns presented prior to the target word. The findings support much of the conclusions of the previous work with respect to the role of offline information (the recurrence of a pattern across trials) and its importance in moderating task performance. While much of the support comes from negative results, the increased number of repetitions replicates our initial finding that shows a significant difference in task performance between a recurring condition and a baseline control. This experiment also replicates the initial finding of a significant difference between an online, repeating masker and its offline counterpart that controls for glimpsing windows. We also observe replications of variance effects seen in the previous paper, supporting the conclusion that offline information plays an important role in individual variability.

Chapter 4 Cognitive tests used to measure individual differences have traditionally resorted to standardizing testing materials because of the onerous nature of creating test items. These tests are generally designed with *equality* in mind: the same "broadly acceptable" items are used for all participants. This approach ignores participants' diverse language experiences that potentially significantly affect testing

outcomes. This has unknown consequences for *equity*, particularly when a single set of linguistic stimuli are used for a diverse population of language users. We hypothesized that differences in language variety would result in disparities in psycholinguistically meaningful properties of test items in two widely-used cognitive tasks (reading span and SPiN), resulting in large differences in performance. As a proxy for individuals' language use, we administered a self-report survey of media consumption. We identified two substantial clusters from the survey data, roughly orthogonal to *a priori* groups recruited into the study (university students and members of the surrounding community). We found effects of both population and cluster membership. We then sought to explain this finding of significant performance differences between clusters on the tests based on their media consumption. To this end, we modeled the language contained in these media sources using an LSTM trained on corpora of each cluster's media sources to predict each test item's target words. We also modeled semantic similarity of test items with each cluster's corpus using skip-thought vectors. We found robust, significant correlations between performance on the SPiN test and the LSTMs and skip-thought models, but not the reading span test.

Chapter 1

Introduction

1.1 The Role of this Dissertation

Speech perception and language comprehension occur in a variety of environments throughout daily life. Despite the fact that many of these environments represent suboptimal conditions in which to perform the task, humans quickly, easily, and precisely decode sensory stimuli to make sense of the incoming signal. This robustness is due in large part to the critical contribution of cognitive mechanisms to several stages of the process. I view this dissertation as an inquiry into the extent to which predictive top-down cognitive processes can facilitate the bottom-up peripheral processing load required to carry out online language comprehension. I probe this question in two domains: speech perception in adverse listening conditions and reading under competing cognitive demands. In the first, I investigate the ability of the auditory system to track, model, and overcome masking (interfering noise) to enable the perception of noisy speech. In the latter, I investigate the benefit to language comprehension of lexical predictions generated by a learned statistical language model and probe these predictions' reliance on the fidelity of the language model to incoming language.

The decision to study these mechanisms under challenging conditions was made to allow insight into their inner workings. When systems operate in normal, ideal conditions, their automaticity and proficiency often do not allow researchers adequate information into what their components are, how they integrate,

and how they operate synergistically to achieve the goal of the entire system. Placing stress on the system, possibly to failure, affords a glimpse into which component bore the stress placed on the system and how this component interacts with other components in the system. Systematically varying the type and amount of stress in different experiments may therefore begin to provide pieces of the mosaic from which our understanding of the system is drawn. The two principal mental processes of interest in the following studies are perception and memory.

In this introduction, I offer a brief overview of perception and memory and how they relate to the work contained in this dissertation. I then discuss verbal working memory and introduce the construct of functional hearing testing used and evaluated in this work. In the following chapters, I present the studies I have carried out along this line of inquiry. Finally, I conclude with directions and considerations for future work to be built upon the findings of the studies contained herein.

1.2 Perception

Perception is the process through which incoming sensory information is captured, bundled, categorized, and interpreted. The process begins with the generation of a distal stimulus in the external world by a physical process, such as from a light source or air perturbation. While many physical processes altering the stimulus may occur en route to the sensory organs, these are not within the domain of perception. Perception continues when the physical occurrence is sensed via a sensory organ (e.g. retina or organ of corti). This excites neural activity, dubbed the proximal stimulus, and the process of transduction begins (Goldstein, 2014). Along the pathway from the peripheral to the central nervous system (CNS), several mechanisms filter and process the proximal stimulus until it is eventually recognized and interpreted as a cohesive percept.

Given that the process originates with an external stimulus that then affects a change in our mental processing, perception may initially seem to be a passive process. However, knowledge and expectations are actively involved in the process of perception as evidenced by the long-standing psychological tradition

of illusions (Gregory, 2004). The active role of the CNS has led to several theories regarding its precise function and necessity in perception. While the field of perception research applies across all sensory modalities, only the auditory modality, with specific focus on the speech signal, will be discussed here for the sake of brevity. Given the predication of our findings on the functioning of the auditory system (particularly in its function of speech perception), the background provided here has great importance for situating the work contained in this dissertation.

1.2.1 The Auditory Pathway

The auditory pathway begins in the peripheral systems that supply the brain with electrical input derived from acoustic energy. These systems begin at the outer ear with the pinna, which acts as a physical filter providing directional information about a sound source and boosts frequencies in the range of the human voice (Middlebrooks and Green, 1991). After the sound enters the ear, its acoustic energy is transformed into mechanical energy by the ear drum. The sound then passes through the impedance matching of the middle ear to the cochlea. In the cochlea, hair cells in the organ of corti then transform the mechanical energy into electrical energy. This signal of electrical energy is then transduced to the CNS. It has been argued that many of these peripheral processes are compensatory evolutionary measures to allow a previous sense of underwater pressure detection to operate in a new environment providing aerial hearing (Christensen et al., 2015). Regardless of their origins, it is certain that sound travels through and is affected by each step in this cascade, and thus it is relevant to the current research.

Upon reaching the CNS, the electrical signal passes through various neural filters that take the raw electrical signal from the organ of corti and transform it into conscious level auditory perception. One of the first operations the CNS performs is using coincidence detectors to perform sound localization by comparing the signal delay between the ears compared with the intra-aural distance. This source triangulation then allows listeners to construct a map of where sound sources originate in space and focus on certain regions while ignoring others (Fritz et al., 2007). This focus on certain regions of space plays a large role in selective attention: the ability of the brain to focus on certain inputs while ignoring or decreasing others.

The topic of selective auditory attention was first studied by Cherry (1953) in his examination of "the cocktail party problem" (see also Cherry 1957), so named because listeners could focus on the conversation they were participating in at parties while ignoring the din of other conversations happening around them. This process relies on subjects' application of their expectation and experience to the active perception of sound in adverse listening conditions and as such is heavily implicated in the auditory experiments in this work. Since Cherry's early work describing the phenomenon, several lines of inquiry have grown out of his findings. One of principal importance here is the role of active rhythm and beat-tracking mechanisms in allowing auditory selective attention (Teki et al., 2011, 2013; Andreou et al., 2011; Sohoglu and Chait, 2016a,b). In these studies, the predictable, regular nature of incoming auditory stimuli allows predictive mechanisms to enable sound segregation. Another closely related offshoot is the ability to encode and reliably track the speech signal across time which is a crucial task in ensuring successful speech perception (Ahissar et al., 2001; Kubanek et al., 2013).

1.2.2 Speech Perception

Given the necessary acoustic building blocks by the auditory system, the speech perception system processes the filtered acoustic signal further to create a form that allows for lexical access. In order to transform acoustic signals into stable lexical items, several adverse, but common, factors must be controlled for. It is important to note that the processes overcoming the following perceptual hurdles do not arise as simple consequences of the auditory system's organization (as some of the aforementioned processes do). They require active application of language experience, which must first be acquired, to facilitate accurate speech perception. This fact must be adequately accounted for in any valid theory.

The first, most widely pervasive factor is the lack of invariance that the system encounters. Simply stated, there is no reliable mapping between the speech signal and the abstract categories whose instances (phonemes) comprise it. This is due to phonological transformations, speaker variation, free variation (e.g. speech rate, carefulness), and random variation. Yet in spite of all this variability, the system adeptly categorizes sounds correctly and often does not notice any difficulty.

One theoretical account for this comes from Exemplar Theory (Johnson and Mullennix, 1997). It posits that the role of the speech perception system is not to match a stimulus to an identical template, but rather to the best match in a listener's previous experience. Therefore, items do not need to exhibit constancy. A related account of best match is given in fuzzy logic (Oden and Massaro, 1978), however, in this theory sub-lexical sounds are perceived independently and assigned a fit value (in a fuzzy fashion) to phoneme prototypes in long-term memory. These fit values are then combined combinatorially to generate the set of possible words. Each possible word is assigned a probability proportional to the goodness of its fit, and the lexical selection chooses a word probabilistically. Similar to the probabilistic fuzzy logic account is the connectionist account (i.e. TRACE: McClelland and Elman 1986; Merge: Norris et al. 2000), in which activation propagates through a network and eventually reaches threshold signifying phonemic identification. As the network is similarly concerned only with proportional activation and which phoneme first reaches threshold, there is no need to match absolutely between acoustics and phonemes.

A different account is that listeners are not actually concerned with the entirety of the speech signal, rather they hunt for reliable acoustic evidence (landmarks) of the component gestures that produced the signal (Stevens, 2002). A similar account but with attention to the entire signal is given by Direct Realism (Fowler, 1986). A related gestural approach lies in the Motor Theory of Speech Perception (Liberman et al., 1967), but this postulates abstract mental representations of gestural units onto which the listener must map the acoustic signal. In all the gestural accounts, because gestural inventories represent a finite, limited set, and their identity can be reliably recovered from the acoustics in the speech signal (either in part or total, directly or through abstraction), the lack of invariance is not a problem.

An additional phenomenon which speech perception theories must account for is categorical perception (Liberman et al., 1957). This is the finding that listeners are particularly sensitive to acoustic variations across phonemic boundaries (exhibiting a nearly immediate boundary), but relatively insensitive to variations within phonemes. While the Motor Theory of Speech Perception was developed in part to explain this finding, similar categorical selection steps of several of the above theories achieve the same end (Exemplar Theory, TRACE, Merge).

It is less clear, however, how Direct Realism accounts for categorical perception given its observed stimulus is the actual dynamics of the vocal tract. In fact, according to the Quantal Theory of Speech (Stevens, 1968), in which gestures typically occur in regions of relative acoustic stability, linear interpolations between phonemes' acoustic cues would be inferred as a large jump towards the gestural boundary. Additionally, for the fuzzy logic account, the probability of phoneme selection should increase linearly with the linear interpolation, rather than the sharp boundary observed. Lastly, given the assumption of Acoustic Landmark theory that acoustic landmarks are stable given their gestural sources, it remains unclear how the perturbation of these landmarks would affect the recovery of gestural information.

1.2.3 Lexical Activation

In addition to the previously discussed phenomena, higher levels of linguistic processing can influence perceptual processing in a top-down fashion. In an example from lexical processing, Ganong (1980) found that categorical boundaries (like those described above) can be systematically shifted toward one phoneme if it completes a word while the other does not (e.g. dash-tash, dask-task). This top-down effect provides evidence for lexical access theories such as the Cohort model (Marslen-Wilson, 1987), in which the task of lexical activation is to select possible words from the entirety of a listener's lexicon given online auditory input. Before any auditory input, the size of the cohort being considered is the entire lexicon, and as more evidence is gathered, fewer words will be possible matches to the partial evidence.

In contrast to the Cohort model, however, the Ganong effect implies that words whose phonemes do not initially match the input are still maintained at some level of activation such that they can be selected given subsequent information. This is necessary to explain why a sound whose decision boundary would normally classify it as a /t/ (e.g. in a dash-tash paradigm) would shift its classification only upon hearing the final sound of the word /ʃ/. This ability may be a repair mechanism of the system to achieve robustness to speech errors or other misclassifications. Given that the effect is demonstrated when bottom-up processing is unreliable (i.e. when acoustic cues exhibit maximum phoneme-category entropy), this may be evidence that as bottom-up cues become less reliable, an increased reliance is placed on top-down processing.

There is evidence that such a repair mechanism exists in the extreme example of the phonemic restoration effect (Warren, 1970). In a similar repair mechanism to that above, when phonemes are deleted from the speech signal entirely (i.e. contribute no acoustic information) and covered with a noise (e.g. a cough), listeners can restore the most probable missing phoneme based on context (syntactic, pragmatic, etc.). Moreover, listeners are so unperturbed by this manipulation that they cannot reliably identify when the noise occurred within the sentence. When combined with the Ganong effect, this provides clear evidence of top-down processes serving as repair mechanisms. As numerous experiments in this work concern the obscuring of speech sounds by noise, these types of top-down repair mechanisms are heavily implicated in enabling successful performance of the task.

In addition to repair, another crucial top-down process at play here is lexical prediction. While repair aids perception after an event has happened, many anticipatory processes facilitate perception's ease and accuracy. Anticipatory processes occur at all levels of language processing. Beginning with early auditory processing, evidence for prediction comes from the mismatch negativity paradigm, in which a regularly occurring pattern is interrupted with a rare stimulus (Näätänen et al., 1978). Neural recordings can be taken using electroencephalography (EEG) and the strength of the response can be measured (Haesen et al., 2011). Experimenters can then vary whether an identical stimulus occurs as part of the repeating pattern or the oddball. The difference in the strength of the EEG response between these two conditions is thus taken as a measure of listener surprisal (Näätänen et al., 2007). Large mismatch negativity signals are evidence that listeners have anticipated the upcoming stimuli, and a violation of that expectation causes an increase in the EEG signal proportional to the prediction strength (Näätänen et al., 2011; Roberts et al., 2011). This low level auditory prediction is central to the listening studies presented here given that their regularity is the experimental manipulation of interest.

An EEG measure similar to mismatch negativity can be taken as a measure of surprisal with regards to lexical prediction. The N400 component is particularly sensitive to items that do not follow from their preceding contexts (Kutas and Federmeier, 2010). This is often used in language studies to measure the effect that a word's likelihood of occurring has on that word's perception. Similar to the measure

above, conditions can then vary whether a word is likely or unlikely given the preceding context. This can be taken as a measure of semantic anomaly or conversely, semantic predictability (Bentin et al., 1985; Holcomb and Neville, 1990; Kutas and Hillyard, 1980, 1984). Given that the tasks employed in chapter 4 stress the importance of sentence final words, (i.e. its probability following a given sentential context), understanding the effects that pragmatic predictions have on online language comprehension is imperative.

In addition to semantic prediction, the N400 has also been used to show that participants predict lexical items on the basis of grammatical feature agreement. This has been shown in Dutch adjective/noun gender agreement (Otten et al., 2007; Van Berkum et al., 2005) and Spanish article/noun agreement (Wicha et al., 2004). These anticipatory effects also apply to expectations generated by phonological rules, as in English's 'a/an' agreement (DeLong et al., 2005). Furthermore, these processes are not modality specific and semantic prediction has also been found in German sign language (Hosemann et al., 2013).

1.3 Memory

Memory is the process through which information is encoded, stored, and subsequently retrieved (Melton, 1963). While traditionally conceptualized of as separate stages, the current view is that the encoding and retrieval phases interact with one another (Neath, 2000). This change was precipitated by studies such as Thomson and Tulving (1970), which showed that absolute statements in the encoding and retrieval of memory are often incorrect because they do not take into consideration the specifics of the encoding and retrieval environments. An extreme example is found in 'recognition failure of recallable words' Watkins and Tulving (1975): in a recall task for test items, the items themselves were not in fact the most reliable cue (e.g. GLUE is a better cue of CHAIR than CHAIR is in remembering whether CHAIR occurred on the test list).

Given that the cognitive tests used in this work either explicitly test memory or implicate its function in related cognitive mechanisms, a survey of memory research is necessary here. Because the tasks focus on linguistic functions of memory, the discussion here will be limited to those aspects of the memory systems

that are implicated in linguistic processes. These linguistic processes do not occur simply at one level of memory and as such, each component of the system discussed in this section has bearing on the work conducted in the following chapters.

1.3.1 Components of the Memory System

Memory models traditionally contain three levels: sensory memory, short-term or working memory, and long term memory (Atkinson and Shiffrin, 1968). In this model, sensory input enters into sensory memory, where the stimulus briefly persists. For echoic memory, the sensory memory of the auditory system, this persistence typically lasts between 1.5-5 seconds (Treisman, 1964a; Glucksberg and Cowen, 1970), though it can persist up to 20 seconds without a new stimulus to overwrite it (Norman, 1969). The visual sensory store, iconic memory, in contrast persists only about 1 second (Sperling, 1960). In the context of the reading experiment in chapter 4, this implies that a participant's auditory feedback from reading the sentence aloud will persist in echoic memory longer than the reading of the text will in iconic memory.

Recently, it has been called into question whether the sensory store is a component of the memory system. The finding of the inverse duration effect by Bowen et al. (1974), indicates that the longer a stimulus lasts, the less time it persists for. Given that this is counter to the notion of a memory store (wherein a stimulus *increases* its persistence with increased duration), researchers have begun to view the sensory store as simply the remains of neural activation caused by sensation of the stimulus (Neath and Surprenant, 2005).

From the sensory store, the information travels to the short-term store. In the short-term store, information can persist without rehearsal for 20-30 seconds (Peterson and Peterson, 1959; Posner, 1966). Rehearsal is the process of actively renewing information in the short-term store by attending to or mentally repeating it. This enables information to continue to persist longer than its normal decay would allow. In contrast to the sensory store, information at this level is not modality dependent (e.g. read information can be preserved by auditory rehearsal). As with overwriting in the sensory store, the short-term store has a limit on how much information can persist in it at one time. This is classically referred to as Miller's

magic number: seven plus or minus two (Miller, 1956). It should be noted that the unit of information here – often called a "chunk" – represents any piece of information that can be treated as a single cohesive entity. Therefore, storing a multi-digit array representing a familiar birthday may be treated as only one piece of information if you instead rehearse the birthday in the short-term store and then transform it back to the multi-digit array on recall. This hierarchical "chunking" is theoretically limitless, and incredible displays of memory consolidation have been observed (Ericsson and Staszewski, 1989; Ericsson and Kintsch, 1995).

Information is continuously being transferred from the short-term store to the long-term store with the strength of the memory trace increasing over time. Thus, rehearsing an item repeatedly in the short-term store can better encode it into the long-term store (although only when processed deeply rather than shallow rote repetition: Craik and Watkins 1973). The path from the short-term store to the long-term store is bidirectional with items transferred or retrieved as needed to attend to in the short-term store. The long-term store is traditionally conceived of as permanent (in the context of a lifespan) and functionally limitless (or at least enormous: Landauer 1986). In fact, an inability to recall something from the long-term store does not imply it no longer exists in this model. A failure to recall information only implies that its retrieval cue – the means by which the item is located and brought back into the short-term store – has deteriorated to the degree to which the item is irretrievable (Capaldi and Neath, 1995). Given that cues deteriorate naturally over time (Thorndike, 1913; Brown, 1958), stored information may be unable to be recalled while still persisting in the long-term store.

Following Atkinson and Shiffrin (1968)'s model, Baddeley and Hitch (1974) proposed a model of working memory which subsumed the short-term store. Two main differences exist between Baddeley and Hitch (1974)'s model and Atkinson and Shiffrin (1968)'s short-term store. Firstly, working memory is more than simply a store: the system handles both storage and processing (computation) of short-term information (Baddeley and Hitch, 1974; Baddeley, 1986; Hitch and Baddeley, 1976). Secondly, working memory consists of several different components. It exists in a hierarchy between a central executive module – controlling information flow and the operations of the working memory system – and the "slave

systems" which are modality specific short-term stores. In its original conception, two slave systems existed: the "Phonological Loop" and the "Visuo-Spatial Sketch Pad". The phonological loop stores auditory information while the visuo-spatial sketchpad stores visuo-spatial memory. Recent formulations have added a third slave system: the "Episodic Buffer". This system stores smell and taste information (Baddeley, 2000). Given its recency and relative lack of study, the Episodic Buffer remains the least well specified slave system.

1.3.2 Verbal Working Memory

Baddeley and Hitch (1974) (see also Baddeley 1986; Hitch and Baddeley 1976) first proposed the model of working memory as a cognitive system which handled both temporary storage of information as well as processing of this information. Their experiments taxed working memory by placing competing demands on the system of simultaneous processing and storing: participants held digits in memory while attempting to comprehend sentences. These two tasks showed performance trade-offs and thus seemed to draw from a common, and fixed, pool of cognitive resources. Regarding online language processing, the "processing" aspect of working memory corresponds to (among others): parsing incoming language, lexical retrieval, and constructing syntactic and semantic structural representations. The "storage" aspect corresponds to preserving previously processed information for later uses, including: to generate syntactic and semantic expectations, to resolve references to or agreements with previous lexical items, and to maintain a coherent discourse narrative.

Given the implication of both functions of working memory in online linguistic processing, it is surprising that traditional measures of working memory were found to be either uncorrelated or only weakly correlated to reading ability. Studies using a digit span or probe digit span task found no correlation to participants' performance on a general reading comprehension test (Hunt et al., 1973; Guyer and Friedman, 1975; Perfetti and Goldman, 1976). Studies using letter strings or similar sounding words found only weak correlations (Rizzo, 1939; Valtin, 1973; Farnham-Diggory and Gregg, 1975).

1.3.2.1 Daneman and Carpenter 1980

In light of this, Daneman and Carpenter (1980) hypothesized that previous measures of working memory (like digit and word span tasks) only involved simple rehearsal and retrieval of common lexical items, and thus were not adequately taxing working memory's processing component. They aimed to create a task that would sufficiently tax both roles of working memory to better approximate the complex demands on the system in its naturalistic functions, thereby illuminating its role in language comprehension. They proposed two working memory language tasks that they found to correlate well with reading performance. Given the parallel nature and purpose of the task to digit and word span tasks, but its emphasis on predicting reading ability, they named the task the "reading span" task. An auditory equivalent named the "listening span" task was also designed.

In the reading span task, participants read sentences printed on index cards aloud in immediate succession. These sentences were grouped in sets of increasing size and participants were asked to recall the final word of each sentence of the set promptly at the set's conclusion. There were three trials at each set length and participants attempted each length until they failed on all three trials of a certain length. The greatest length at which they had scored at least two of the three trials correct was taken as their "reading span".

This test (and the listening span test) produced high correlations with measures of reading ability, which was assessed by participants' verbal SAT scores and their performance on comprehension questions following short vignettes. These questions are common in reading comprehension tests (Carroll, 1971; Davis, 1968, 1944), and tested participants' understanding of the passage's theme and resolving certain grammatical constructions like pronominal reference. The authors believe these types of questions approximate the processing demands that discourse tracking and syntactic parsing place on verbal working memory during naturalistic language processing.

In attributing the observed individual differences to differences in working memory ability, the authors cite several reasons why this account is plausible. They claim both the time course and information load of the task are compatible with working memory. They also claim that the amount of variance captured

by, and the strength of the correlation between, reading comprehension and reading span suggest that the limitation of an individual's working memory is the common underlying factor, making it an important source of individual differences in reading ability. Additionally, given the high correlation between their reading span and listening span tasks, they hypothesize verbal working memory to underlie language processing in general, regardless of input modality.

1.3.2.2 Just and Carpenter 1992

Following Daneman and Carpenter (1980)'s development of the reading span and listening span tasks and several subsequent studies, Just and Carpenter (1992) present a computational theory of limited working memory capacity and its effect on several aspects of language processing, including those measured by Daneman and Carpenter (1980). Their work first models the results of numerous language processing experiments and attributes their results to individual differences in participants' working memory capacity. Given their hypothesis that participants' working memory capacity is the limiting factor in language comprehension ability, they name their theory "capacity constrained comprehension". They then present a computational model that matches the predictions of their theory in which the only manipulation between the high and low performance conditions is the amount of working memory capacity available to the system.

In keeping with previous work, the authors conceptualize of, and model, working memory as subserving both the storage and processing of linguistic information, with the two competing for shared activation resources. The authors claim that use of an activation based account shares aspects of both symbolic (e.g. Anderson 2013) and connectionist (e.g. McClelland and Rumelhart 1989) models, capturing important aspects of both. The activation capacity comes into play when the system begins to run out of activation energy. If processing is deemed more important than storage, the system ceases the maintenance of older information in memory to recoup their activation. If conversely storage is more important, the system maintains the activation in old information but processes incoming information at a much slower rate (having less activation available to it at each time step). Models with lower capacity must therefore make

trade-offs that models with higher capacity do not have to make. Therefore, higher capacity models exhibit better performance than lower capacity ones.

To explain why simple tasks do not eventually deplete the activation and cause processing difficulties over long periods of time, they posit cognitive mechanisms that incrementally reduce storage demands as time goes on. They cite mechanisms that automatically and incrementally attenuate propositions from past sentences not central to the discourse narrative (Glanzer et al., 1984; Kintsch and Van Dijk, 1978; Van Dijk et al., 1983), as well as those which aid in processing demands by predicting and pre-activating concepts and structures likely to occur in upcoming sentences (Sanford and Garrod, 1981; Sharkey and Mitchell, 1985). They also posit a reduction of storage demands through the immediacy of processing, in which each lexical item is processed to the fullest extent possible immediately rather than maintaining multiple possible representations of the item in working memory (Carpenter and Just, 1983; Just and Carpenter, 1980). Lastly, they posit that once higher level linguistic structures have formed completely, their constituent unresolved parts can safely be inactivated. To support this, they cite behavioral evidence that readers do not maintain much lexical or syntactic information from past sentences as they read a text (Huey, 1908; Jarvella, 1971; Sachs, 1967).

While they argue that the main dimension of individual differences lies in the amount of working memory capacity available, they do not exclude the role that processing efficiency may play in separating good from poor readers. They suggest that a particularly inefficient mechanism of a poor reader could place undue demands on overall working memory capacity thus causing a bottleneck in processing time (and accuracy) by consuming so many resources (see Perfetti and Lesgold 1977). They do note, however, that the differential performance they observe behaviorally only manifests when comprehension tasks become demanding, and that differences in processing efficiency alone should manifest regardless of the difficulty of the comprehension task. They further draw on evidence that individual differences in reading ability tend to be associated with the speed of a variety of comprehension processes rather than just a single component process (Frederiksen, 1981).

1.3.2.3 Waters and Caplan 1996a

In response to Just and Carpenter (1992), Waters and Caplan (1996a) theorizes that two separate working memory resources underlie language comprehension (Caplan and Waters, 1990; Waters et al., 1995; Waters and Caplan, 1996b). The first working memory system underlies the fundamental, automatic, processes enabling transformation of speech or text into its discourse level representation (e.g. lexical activation, syntactic parsing, semantic composition). The second working memory system enables higher level, consciously controlled, aspects of language comprehension such as intentionally searching through long-term to recall a piece of information or attempting to reason explicitly about incoming linguistic information. They take these systems to be separate in their operations as well as in the resource pools available to them. In response to this dichotomy, Just et al. (1996) note that language processing occurs on a continuum between conscious and unconscious processing, and thus binarizing consciousness into these two systems is problematic.

Their theory claims that given the separate capacities of automatic and conscious working memory, the processing and storage constraints should not interfere across automatic and conscious levels. In support, they cite behavioral findings that if digits are presented prior to sentence presentation, no task interference is found (Waters et al., 1995; Caplan and Walters, 1996). When material to be stored and recalled interrupts the comprehension task however, interference effects are observed (Wanner and Maratsos, 1978; King and Just, 1991; Waters, 1996). They argue it is the interruption of the automatic system's processing, such as that in Daneman and Carpenter (1980), not intrusion by the conscious system that interferes with its functioning. They note that the lack of observed interference was not due simply to the task being too easy for memory constraints to come into play, as participants performed below ceiling on both recall and comprehension tasks (Caplan and Walters, 1996; Waters et al., 1995).

In presenting an account of separate language processing, Waters and Caplan (1996a) align themselves closely with the modular accounts of generative linguistics (Chomsky, 1965; Fodor, 1983; Frazier, 1987) and its counterpart in neuropsychology of specialized language systems (e.g. Martin et al. 1994). While

Waters and Caplan (1996a) does make substantive revisions of the role of working memory in language comprehension, their underlying theory and architecture do not differ significantly from Just and Carpenter (1992), save for positing a division of labor for conscious and automatic processes.

1.3.2.4 MacDonald and Christiansen 2002

In stark contrast to the above accounts separating working memory from the linguistic knowledge it operates on, MacDonald and Christiansen (2002) presents a unified connectionist account. The authors build upon previous work criticizing cognitive capacity accounts as inherently flawed and lacking explanatory power (Navon, 1984), neglecting the crucial role of practice and skill (Ericsson and Kintsch, 1995), and not tracking performance gains with cultivation of expertise rather than increases in working memory in child development (Munakata et al., 1997; Roth, 1984).

In their theory, a naive network undergoes supervised learning and gradually adjusts its parameters to improve performance over time. In their model, linguistic knowledge, its processing, and the capacity of that processing, are all consequences of network structure and parameters, rather than separate overtly modeled components. Their network thus lacks the ability to independently manipulate a concept such as activation capacity. This captures their theoretical claim that linguistic knowledge and processing are inextricably linked, and that individual differences arise not from working memory capacities, but from differences in biological factors and language experience. Given this, they consider linguistic working memory tasks and language comprehension tasks to be different measures of the same underlying skill of linguistic processing. They further state that language experience goes beyond mere vocabulary size (Carpenter, 1994; Daneman, 1988), influencing the online processing of probabilistic constraints (Pearlmutter and MacDonald, 1995), and the formation of syntactic structure (Chang et al., 2000). They claim that their simplified model provides greater falsifiability and fewer possibilities for ad hoc modifications compared with Just and Carpenter (1992)'s model.

In this light, the authors view the reading span task as simply a measure of language processing skills like lexical decision tasks, reading speed, etc. (see also Ericsson and Kintsch 1995; Martin 1995). As such,

the variability it captures in the normal population is explained by: 1) Differential exposure to and practice on language comprehension tasks, particularly reading, and 2) Biological differences such as the encoding fidelity of phonological representations. They thus explain the correlations between reading span and other language processing tests as caused by sharing the above two factors and *not* by a shared verbal working memory capacity. This parallels work by Ericsson and Kintsch (1995) showing that experts' increased working memory capacity does not extend beyond their domain of expertise, whether that be chess or reading comprehension.

Given the inextricable nature of working memory and knowledge in a connectionist system, the authors attribute limitations on processing capacity to differences in network architecture, including: the number of hidden units (Harm and Seidenberg, 1999; Patterson et al., 1989), the propagation efficiency (Dell et al., 1997), and the amount of noise in the input (St John and Gernsbacher, 2013). They also stress the role of training duration on network performance (Munakata et al., 1997) as this serves as both the central manipulation of their simulations and their primary account of previous behavioral data. They note that all of these variations account for individual differences in processing capacity, but that these affect the network holistically, rather than individually affecting storage or processing.

1.4 Auditory Masking and Speech Perception in Noise

Dialogue occurs in a variety of environments throughout the course of daily life. While some of these auditory scenes provide a proverbial blank canvas for linguistic interchange, most contain at least some level of background noise. The auditory system is remarkably resilient in adverse listening conditions, particularly for speech (for reviews see Guediche et al. 2014; Mattys et al. 2012). On busy streets, in noisy bars, and at crowded concerts, people still manage to converse with each other. They maintain this necessary level of speech perception ability despite the plethora of competing acoustic information reaching their ears and vying for their auditory attention.

The obstruction of an interfering stimulus in the auditory system's detection or comprehension of a target signal is known as auditory masking. Auditory masking can be either asynchronous (the target preceding or following the masker) or synchronous (the target being partially or wholly contained within the masker). Asynchronous maskers are described by their position with respect to the target: forward maskers that occur before the target (Widin and Viemeister, 1979; Smiarowski and Carhart, 1975), and backward maskers that occur after the target (Kidd and Wright, 1994). As the listening tests in this dissertation focus on synchronous maskers, the background here will be limited to those. Additionally, because the hearing tests used in chapters 2 and 3 are situated conceptually between canonical masking types, each of these types and the mechanisms that overcome them are introduced here.

Synchronous maskers are traditionally divided into two groups depending on the way in which they obstruct the perception of the target signal. The first type, energetic masking, is thought to hinder perception by physically obscuring the target with noise. The difficulty experienced with energetic masking is attributed to an overlap of the noise and target in exciting the peripheral sensory organs (Brungart, 2001; Durlach et al., 2003). Though the characteristics of the signal and noise are quite different, as in the frequencies of speech and white noise, areas of sensory organs activated by the target are also activated by the noise, and the signal is drowned out and useless to downstream processes. Because of its peripheral nature, we find energetic masking to utilize low level confounds such as exhibiting similar spectral characteristics to speech's long term average spectrum (LTAS) but without the temporal information included in its envelope (Brungart et al., 2006). The closer the noise approximates which parts of the periphery the target will activate, the more interference and blocking to sensory resources the masker can provide.

The second type of masker, informational masking, concerns the presentation of competing information similar to the target signal. Informational maskers confound auditory processing mechanisms by slipping through attentional filters and placing competing cognitive demands on the systems processing the target signal. While the target is still perceivable, the similarity of target and masker makes it difficult for cognitive processes to tease them apart. Top-down processing thus plays a crucial role in the release from informational masking: the ability to selectively devote attentional resources allows the sensory system

to locate, perceive, and process the target information (Zhang et al., 2014). Its predication on top-down processing implies it is a more central cognitive process occurring further downstream the auditory transduction pathway (Leek et al., 1991; Scott et al., 2004). Given its reliance on central processing, increasing perceptual similarity subsequently increases task difficulty. This is evidenced by a synthetic "speech-like" masker of alternating harmonic (vowel like) and non-harmonic (consonant like) speech shaped noise (SSN) providing more masking than either on their own (Chen et al., 2012). Additionally, the distance in fundamental frequency between a target and masker is inversely proportional to the task difficulty (Leek et al., 1991). In line with these, we find the hardest informational masking task to be segregating multiple streams recorded from a single speaker (Cherry, 1953).

Traditionally, informational masking and energetic masking have been relegated to separate domains: informational masking confounds higher central cognitive processes and energetic masking confounds lower peripheral sensory processes. With the exception of space providing a common release from both types of masking, (Best et al., 2005; Arbogast et al., 2002; Ihlefeld and Shinn-Cunningham, 2008b), the interaction between the two noise types and where one's domain ends and another's begins is still largely unknown despite its being quite an old problem (Miller, 1947; Tanner, 1958). While most continue to conceptualize of the two types of masking as separate phenomena, a few have begun to probe the question of how separate the two really are, and whether they can be combined to create novel masking effects. Stone et al. (2012) contrasted the effects of a noise masker with a near-constant envelope and those with various levels of sinusoidal amplitude modulation. They found that the near-constant envelope did not provide as much masking power as the 8Hz sinusoidal amplitude modulated masker. Stone and Moore (2014) extended this line of inquiry by examining which envelope oscillation frequencies provide the greatest masking effect. They found fluctuations in the range of 4-16Hz to provide the most masking. This 4-16Hz range fits well within popular frequency ranges cited for several neurophysiological oscillation patterns related to speech perception (Ng et al., 2012; Riecke et al., 2015). The relative prominence of auditory processing in this range helps explain why maskers within this range are particularly adverse for speech perception mechanisms.

1.4.1 Overcoming Masking: Glimpsing

One prominent theory of how listeners overcome auditory masking is provided by glimpsing (Cooke, 2003, 2006). In the glimpsing model, listeners take advantage of periods of locally favorable signal-to-noise ratios to "glimpse" the target signal beyond it. These glimpses can occur at periods of low noise power or high signal power (or at frequencies where the masker is weak or the signal is strong). This is taken to its extreme in the "Picket Fence" pattern of Licklider and Miller (1948) (see also Miller and Licklider 1950). The picket fence masker alternates isochronous periods of masker and silence, thus allowing listeners to perceive regular predictable portions of the target signal unobstructed. Listeners then piece together the acoustic information they were able to perceive and attempt to reconstruct the message contained in the signal.

Several other manipulations of noise masker envelopes have been performed to allow for differential amounts of glimpsing. A variation of the picket fence pattern was performed by Pollack (1955), who varied the level of the noise masker between the "pickets" between 0% and 100% of the volume of the target speech. He found that increased noise led to decreased intelligibility. Festen and Plomp (1990) manipulated the noise envelope not in an isochronous pattern, but by correlating the envelope and LTAS of the noise to non-target speech. They found that these envelope fluctuations provided release from masking relative to a steady-state masker. Iyer et al. (2007) also performed a variant on the picket fence pattern in which the pickets were portions of a distractor speech signal. In this variation, they tested both an interrupted and a continuous speech distractor and found performance was better in the continuous condition than the interrupted condition. This implies that subjects may be using temporal information from the distractor speech to segregate the target and distractor speech. This finding is in sharp contrast to the interrupted noise maskers mentioned above, a finding they replicate.

1.4.2 Overcoming Masking: Auditory Scene Analysis

Another prominent theory of how listeners overcome auditory masking comes from auditory scene analysis (Bregman, 1990). Auditory scene analysis is the binding of temporally and spectrally disparate acoustic information into cohesive auditory percepts (for a review see Bizley and Cohen 2013). These percepts are referred to as auditory objects and their formation is key in interacting with the auditory world. In a complex auditory scene there can be a dozen or more auditory objects present at once. Once distractor objects have been perceptually separated from the target object, listeners can selectively attend to the target, thus overcoming masking. The question of what mechanisms are recruited in this process has led researchers to examine the criteria necessary for formation and separation of these auditory objects.

Two very important variables in auditory stream segregation are time and attention, that is, auditory stream segregation is an online process that takes time to occur and is a process that must be attended to (Snyder et al., 2006; Shamma et al., 2011). It does not occur instantaneously, and if attention is shifted away from an auditory object, streaming rapidly resets and the build up of streaming must start all over again (Cusack et al., 2004). Another very important criteria is that of spatial location (Ihlefeld and Shinn-Cunningham, 2008a). The ability to localize sounds enables spatially separating the target from distractors, facilitating their segregation into distinct streams.

Auditory object formation allows listeners to isolate a target from a complex acoustic environment (e.g. those in Teki et al. 2013). While auditory attention is heavily implicated in this process (Carlyon et al., 2001), recent work suggests segregation may be robust to attentional demands (Masutomi et al., 2016). In contrast, the mechanisms downstream of segregation that track signal dynamics and construct a stream from them do rely on selective attention (Snyder et al., 2006; Petsas et al., 2016). Additional evidence suggests that these attentional demands are modality specific (Chait et al., 2012), supporting Baddeley and Hitch (1974)'s hypothesis of separate visual and auditory working memory systems (see §1.3.1 for more detail).

1.5 Functional Hearing Testing

Functional hearing tests are used in the clinic where they seek to model the adverse listening conditions encountered in daily life. These tasks simulate certain adverse listening conditions and then present spoken language (either sentences or words alone) and require the listener to report the words they hear. These tests are particularly useful at identifying functional hearing loss due to the increased difficulty, and subsequently decreased performance, this population faces when performing the task. As such, the tests represent a more ecologically valid approximation of the demands listeners face than traditional hearing tests. In all incarnations of the task, various cognitive mechanisms play a significant role in task performance and can even provide compensatory benefit in hearing impaired listeners. While much is known about the implicated compensatory mechanisms, they remain a critical target for study due to the possibility of their training to improve functional hearing outcomes.

It is these functional hearing tests that are the focus of this dissertation. Given the significant benefits they offer in predicting functional outcomes, mechanisms that jeopardize their utility should be identified and mitigated. The mechanisms identified in this dissertation are only a few possible confounds present in current testing procedures. Despite the current confounds, the paradigm is promising and if carefully administered can significantly improve the diagnostic power of clinical hearing testing.

1.5.1 The Importance of Functional Hearing Testing

The topic of speech perception in adverse listening conditions begs study for a few reasons. The most directly and immediately applicable to the public is the clinical advances that can be gained in the testing and training of those with normal hearing and hearing impairment (Burk and Humes, 2007, 2008). Difficulties with communication are associated with reduced quality of life, exclusion, depression, and severity of dementia symptoms (Bernabei et al., 2014). Populations that can benefit from these advances include the elderly (Prosser et al., 1990), children (Leibold et al., 2016; Hall III et al., 2002), those with cognitive differences(Dole et al., 2012; Calcus et al., 2015), and those with peripheral hearing impairment

(Goldsworthy, 2015; Moberly et al., 2014), or frequently those that fit into several of these categories. Studies on speech perception in adverse listening conditions have provided standardized testing practices for the field of audiology, such as the words-in-noise test (Wilson et al., 2007) or the Speech Understanding in Noise test (Paglialonga et al., 2013), to aid in diagnostic decisions such as whether to prescribe an assisted listening device or to help uncover more serious pathologies or neurological deficits.

In addition to the clinical knowledge gained from functional hearing tests, understanding functional hearing informs the creation of new technologies. The two main fields of technology that make use of these findings are the development of automatic speech recognition systems (including automated noise reduction algorithms) and the design of assisted listening devices ranging from auditory brainstem implants to hearing aids. Automatic speech recognition systems have become an extremely popular method of human computer interaction recently. While performance of these systems excels in ideal listening conditions, they still struggle to extract the target speech signal in noisy environments. Because of this, noise reduction methods have been a popular line of research for many years, often drawing inspiration from the human auditory system (e.g. Wang and Brown (2006)).

With respect to assistive listening devices, knowledge of how the auditory system functions is crucial to interfacing with it by means of technology. While incredible strides have been made in this field, many devices still struggle in noisy environments. Because of this, voluntary assistive devices, such as hearing aids, are sometimes elected against or left unused. Therefore, understanding how the healthy auditory system overcomes noise is crucial to these devices' function and in restoring the quality of life of those who depend on them.

1.5.2 Improvements over Traditional Hearing Testing

A main complaint of audiology patients is difficulty understanding speech in adverse listening conditions (Pekkarinen et al., 1990; Wiley et al., 1998). However, communication in adverse listening conditions can be impaired even in individuals whose hearing is within the normal range on standard audiological tests (i.e. "hidden hearing loss": Liberman et al. 2016; Hind et al. 2011). Because basic hearing profiles

are not deterministic of functional ability to communicate in complex environments, patients with mild to moderate hearing loss exhibit a wide range of functional outcomes (Divenyi and Haupt, 1997; Boxtel et al., 2000; Taylor, 2007; Cord et al., 2000). Furthermore, patients with identical audiograms can display disparate outcomes (Erdman and Demorest, 1998).

What underlies the variability in functional outcomes is currently not well understood and conventional audiological measures do not have adequate predictive power to diagnose or treat these patients. Cognitive abilities such as attention and tracking a distracting sound's temporal and spectral dynamics contribute in part to this functional ability (see §1.4.2 for more detail). Impairment in functional ability is not restricted to those individuals with hearing loss, extending to patients with age-related cognitive decline and other cognitive deficits. Due to the plastic nature of the auditory system (e.g. for higher-level auditory attention: Soveri et al. 2013), delineating the specific cognitive mechanisms underlying functional hearing can help identify and treat patients whose canonical hearing ability is within the healthy range. Similar targeted auditory training has previously been shown to retune plastic processes such as phoneme categorization (Lim and Holt, 2011).

1.5.3 Training to the Test

As discussed above, functional hearing testing includes several cognitive mechanisms capable of compensating for peripheral hearing loss. While these mechanisms are desirable to capture for determining functional outcomes, this is only true when they serve as faithful representations of the mechanisms used in daily life. While this may be reasonable to assume for mechanisms such as auditory attention, for online language processing mechanisms, patients' daily experience can differ greatly from the clinical environment. Therefore, patients whose linguistic experience is in a language variety similar to test materials have unknowingly been training for tests of functional hearing.

Capacity theory (see §1.3.2.1) posits that additional working memory capacity can afford participants access to abilities above and beyond other participants. These abilities enable them to accomplish additional tasks during language processing, decreasing difficulty and increasing accuracy. Connectionism,

on the other hand, posits that the observed performance benefit in some participants stems from their increased experience with the type of language comprehension being tested. Therefore, the mechanisms facilitating the task have received increased training in these participants compared to those with a relative lack of experience.

These differing accounts both adequately explain numerous behavioral findings regarding discrepancies in the parallel domain of reading comprehension. Participants who excel at the reading span test (Daneman and Carpenter, 1980) routinely outperform those who do not on other measures of reading ability. The first example is increased accuracy and speed of comprehension on complex sentences (King and Just, 1991). The second is the ability to resolve temporary ambiguity in reduced relative clauses (e.g. "*The experienced soldiers warned* about the dangers conducted the midnight raid": MacDonald et al. 1992). The third is the ability to use semantics and pragmatics to resolve temporary ambiguity (e.g. animacy, "*The defendant examined* by the lawyer shocked the jury." vs. "*The evidence examined* by the lawyer shocked the jury.": Ferreira and Clifton Jr 1986). The fourth is the distance at which participants can resolve pronominal reference (Daneman and Carpenter, 1980). The final is performance ability under extrinsic memory load (Baddeley et al., 1984).

To resolve the debate between capacity theory and connectionism, Reali and Christiansen (2007) attempted to find further evidence for the role of experience in language comprehension. A corpus analysis revealed object relative clauses to be more common when the clause contained a personal pronoun (e.g. "The lady that *I visited* enjoyed the meal."), but subject relative clauses to be more common when the clause contained an impersonal pronoun (e.g. "The studies that *motivated it* converged on similar results."). They hypothesized that if experience moderated comprehension ability, participants should perform better on the more common structures. However, if verbal working memory or innate linguistic constraints were at play, participants should behave similarly in the two conditions given their parallel structure. They observed a strong interaction between structure and pronoun type correlating with the corpus frequency of the constructions. This provided strong evidence for the role of experience in language processing.

Following the interaction effects found by Reali and Christiansen (2007), Wells et al. (2009) hypothesized that the role of experience in clause type processing could not only be correlated, but could be experimentally manipulated. The authors trained participants on relative clauses while a control group trained on parallel sentences whose structure had been converted to avoid relative clauses. The target group (but not the control) displayed a significant speed up in reading time for relative clauses following training compared to pre-training measurements. Additionally, this speed-up improved object relative clauses (the rarer kind) significantly more than subject relative clauses despite participants' equal training on both. The authors attribute this asymmetry to an increased effectiveness of each object relative training instance given their relative infrequency in naturalistic material. This provides strong experimental evidence for the role of experience in language processing.

In light of the evidence of experience's role in reading comprehension ability, a parallel phenomenon is likely to occur in functional hearing testing. Given that every participant's language experience is different, participants bring into the clinic an unimaginable variability in language processing ability. Functional hearing tests' reliance on linguistic stimuli are therefore simultaneously their best and worst feature. If variability in language experiences can be accounted for, the use of linguistic material becomes a clear advantage over traditional hearing testing in determining functional outcomes. Until then, however, participant language experience remains a persistent confound jeopardizing the validity of functional hearing testing.

1.6 Chapter Summary

Chapter 2 probes the question of which mechanisms release from energetic masking could utilize, given temporal regularities in the noise masker. We create and test a new paradigm by creating patterns of glimpsing windows. This introduces several complex timing structures into the noise signal, something previous studies have not done. This new masking paradigm controls several previous confounds, namely:

the categorically different sources producing energetic and informational masking, the variable signal-to-noise ratio and overall noise power between maskers, and the variable glimpsing windows afforded.

Chapter 3 seeks to situate chapter 2's novel maskers in the space between canonical informational and energetic masking. It further characterizes the effects of chapter 2's novel paradigm as well as tests its finding of cross-trial learning. It manipulates maskers in the spectral domain as well as the temporal domain and experiments with the timecourse of the effects observed in chapter 2.

Chapter 4 questions the role that participants' language backgrounds play in moderating their performance on clinical tests relying on language. We obtain proxies of language experiences using a self-report survey of the media participants report consuming and test participant performance on two language-based clinical tests: the Speech Perception in Noise task (Kalikow et al., 1977) and the reading span task (Daneman and Carpenter, 1980). To approximate the diversity of clinical populations, we recruit participants from both on- and off-campus populations which differ along dimensions a priori predicting differences in language exposure and use.

Finally I discuss the implications of this dissertation for the topics discussed in this introduction as well as those implicated in the three chapters mentioned above. I then discuss future directions for this line of research and conclude with a reflection on this body of work.

Chapter 2

Speech perception with temporally patterned noise maskers.[1]

2.1 Introduction

Speech perception must regularly overcome the noisy nature of the real-world contexts in which it occurs. Previous studies have delineated two distinct sources causing processing difficulty: interfering background noise, energetic masking, and distracting signals that are meaningful but irrelevant, informational masking (Pollack 1975, for review see Kidd et al. 2008). They differ, for example, in that the degree of interference from energetic masking is highly correlated with the level of background noise, whereas informational masking is only weakly related to the relative level of the competing signal (e.g. Rosen et al. 2013). When maskers are quiet relative to targets, informational masking is more difficult, as the additional information is distracting. However, at higher levels, the ability to treat the competing signal and target as distinct auditory objects enhances the ability to specifically attend to the target, making informational masking easier. Energetic and informational masking also seem to engage separate networks in neuroimaging studies (e.g. Scott et al. 2004).

One issue with characterizing the distinction between energetic and informational masking is that the typical paradigms for studying the two differ on many dimensions, preventing direct comparisons between them. Energetic masking is generally created with white or speech-shaped noise, whereas informational

[1] This chapter draws on work from: Courtland, M., Goldstein, L., and Zevin, J. D. Speech perception with temporally patterned noise maskers. In Prep. **and** Lander-Portnoy, M. (2016). Release from Energetic Masking Caused by Repeated Patterns of Glimpsing Windows. Interspeech 2016, 1672–1676. https://doi.org/10.21437/Interspeech.2016-1571

masking typically involves one or more additional talkers as an interfering mask (Mattys et al., 2012). As a result, any inferences we draw about informational masking and its difference from energetic masking are constrained to cases in which the competing information is also a linguistic signal. Thus, Scott et al. (2004) attribute their findings of increased Superior Temporal Gyrus activity during informational masking to the need to process more speech. Because informational masking stimuli are always linguistic, it is unclear what the relative contributions in overcoming informational masking are of language-specific processing (i.e. filtering unwanted speech by content, à la Treisman 1964b), or general auditory scene analysis (i.e. segregating masker and target streams using object-based attentional mechanisms, which would also be engaged to selectively attend to speech in an array of non-speech auditory objects, à la Bregman 1990).

To address this question, we create a patterned, non-linguistic stimulus to isolate the role of non-linguistic attentional mechanisms in informational masking. To allow direct comparison to energetic maskers, we construct maskers using speech-shaped noise correlated with the long-term average spectrum of our target words, and then produce temporal structure by inserting regularly patterned, non-isochronous, silences into the masker (see Sheft and Yost 2008 for a discussion of information v. structure). The complexity of the masker compared to previous isochronous studies (e.g. Miller and Licklider 1950) aims to recruit the same mechanisms that process the complex temporal regularities of an interfering speech signal in canonical informational masking. The regularity of the masker, on the other hand, differs from previous complex maskers (e.g. Festen and Plomp 1990) and provides the temporal structure present in stimuli designed to induce auditory stream segregation (Bregman, 1990). In contrast to previous studies, our maskers also precisely control for both global and local SNR, ensuring that performance differences arise solely from the presence or absence of information in the maskers.

We hypothesized that listeners would be able to take advantage of the regular temporal structure of the masking noise, providing a processing advantage over energetic masking. This was supported by work such as Andreou et al. (2011), which found that tone-on-tone maskers' temporal regularity facilitates target perception when tones repeat every 100-250ms and spectral separation alone is insufficient to segregate streams. The parameters of our maskers adhere to these criteria for facilitating target perception: the

maskers are interleaved with silences in 125ms windows, and their spectral profiles match the long-term average of the targets they mask. We predicted that speech perception would analogously improve when our maskers were temporally regular.

An initial experiment provided strong evidence for this hypothesis (Lander-Portnoy, 2016), but conceptual replications and a direct replication reported here (Experiments 2-4) did not produce significant differences in mean accuracy. In this paper, we examine measures of variability and patterns of correlation across conditions in both the initial and follow-up experiments. Interestingly, we find that these measures are very similar across experiments, suggesting that some as yet unidentified difference among participants is responsible for the systematic variability in performance we observe across conditions. We then discuss the robustness of our findings across experiments and identify mechanisms implicated in this systematic variability.

2.2 Experiment 1

In this experiment, we asked participants to identify words presented in the presence of maskers comprising patterns of speech-shaped noise interrupted by short silent intervals. The duration and timing of the silent intervals were irregular, and in a subset of the trials, was repeated twice before the presentation of the masked speech stimulus. We predicted that participants would be able to make use of the repeating pattern in the stimulus to improve their performance on the word recognition task.

2.2.1 Methods

2.2.1.1 Participants

Sixty undergraduate students at the University of Southern California participated in exchange for course credit. No personally identifiable information was gathered about participants. All participants had self-reported normal hearing and were monolingual speakers of English. None had previously participated in speech perception experiments in the lab.

2.2.1.2 Stimuli

The 108 target items were polysyllabic words selected from The Nationwide Speech Project Corpus (Clopper and Pisoni, 2006). Speakers were selected at random to include 9 speakers (4-5 female, 4-5 male) from each of the 6 United States dialect regions. The corpus consists of white 18-25 year old native speakers of English with no history of hearing or speech disorders.

Target items were manually extracted from the corpus' sound files using Praat (Boersma and Weenink, 2012) to allow precise control over onset timing of the target word. Targets played from the two second mark of the stimulus until the termination of the word, while masking continued until termination of the three second trial. Stimuli were masked by speech shaped noise (SSN) correlated to the long term average spectrum (LTAS) of the aggregate of all targets. A 100^{th} order linear predictor coefficient (LPC) filter was calculated from the concatenated targets. For each trial, 3 second white noise samples were generated and filtered using this 100^{th} order LPC filter to match its spectral profile to the LTAS. It should be noted that this spectrum differs somewhat from an individual target's LTAS. While matching SSN to target LTAS increases masker fit, it undesirably provides participants anticipatory spectral information about the upcoming target. After filtering, maskers' volumes were set at a signal-to-noise ratio (SNR) of -10dB relative to their targets.

To encode regular temporal information in the noise, we divide the noise into 8Hz windows (125ms each). In each window, we silence a portion of the previously constant masker, creating periods in which to glimpse the target unobstructed (see Cooke (2003) for Glimpsing). We then create 4 window types varying in the silence's duration and location within the window (see Fig. 2.1, left). The first of these window types contained 25ms (20%) silence at the beginning of the window, with the remainder SSN noise. Type 2 contained 100ms (80%) silence at the beginning. Type 3 and 4 contained 25ms and 100ms of silence respectively in the middle of the windows. We then randomly permute two of each window type and concatenate them to create a 1 second rhythmic pattern (for example, see Fig. 2.1, right). Given that

each pattern's information lies wholly in the ordering of windows and not in their composition or selection, the amount of noise power, global SNR, and total noise duration remain constant across patterns.

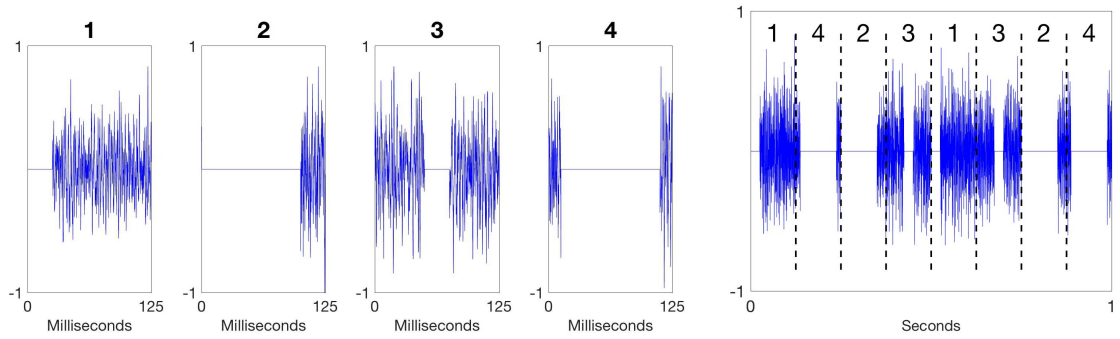

Figure 2.1: We utilize four different types of 125ms windows (left) to construct noise patterns (right). Windows 1 and 3 contain 20% silence (25ms) and 80% noise (100ms), while windows 2 and 4 contain 80% silence (100ms) and 20% noise (25ms). Additionally, in windows 1 and 2, the beginning of the silence is aligned with the beginning of the window, whereas in windows 3 and 4, the middle of the silence is aligned with the middle of the window. We then create a pattern by randomly permuting two of each window type (denoted at top). As windows are 125ms long, and each pattern contains 8 windows, patterns are 1 second long.

For each target, we create three conditions to occur in different lists: one that affords masking pattern prediction, one that does not but provides the same glimpses (i.e. local SNR) and forward and backward masking effects, and one with unrelated random patterns (see Fig. 2.2 for visuals). The first condition repeats the masking pattern twice in the preamble allowing participants to recognize it prior to the onset of the masked target in the third second. The second condition utilizes the same masking pattern as the first condition to control for glimpses it affords, but contains two random preamble patterns providing no anticipatory information. The third condition utilizes three different patterns that serve as a task performance baseline. The two random preamble patterns in condition two and all patterns in condition three only occur once per experiment and are thus novel when participants hear them. In contrast, the noise pattern that serves as masker in conditions one and two, which we call "Pattern A", recurs to allow cross-trial learning of its characteristics. Because condition one comprises a threefold repetition of Pattern A, we call this condition AAA. Because condition two contains two random patterns followed by Pattern A, we call this condition XYA. Lastly, because condition three contains three random patterns, we call this condition XYZ. Given that AAA contains all temporal information, XYA just cross-trial information, and

XYZ no information, we expect participants' performance in these conditions to correlate with the amount of available temporal information (i.e. $AAA > XYA > XYZ$).

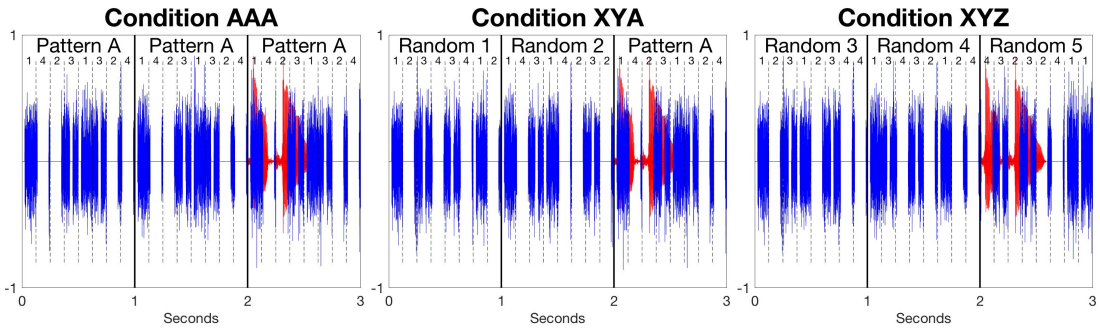

Figure 2.2: Wave forms showing the three experimental conditions with SSN in blue and target word in red. AAA (left) affords masking pattern information twice before target presentation. XYA (center) affords the same glimpsing windows as AAA, but no information about the upcoming masking pattern. XYZ (right) provides the same global SNR as AAA & XYA but contains no temporal regularity.

2.2.1.3 Procedure

We divided the 108 targets equally into each condition (36 each) and grouped trials into 4 blocks (27 each). Each block contained 9 triplets of one trial per condition permuted randomly to provide uniform exposure to each condition. This resulted in 12 lists (3x4 Latin square, 5 participants per) counterbalancing block and condition.

Testing occurred in a noise attenuating booth. Stimuli were presented over headphones in mono at a sampling rate of 44,100 Hz at 16 bits. Volume was set at a comfortable level consistent across participants. Participants typed their responses in a free response text box and advanced to the next trial when ready. Responses were automatically checked for correctness and those marked incorrect were checked by hand for misspellings. No partial credit was given. Initial separate scoring for inflectional errors and lemma identification proved uninformative, so we drop it from the analyses presented here.

2.2.1.4 Ancillary cognitive measures

Participants were also tested in measures of working memory and inhibitory control. The working memory task was an ordered digit span recall task in which digits (0-9) were presented sequentially on screen

for 250ms each, with a 750ms pause after each digit. Digits were sampled uniformly at random with replacement and a correct response consisted of the digits ranked in ascending order. Trials gradually increased in length from three digits to eight digits with five trials occurring per span length.

Performance on the Digit Span task was calculated as a simple percentage of total trials correct. No partial credit was given.

The inhibitory control task presented a square that was red or yellow on the left or right side of the screen. Participants clicked the left or right mouse button according to color and *not* screen position. Incongruent trials are those when correct mouse response and screen position are opposite.

Performance on the inhibitory control task is measured by the difference between the mean response time of correct congruent trials and correct incongruent trials. We exclude responses less than 200ms or more than 1500ms due to their probable cause by inadvertent click and distraction, respectively. We then exclude outliers further than 3σ away.

2.2.2 Results

2.2.2.1 Hypothesis Testing

Distributions and means for each condition are shown in Figure 2.3. To test for differences between conditions, we use a multilevel mixed-effects logistic regression model with accuracy as the dependent variable, condition as a fixed effect, and participant and item as random effects modeled as random intercepts.

The model reveals a significant main effect of condition, $\chi^2(2, N = 60) = 16.41, p < .0005$. Pairwise comparison between conditions indicate significant differences between AAA and XYA, $\beta = -.25, z = -3.36, p = .001$ as well as AAA and XYZ, $\beta = -.27, z = -3.67, p < .001$ but no significant difference between XYA and XYZ, $\beta = -.02, z = -0.30, p = .76$.

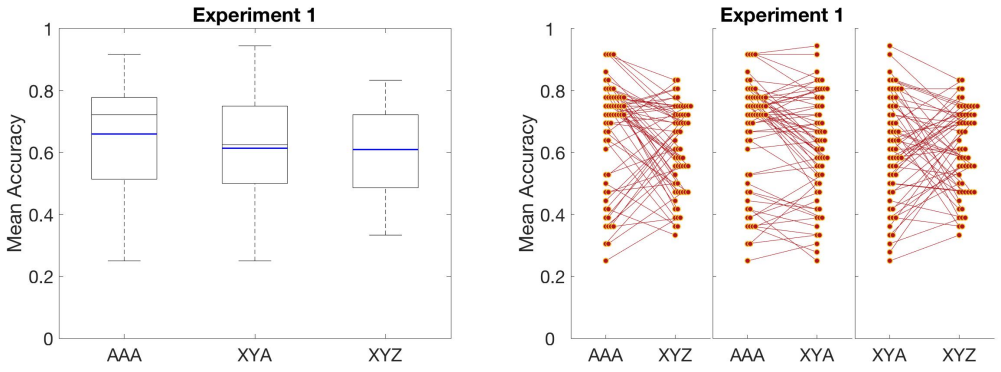

Figure 2.3: Results of the word recognition task performance by condition. Significant differences in sample means (blue, left) are found between AAA and XYZ as well as AAA and XYA, but not XYA and XYZ. Performance distributions (right) are represented as vertical histograms (i.e. histograms rotated clockwise 90°) with each circle representing a participant. These distributions show significant variance differences between AAA and XYZ, as well as XYA and XYZ, but not AAA and XYA. Lines connect participants' performances in each condition and visually represent correlations between conditions.

2.2.2.2 Exploratory Analyses

Following the unexpected results of several follow-up experiments, we reanalyzed the results of previous experiments to help explain our findings. These analyses are presented in the "Exploratory Analyses" sections throughout this paper.

Variance Differences As shown in Figure 2.3, large differences in variance were observed among the conditions. We tested differences in variance using Bartlett's Test (Bartlett, 1937), and found a significant difference between the variances of AAA, $s = .18$, and XYZ, $s = .14$, $T(1) = 3.99, p < .05$, as well as a marginally significant difference between XYA, $s = .17$, and XYZ, $s = .14$, $T(1) = 2.65, p = .10$. No difference was observed between AAA, $s = .18$, and XYA $s = .17$, $T(1) = .14, p = .71$.

Correlations between conditions As shown in Figure 2.4, performance is correlated across conditions. There is a strong correlation between performance in the AAA condition and performance in the XYA condition, $r(58) = 0.79, p < .0001$, whereas the correlation between AAA and XYZ, $r(58) = 0.32, p < .01$, and XYA and XYZ, $r(58) = 0.26, p < .05$, are moderate at best.

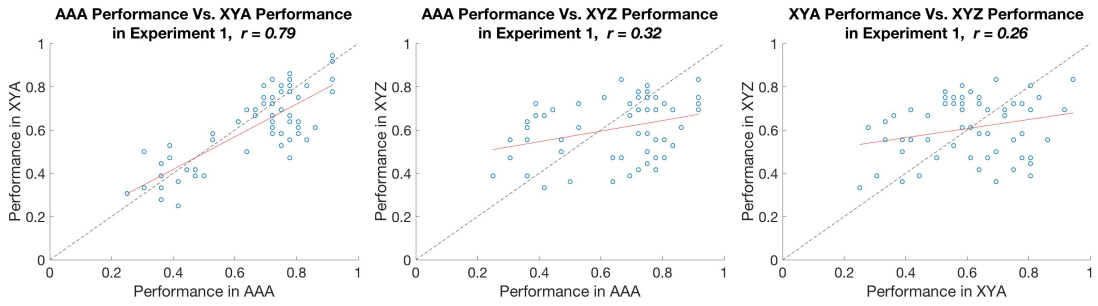

Figure 2.4: We find a strong correlation in exp. 1 between the conditions containing offline information but not between either of these and baseline. A least mean squares linear regression line is drawn on the plots in red and the identity line in black.

Using a Fisher z-transformation (Fisher, 1915; Lowry, 2019), we compare the correlations between conditions found above. We find AAA/XYA to be a significantly stronger correlation than AAA/XYZ, $z = 3.95, p < .0001$, and also significantly stronger than XYA/XYZ, $z = 4.30, p < .0001$. We do not find a significant difference in strength between AAA/XYZ and XYA/XYZ, $z = .35, p = .36$.

Finally, we perform hierarchical regression to test whether XYA explains additional variance in AAA performance beyond baseline (XYZ). We find that the model using both XYA and XYZ as predictors explains significantly more variance in AAA than the model using XYZ alone, $F(1, 57) = 85.89, p < .0001$. While the model using XYZ alone explains a significant amount of AAA variance, $F(1, 58) = 6.54, p = .01$, the addition of XYZ as a predictor does not explain significantly more variance than a model with XYA as the sole predictor, $F(1, 57) = 2.15, p = .15$. XYA also explains a significant amount of AAA variance alone, $F(1, 58) = 97.91, p < .0001$.

2.2.2.3 Cognitive Ability Measures

Inhibitory Control We observe a mean incongruence effect of 29ms (95% CI = $(22ms, 37ms)$). A t-test reveals this mean difference effect to be significant, $t(58) = 7.54, p < .0001$. Performance in the inhibitory task is not significantly correlated with overall performance in the word recognition task, $r(58) = -.21, p = .11$, nor with the benefit from repeated masking patterns (i.e. the difference between mean scores in the AAA and XYZ conditions), $r(58) = -.09, p = .52$.

Digit Span Participants' mean score for digit span recall was 76%, (95% CI $= (72\%, 79\%)$). We observe no significant correlation of digit span with overall word recognition performance, $r(58) = .11, p = .39$, nor with a difference score computed to indicate the benefit of repeating the masking pattern (condition $AAA - XYZ$), $r(58) = .04, p = .76$. Similarly, no evidence was observed for a relationship between inhibitory control and working memory capacity, $r(58) = -.05, p = .70$.

2.2.3 Discussion

Participants in this experiment were more accurate on trials when the temporal pattern was repeated (AAA) than in two control conditions. To foreshadow the results of the following studies, this difference in means was not reproduced in either conceptual or direct replications. Exploratory analyses, however, revealed that trials on which the stimuli were masked by the pattern repeated throughout the experiment (AAA and XYA) both produced increased individual variability, and further, that performance in these conditions was strongly correlated, explaining a significant amount of performance variability over baseline. Some participants improved relative to the baseline (XYZ) condition, whereas other participants performed more poorly when the masker was repeated. We collected working memory and inhibitory control measures, which have been shown to correlate with speech perception in noise in prior research (for review see Dryden et al. 2017; Akeroyd 2008), but these did not correlate with overall performance, nor with the repetition advantage. Nonetheless, the results suggest that some stable difference between participants makes the repetition of the noise pattern beneficial for some and detrimental to others.

2.3 Experiment 2

Based on the results of Experiment 1, we hypothesized two possible mechanisms responsible for the significant performance difference between the repeated condition (AAA) and the control condition (XYA). The first is a global perceptual grouping mechanism akin to stream segregation (Bregman, 1990) by which the noise pattern is perceived as distinct from the target speech. If the target speech is perceived as a

separate object, attending to it will increase perceptual ability and thus task performance. The alternative hypothesis is a local computational mechanism akin to beat induction and matching (Honing, 2012; Winkler et al., 2009) which calculates intervals between glimpsing windows so that attention can be allocated to the unmasked portions of the speech signal (Cooke, 2003). This experiment was devised to distinguish between these two hypotheses by interrupting streaming processes using short silences.

Auditory object formation is an online process that requires signal continuity and attention to construct a representation of a sound (Snyder et al., 2006; Shamma et al., 2011). When attention shifts away from the sound or the signal is interrupted by silence, the process rapidly resets and must begin anew (Cusack et al., 2004). We use this rapid resetting after interruption to defeat perceptual grouping by inserting silences between noise patterns. Because the silences are regular, the temporal predictability of glimpses remains unaffected. This selective affordance for local computation allows us to probe which phenomenon underlies the advantage observed in Experiment 1. If perceptual grouping is at play, we expect a null result. If local computation is at play, we expect to replicate Experiment 1.

2.3.1 Methods

Participants and Procedure

60 new participants were recruited from the same population as Exp. 1 in the same manner. The procedure is identical as well except for the removal of the inhibitory control and digit span tasks.

Stimuli

We generated stimuli in the same manner as Exp. 1 with identical window types, noise spectrum, and signal-to-noise ratio. In this experiment, however, we insert 433ms silences between noise patterns causing trials to last 3.866s rather than 3.0s. We refer to the conditions in the same manner as Exp. 1 but use the "_" character to denote a silence: A_A_A, X_Y_A, and X_Y_Z.

2.3.2 Results

2.3.2.1 Hypothesis Testing

We use the same mixed-effects logistic regression model with subjects and words as random intercepts as in Exp. 1 (see §2.2.2.1 for more detail). The model reveals no significant main effect of condition, $\chi^2(2, N = 60) = 3.08, p = .21$.

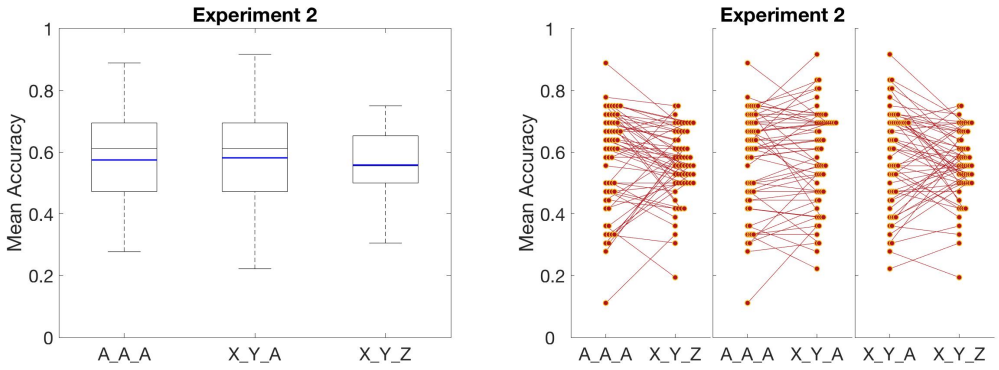

Figure 2.5: Results of word recognition task with inserted silences (denoted by "_"). We find no significant difference in sample means (blue, left) between A_A_A and either X_Y_A or X_Y_Z, but find a marginally significant difference between X_Y_A and X_Y_Z. Distribution variances are again significantly different between A_A_A and X_Y_Z, as well as X_Y_A and X_Y_Z, but not A_A_A and X_Y_A. Lines (right) again connect a participant's performances and represent correlation between conditions.

2.3.2.2 Exploratory Analyses

Variance Differences Bartlett's test (see Fig. 2.5 for visual), reveals a significant variance difference between A_A_A ($s = .16$) and X_Y_Z ($s = .11$), $T(1) = 6.56, p = .01$ and between X_Y_A ($s = .16$)and X_Y_Z, $T(1) = 6.24, p = .01$. No difference in variance was observed between A_A_A and X_Y_A, $T(1) < .01, p = .95$. Thus, as in Experiment 1, variance was greatest when the masking pattern was repeated throughout the experiment, whether it was repeated in the course of a single trial or not.

Condition performance correlations We observe a strong correlation between A_A_A and X_Y_A, $r(58) = .79, p < .0001$, as shown in Figure 2.6. Moderate correlations were observed between A_A_A and X_Y_Z , $r(58) = .51, p < .0001$ as well as between X_Y_A and X_Y_Z, $r(58) = .53, p < .0001$.

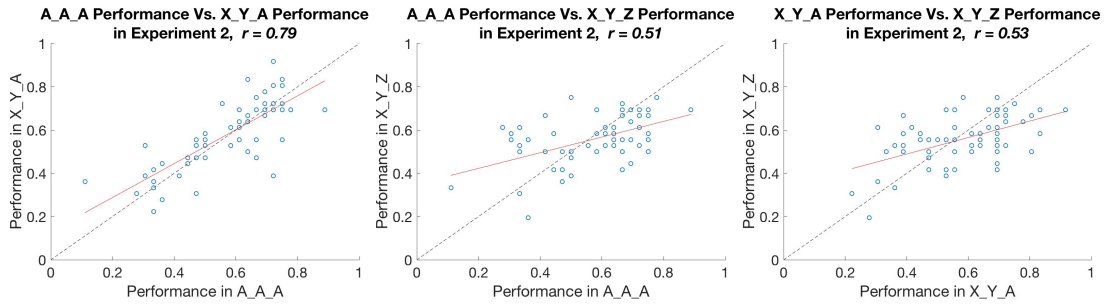

Figure 2.6: We again find a strong correlation between performances in the offline information conditions (A_A_A and X_Y_A) in Exp. 2. Interestingly, correlation of both A_A_A and X_Y_A with X_Y_Z increases relative to Exp. 1. Despite this, the correlation of offline conditions with each other is still significantly stronger than either with baseline. A LMS regression line is drawn in red and the identity line is drawn in black for reference.

We again test for significant differences in correlation strengths using Fisher's z-transformation. We find A_A_A/X_Y_A to be a significantly stronger correlation than A_A_A/X_Y_Z, $z = 2.74, p < .005$, and than X_Y_A/X_Y_Z, $z = 2.58, p = .005$. We do not find evidence of a significant correlation strength difference between A_A_A/X_Y_Z and X_Y_A/X_Y_Z, $z = -.16, p = .56$.

Finally, we again perform hierarchical regression to test whether X_Y_A explains additional A_A_A variance beyond baseline alone. Using both X_Y_A and X_Y_Z as predictors explains significantly more variance than using X_Y_Z alone, $F(1, 57) = 59.13, p < .0001$. Using X_Y_Z alone again explains a significant amount of variance, $F(1, 58) = 20.08, p < .0001$, but adding X_Y_Z does not explain significantly more variance than a model with X_Y_A as the sole predictor, $F(1, 57) = 1.74, p = .19$. X_Y_A alone again explains significant A_A_A variance, $F(1, 58) = 96.37, p < .0001$.

2.3.3 Discussion

We did not find differences in means across conditions. These data must be interpreted in the context of the failure to reproduce the effect of Experiment 1 in replications (Experiments 3 and 4, below), however. We cannot ascribe the lack of an advantage in the AAA condition to a disruption of streaming by the addition of silences. In contrast, we did replicate the pattern of change in variances in both the AAA and XYA conditions. We also replicated the strong correlation between performance in these conditions, further

supported by the hierarchical regression findings. This is further evidence that whatever mechanism is driving the individual differences in performance does not depend on streaming, and instead may be a result of long-term memory for the repeated "A" pattern.

2.4 Experiment 3

This experiment was designed based on results of the tests of a priori hypotheses from Experiments 1 and 2, and without the benefit of the exploratory analyses. We reasoned that the lack of an advantage for the AAA condition in Experiment 2 could be due to interruption of streaming, and sought to test whether object formation could be driven at the single trial level by repeating a masker that the participant has never encountered before.

Attending to objects and encoding them into memory allows participants to identify the object in future trials, thus facilitating task performance. This object persistence in memory with subsequent recall is attested visually in Hollingworth (2005)'s change detection task, where participants performed equivalently whether the object identification was online, in the next trial, or at the end of the experiment. This memory encoding represents features of the object and in relation to the object's context (Sun and Gordon, 2010). Encoding and retrieval of object representations in memory, and their strengthening with exposure, is strongly implicated in our experiments. Given our previous inconclusive results, we perform an experiment to delineate the role of isolated offline information in affording the initially observed advantage. If previous exposure to the object in question (pattern A) is necessary to segregate streams to defeat masking, then we expect higher performance only in the condition containing offline information. If, however, the mechanism relies solely on repetitive structure, both conditions containing online information should be distinct from baseline, but not each other.

2.4.1 Methods

Participants and Procedure

60 new participants were recruited from the same population and in the same manner as previous experiments. The word recognition in noise task proceeded the same as in previous experiments. Given the need to recognize and encode regular temporal structure in our experiments, we hypothesize that musical experience may be a good predictor of performance. We therefore include surveys of musical experience that cover musical training and performance, attitudes toward music, and music consumption behavior.

2.4.1.1 Stimuli

Maskers for this experiment were synthesized in the same way as previous experiments, with identical parameters, by inserting silences into long-term average spectrum speech shaped noise to create rhythmic patterns (see §2.2.1.2 for details). In this experiment, we include conditions AAA, containing both online and offline temporal information, and XYZ, a baseline containing neither, constructed as in Experiment 1. To test the isolated benefit of online temporal information, we also include a condition exhibiting only the repetition of noise patterns within a trial but not their reuse across trials. This condition contains threefold repetition of a novel pattern each trial which we denote with the variable X, giving us condition "XXX".

2.4.2 Results

2.4.2.1 Hypothesis Testing

The same model as Experiments 1 and 2 reveals no significant main effect of condition, $\chi^2(2, N = 60) = 2.07, p = .36$.

2.4.2.2 Exploratory Analyses

Variance Differences Bartlett's test reveals a marginally significant difference between AAA, $s = .16$, and XYZ, $s = .13$, $T(1) = 3.02, p = .08$, and a marginally significant difference between AAA and XXX

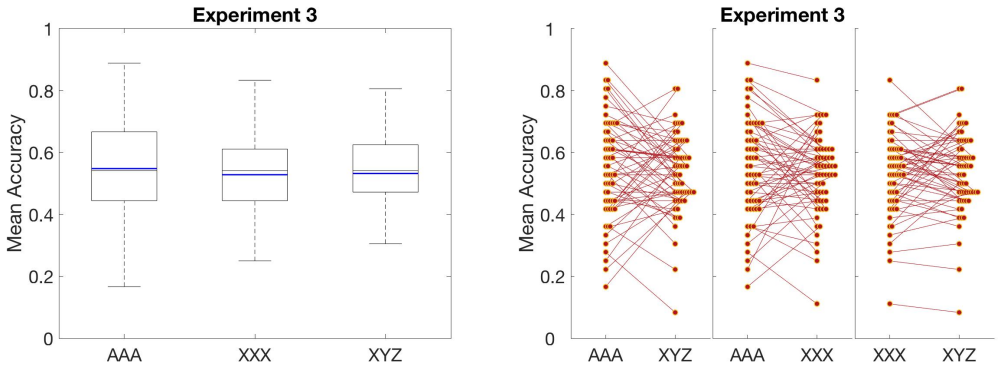

Figure 2.7: Neither the novel isolated online information condition, XXX, nor the previously significant online and offline condition, AAA, differ significantly in their sample means (blue, left) from baseline, XYZ, or from each other. The distribution variance of AAA differs significantly from both XXX and XYZ but the two do not differ from each other. Lines (right) connect a participant's performances and represent correlation between conditions.

$s = .13$, $T(1) = 2.69, p = .10$. No difference was observed between XXX and XYZ, $T(1) < .01, p = .92$.

Condition performance correlations We found weak to moderate correlations between performance in the AAA and XXX conditions, $r(58) = .38, p = .001$, and between AAA and XYZ, $r(58) = .29, p = .01$. A somewhat stronger correlation was found between XXX and XYZ, $r(58) = .56, p < .0001$, perhaps suggesting that variance in conditions where the masker is not repeated is driven by factors related to overall skill with word recognition in noise, whereas an additional factor is at work in conditions where the masker is repeated.

Fisher's z-transformation reveals XXX/XYZ to be a significantly stronger correlation than AAA/XYZ, $z = 1.75, p = .04$, but not significantly stronger than AAA/XXX, $z = 1.22, p = .11$. We find no significant difference in correlation strength between AAA/XXX and AAA/XYZ, $z = .52, p = .30$.

Hierarchical regression reveals that XYZ again explains significant AAA variance, $F(1, 58) = 5.51, p < .05$, but XXX explains significantly more when included, $F(1, 57) = 4.60, p < .05$. XXX itself explains significant AAA variance, $F(1, 58) = 9.86, p < .005$, beyond which including XYZ does not explain significantly more, $F(1, 57) = .65, p = .42$.

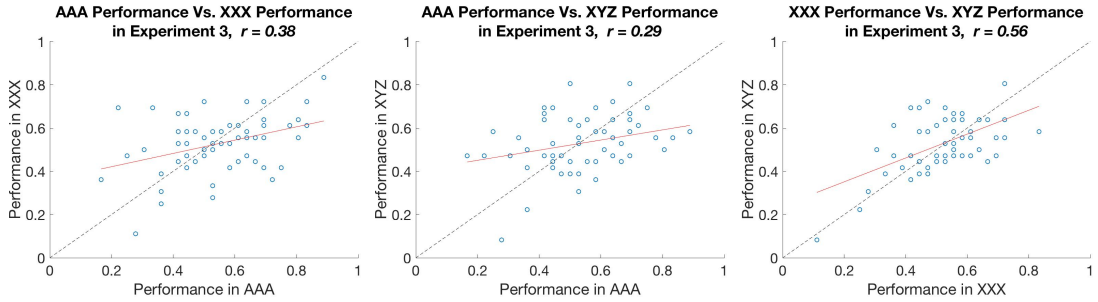

Figure 2.8: In general, we find weaker correlations in Exp. 3 than in previous experiments. The weak correlation between AAA and XYZ replicates Exp. 1's findings. XXX exhibits a modest correlation with AAA, though it is more similar to AAA's relationship with XYZ than with XYA (in previous experiments). XXX is moderately correlated with XYZ, though notably more weakly than previous correlations along offline informational content (AAA and XYA). Again, a least mean squares linear regression line is drawn in red and the identity line is drawn in black on the plots for reference.

2.4.2.3 Musical Experience Survey

We use three variables measuring participants' musical experience in our analysis: years of formal training, hours per day spent listening to music, and current involvement in music. These respectively capture the amount of learned skills, degree of passive exposure, and condition of acquired skills.

We find no significant correlation between overall task performance and years of musical training, $r(58) = .15, p = .12$, or hours per day spent listening to music, $r(58) = -.11, p = .80$. We also find no significant correlation in the difference between AAA and XYZ and years of formal training, $r(58) = -.05, p = .65$, nor average hours per day listening to music, $r(58) = -.15, p = .87$. Musical involvement was measured with a yes/no question and analyzed with a logistic classfier. The decision boundary maximizing accuracy simply classifies the data in one class, producing an accuracy of .77, its underlying class distribution, for both overall performance and AAA-XYZ difference, indicating no relationship with either.

2.4.3 Discussion

Repetition of the masker within a trial appears to have little impact on performance in the word recognition task, based on the lack of an effect on the means and the lack of any detectable effect on variances.

Trials containing a stimulus repeated over the course of the experiment, in contrast, differed (although marginally) from both of the other conditions in the variance observed, and was only weakly to moderately correlated with the other conditions, suggesting that the repetition of the masker over the course of the experiment may be more important in driving individual differences. It is interesting to note that XXX explains significantly more AAA variance than baseline alone, despite their mutual lack of offline information. While this hints at a possible effect of online information, the effect is not strong enough to drive significant findings in any of the other tests of central tendency or variance we perform in this paper (Exp. 1 mean effect notwithstanding). Lastly, musical experience did not correlate with with overall word recognition performance nor with differences between the AAA condition and the XYZ baseline condition. Two conditions in this experiment (AAA and XYZ) were identical to Experiment 1, but did not produce the same difference in means, prompting us to conduct a direct replication of Experiment 1.

2.5 Experiment 4

Experiment 4 is an attempt at a direct replication of Experiment 1.

2.5.1 Methods

The experiment's design, stimuli, and procedure are identical to Exp. 1's (see §2.2.1 for details) except the substitution of a musical experience survey for the tests of working memory and inhibitory control. 60 new subjects were recruited from the same population and in the same manner as previous experiments.

2.5.2 Results

The same mixed-effects logistic regression model from Exp. 1 (see §2.2.2.1 for details) reveals no significant mean difference between AAA and XYA, $\beta = -.01, z = -0.21, p = .84$, and only a marginally significant difference between AAA and XYZ, $\beta = -.12, z = -1.79, p = .07$. No significant difference was observed between XYA and XYZ, $\beta = -.11, z = -1.58, p = .11$.

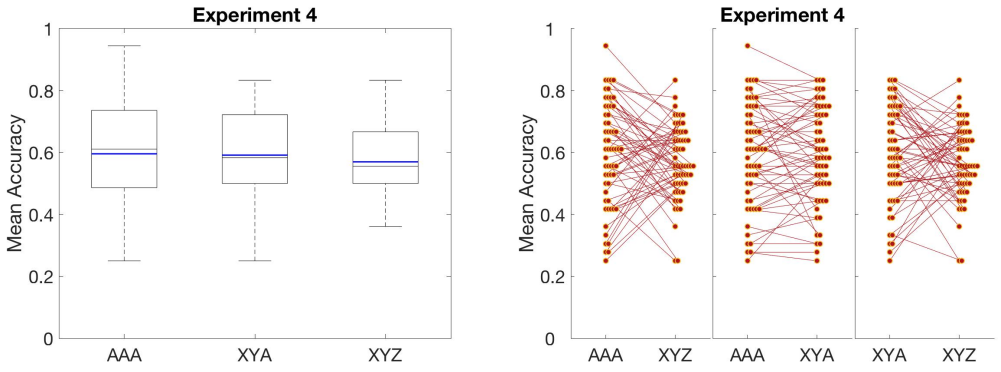

Figure 2.9: Neither previously significant difference between condition means (blue, left) replicates. The online and offline condition, AAA, does not differ significantly from the offline only condition, XYA, and differs only marginally from baseline, XYZ. The lack of significant difference between the latter two conditions replicates from Exp. 1. The distribution variances of AAA and XYA both differ significantly from XYZ but not each other, replicating the findings of Exp. 1 and 2. Lines (right) again connect participant performances and represent correlation between conditions.

Variance Differences Bartlett's test reveals the variance of AAA, $s = .17$, to be significantly greater than that of XYZ, $s = .12$, $T(1) = 6.61, p = .01$, and the variance of XYA, $s = .16$, to be significantly greater than XYZ as well, $T(1) = 4.33, p = .04$, as shown in Figure 2.9. The variances of AAA and XYA do not significantly differ, $T(1) = .25, p = .62$.

Condition performance correlations As in Experiment 1, AAA performance is a strong predictor of XYA performance, $r(58) = .7, p < .0001$, (Fig. 2.10). No significant correlations are observed between XYZ and AAA, $r(58) = .11, p = .20$, and between XYZ and XYA, $r(58) = .07, p = .29$.

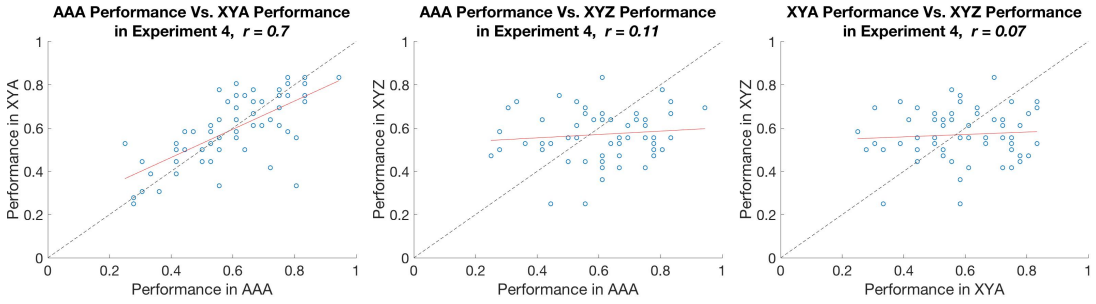

Figure 2.10: Exp. 4 replicates the strong correlation between conditions with offline information (AAA and XYA). The finding that both AAA and XYA are poor predictors of baseline performance (XYZ) also replicates. A least mean squares linear regression line is drawn in red and the identity line is drawn in black on the plots for reference.

We replicate the correlation findings of Experiment 1 and again find AAA/XYA to be a significantly stronger correlation than AAA/XYZ, $z = 4.00, p < .0001$, as well as XYA/XYZ, $z = 4.20, p < .0001$. We do not find a significant difference in the strength of correlations between AAA/XYZ and XYA/XYZ, $z = .20, p = .42$.

We replicate the hierarchical regression of Experiment 1 and again find XYA to explain significant variance in AAA performance, $F(1, 58) = 54.55, p < .0001$, and significantly more variance than XYZ alone, $F(1, 57) = 53.02, p < .0001$. Here, XYZ again does not explain significantly more variance than XYA, $F(1, 57) = .38, p = .54$, and moreover fails to explain a significant amount of variance alone, $F(1, 58) = .70, p = .41$.

Musical Experience

We replicate Exp. 3's findings (see §2.4.2.3) that amount of musical training is a poor predictor of AAA benefit, $r(58) = .17, p = .10$. Additionally, average daily listening time is again a poor predictor of AAA benefit, $r = -.03, p = .60$ as well. Current musical involvement is also uninformative as a predictor of AAA benefit, once again returning all the data as one class and achieving the underlying class distribution, .7, as its accuracy.

2.5.3 Discussion

When considering differences in measures of central tendency, a very different pattern was observed here than in Experiment 1, where the condition in which maskers were repeated within the trial (AAA) differed from both a condition that comprised all novel maskers (XYZ), and a condition in which the masker was repeated over the course of the experiment (XYA). In Experiment 4, performance did not differ significantly between AAA and XYA, and neither differed significantly from XYZ, although they did differ numerically. This failure to replicate prompted the exploratory analyses reported throughout the paper, i.e., comparisons of the variance across conditions, and the correlation among conditions. These analyses produce similar results in Experiment 1 and the replication: conditions in which the masker is repeated

throughout the experiment (AAA and XYA) produce greater variability between participants, and performance in these conditions is strongly correlated, whereas performance in the condition in which the masker is not repeated is only weakly correlated with both. This is supported throughout by hierarchical regression analyses, which are most strongly seen here: baseline performance does not explain significant variance in AAA performance but XYA both explains significant variance and significantly more than baseline.

2.6 General Discussion

Our initial hypothesis was that participants would be able to use repeating temporal information within a trial (online information) to overcome masking and improve their performance in a word recognition in noise task. Although initial analyses of measures of central tendency supported this hypothesis in Experiment 1, the overall evidence does not support this hypothesis. As seen in Table 2.1, the advantage for the AAA condition over other conditions observed in Experiment 1 is not observed consistently across experiments. In contrast, analyses of differences in variance and correlations between conditions produce similar patterns across experiments (Tables 2.3 and 2.4, respectively). Specifically, conditions in which the masker was repeated throughout the experiment (AAA and XYA) were consistently found to increase variance when compared with conditions in which the stimulus was not repeated (XYZ), or was repeated only within a trial (XXX). Further, in all three experiments in which both conditions were included, strong correlations were observed between the AAA and XYA conditions. This is supported by hierarchical regression which repeatedly finds XYA to explain significantly more variance in AAA performance than baseline alone.

On balance, our findings suggest that the availability of recurring information across trials (offline information), rather than repeating information within trials (online information), is important in determining performance in each condition. Direct tests designed to evaluate the contribution of streaming within a trial, as when silences were inserted between maskers in the preamble in Experiment 2, or the repetition of the masker within a trial, as in Experiment 3, did not produce evidence for the use of online

information. Further, we note that in Experiment 3, the "A" masker was repeated less frequently than in the other experiments, and in that experiment we observed the smallest effect sizes for comparisons between the AAA condition and others, in terms of both differences in the mean and the variance, as shown in Tables 2.2 and 2.3. Experiment 3 presents another noteworthy finding – the correlation between conditions that lack offline information (XXX and XYZ), while sizeable, is much smaller than between those that contain it (AAA and XYA, see Table 2.4). This correlation may be weakened due to the restricted range caused by the decreased variance we observe in these conditions (Nikolić et al., 2012; Alexander et al., 1984; Aguinis and Whitehead, 1997).

Performance Means by Condition

	AAA	XYA	XXX	XYZ
Exp. 1	.66	.61	-	.61
Exp. 2	.58	.58	-	.56
Exp. 3	.55	-	.53	.53
Exp. 4	.60	.59	-	.57

Table 2.1: Mean performance by condition appears fairly variable and unreliable across experiments. Empty cells indicate the condition did not occur in that experiment.

Cohen's d Across Experiments

	Exp	AAA	XYZ
XYA	1	.26	.03
	2	-.04	.17
	4	.03	.16
XXX	3	.14	-.03
XYZ	1	.31	-
	2	.12	-
	3	.11	-
	4	.18	-

Table 2.2: Cohen's d reveals that normalized mean *differences* also vary greatly across experiments. The XYA vs. XXX comparison is absent as the two were mutually exclusive. Together with Table 2.1, these results indicate that central tendency is not a robust indicator in our experiments.

We collected ancillary data about people's working memory and executive function abilities (Experiment 1) and about their musical training and experience (Experiments 3 and 4), but found no evidence for correlations of these measures with the size or direction of the effect of repeated information in the masker on word recognition. One goal of future work with these phenomena will be to identify other sources of variability that contribute to the increased variance in performance when maskers are repeated throughout

Standard Deviations by Condition

	AAA	XYA	XXX	XYZ
Exp. 1	.18	.17	-	.14
Exp. 2	.16	.16	-	.11
Exp. 3	.16	-	.13	.13
Exp. 4	.17	.16	-	.12

Table 2.3: Standard deviation by condition is relatively stable, in contrast with condition means. These results, and their levels of significance, provide evidence that variance is a robust metric, grouping conditions along the basis of offline information: AAA with XYA and XXX with XYZ. Empty cells indicate the condition did not occur in that experiment.

Condition Correlations

	Exp	AAA	XYZ
XYA	1	.79	.26
	2	.79	.53
	4	.70	.07
XXX	3	.38	.56
XYZ	1	.32	-
	2	.51	-
	3	.29	-
	4	.11	-

Table 2.4: Condition correlations are reliable across experiments in their strength (AAA & XYA) or weakness (AAA & XYZ and XYA & XYZ). These robust findings, as well as Exp. 3's weak correlation of XXX with AAA and moderate correlation with XYZ, support the offline information account provided by variance (see Table 2.3). Exp. 2's higher correlations for AAA & XYZ and XYA & XYZ may be due to the silences between patterns, making it not directly comparable to other experiments. The XYA & XXX correlation is absent as the two conditions were mutually exclusive.

an experiment. For example, offline information necessitates a long-term memory representation of the object, given the time-course of the experiment and the inability to rehearse noise to keep it active in working memory (Demany et al., 2001; Kaernbach, 2004). Furthermore, this memory storage must represent pattern A alone so that the speech with which it co-occurs in previous trials does not interfere with its application in further trials, analogous to the separate storage of each voice in polyphonic music (Fujioka et al. 2005, for review see Trainor et al. 2014). Given the many dimensions along which our maskers differ from the target speech (timbre, envelope, spectral constancy, etc.), it is quite possible that they are treated as a separate "voice" – or at least a distinct object – for the purposes of memory storage.

Given the shape of AAA's distribution (see Fig. 2.3), we hypothesize that the increased variance may be due to differential effects of repeating the masking pattern on two distinct subsets of participants. Individuals may differ in their ability to form distinct memory representations for simultaneously presented

objects (words and masking patterns), so that for some participants the memory of the masker improves performance, whereas for others, memory of the prior presentation of the masker produces proactive interference. A possible mechanism for this differential account comes from load theory (Lavie 1995, 2000, for review see Lavie 2010, 2005). Under Load theory, when cognitive load remains below a person's capacity, they process all information automatically and non-selectively. It is only when load exceeds capacity that selection occurs and processing is narrowed to the object of focus. As it relates to our task, participants whose cognitive capacities are not met by task demands automatically process all information in the stimuli – both the speech and the noise. However, participants whose capacities are exceeded by task demands will necessarily constrain attention to only the target task of speech perception – disregarding the information contained in the noise. In this account, below-capacity participants can devote resources to capitalizing on the information present in AAA and XYA, while above-capacity participants are likely to be adversely affected by additional surplus information they are unable to make use of.

Given the Load theory account presented above, this individual variability is likely to occur along the lines of cognitive capacity. It is thus surprising that the working memory and executive function measures we obtained were not predictive of offline information benefit. This is likely explained by the visual modality of the ancillary tasks, which differs from the auditory modality in capacity and processing (Cohen et al., 2009). Additionally, Cohen et al. (2011) found that musicians' superior auditory memory did not extend to the visual modality, implying that we did not measure the dimension of interest for an auditory task like ours. Our results support Stenbäck et al. (2015)'s findings that similar measures of executive function and working memory were uncorrelated with each other and parallel conditions in their word recognition task (cf. Rönnberg et al. (2010); Sörqvist (2010)). Given musicians' superior auditory memory (Cohen et al., 2011), musical experience also represented a good candidate for an informative covariate. However, our measurements of music education, involvement, and consumption all proved to be uninformative with regards to task performance. This may be explained in part by Honing et al. (2009)'s findings that beat induction and rhythm tracking mechanisms are innate, and are not in fact preferentially present in musicians over non-musicians (cf. Thompson et al. 2015). While we measured musicianship

as a proxy for beat tracking ability, we plan in future studies to directly assess this ability and correlate it with task performance. This is motivated by findings that both beat tracking and beat production ability moderate speech perception in noise performance (Slater and Kraus 2016; Slater et al. 2018), possibly by mitigating the effect of backward masking (Tierney and Kraus, 2013). These effects were only present for sentences (not words) however, so a modification in our target stimuli is needed.

Chapter 3

Exploring the gap between informational and energetic masking with select parameter manipulations.[1]

3.1 Introduction

Speech perception becomes a challenging task when performed in a noisy environment. Decades of previous research have discovered two canonical sources of difficulty presented by different types of noise. The first, energetic masking, is caused by a steady stream of random noise interfering with proper reception of the desired speech signal at the periphery of the auditory system (e.g. at the basilar membrane: Scott et al. 2004). Energetic maskers are devoid of informational content and thus represent noise in the traditional sense. In contrast to this, informational masking, the second type, represents an informative signal in its own right, but one that is not the desired signal in a particular context. This additional information is distracting and interferes with selective processing of the target signal at higher levels of the auditory pathway (Dirks and Bower, 1969). Although informational maskers may overlap spectro-temporally in the periphery incidentally causing energetic masking (Brungart et al., 2006), this is not necessary and their damaging effects may be purely at higher levels of processing.

Despite the rigorous delineation of energetic masking and informational masking (e.g. for review see Mattys et al. 2012), research bridging the gap between the two is fairly rare. While the two masker types

[1] This chapter draws on work from: Courtland, M., Goldstein, L., and Zevin, J. D. Speech perception with temporally patterned noise maskers. In Prep.

are traditionally treated as categorically different phenomena, the two can actually be thought of as two extremes in opposite corners of a parameter space (akin to two opposite corners of a unit square in 2-d parameter space). Conceived of in this way, intermediary maskers can be generated at other corners of the parameter space that exhibit the properties of one or the other type of masker. This allows careful experimental manipulation along dimensions of interest in the parameter space, affording the ability to characterize what effect each dimension (or subset of dimensions) has on masked speech perception and the auditory system in general. This addresses the issue that the two masking types differ along many dimensions, making it hard to determine which dimensions cause which observable differences. These differences include: differential sensitivity to signal-to-noise ratio (Brungart, 2001), different patterns of neural activation (Scott et al., 2004), and differential sensitivity to attentional demands (Leek et al., 1991), among others.

With an incremental walk through masking parameter space, we can begin to probe which differing properties of the two masking types cause the different behavior observed in previous studies. These isolated manipulations of individual masker properties allow for the delineation of masking at previously infeasible levels of specificity. This paper represents an exploration into several novel portions of the parameter space and documents the effect of each on a word identification in noise task. We initially undertook an isolated manipulation of temporal regularity in Courtland et al. (2019). Here, we build on our findings using variations on the stimuli created for that paper.

3.2 Experiment 1

Given the results of Courtland et al. (2019)'s experiments 1 and 2, we attributed our findings to Auditory Scene Analysis (ASA, Bregman 1990) and its ability to provide release from informational masking (for review see Leibold 2012). Here, we test this hypothesis by leveraging several findings about the limitations of ASA in the spectral domain. The most relevant finding is that in order to construct an auditory object from the components present in an acoustic signal, the components must all be within a certain distance in

the frequency domain (a "critical band"). Components that are too spectrally distant will not bind into a single object and thus cannot leverage ASA to provide release from masking. In addition to the need for a constrained frequency range, acoustic components must regularly repeat to bind into an auditory object and provide release from masking (Bregman, 1990).

In this experiment, we vary the spectral profiles of our maskers with both a low-variance (i.e. within a critical band) and high-variance (i.e. outside a critical band). Additionally, as in Courtland et al. (2019), we vary whether the masker repeats or is random. In contrast to experiments 1 and 2 of Courtland et al. (2019), the repeated masker does not recur across trials here. This eliminates a possible confound of learning the recurring pattern but should not affect ASA, which does not depend on recurrence. We then predict that repeated conditions whose components are within a critical band will be able to be bound into an auditory object providing masking release, while those whose components are outside will not and thus will be more effective maskers. We additionally predict that the lack of repetition in both high- and low-variance non-repeated conditions will prohibit the formation of auditory objects also making them more effective maskers than the repeated low-variance masker. Lastly, because neither non-repeated masker types should bind into auditory objects, the low-variance masker should provide more energetic masking than the high-variance one as its spectral profile will more closely overlap with the target speech.

3.2.1 Methods

3.2.1.1 Participants and Procedure

We recruited 60 USC undergraduates to participate in the experiment in exchange for course credit. All were monolingual English speakers and had self-reported normal hearing. None had previously taken part in this type of experiment in our lab. We collected no personally identifiable information from subjects.

Participants performed the word recognition task on a desktop computer in a noise-attenuating booth. Stimuli were presented in 16-bit mono at a sampling rate of 44.1kHz over headphones at a comfortable volume constant across participants. Participants responded using a free-response text box and advanced

to the next trial when ready. Responses were hand-checked for misspellings and no partial credit was awarded.

3.2.1.2 Stimuli

Stimuli are described at length in Courtland et al. (2019), but we offer a brief overview here. The 108 target words for the word recognition task come from Clopper and Pisoni (2006)'s polysyllabic word list. We randomly selected speakers balanced across gender and regional dialect. Clopper and Pisoni (2006)'s speakers are white 18-25 year-old native English speakers with no hearing or speech pathologies.

We create the maskers by synthesizing 3-second samples of white noise and filtering it so its spectrum matches the long-term average spectrum (LTAS) of all the target speech samples. We then segment the LTAS noise into 125ms (8Hz) sections and insert a period of silence into each section. These silences are either long (100ms) or short (25ms) and aligned to either the left-edge or center of the 125ms window. This 2x2 manipulation yields 4 window types from which we make 1-second patterns by randomly permuting 2 of each window type. In this experiment, the 3-second maskers either contain a threefold repetition of a novel pattern, a condition we call XXX (X representing a variable pattern novel across trials), or 3 random novel patterns, a condition we call XYZ (X, Y, and Z each representing different and novel patterns). After the spectral manipulation described below, the maskers are normalized to a -10dB signal-to-noise ratio (SNR). Crucially, because the experimental manipulations only change the order of presentation of the masking windows, the global SNR of our maskers remains the same across conditions. To create the stimuli, we present targets beginning at the 2-second mark of the patterned masker. All targets last less than 1 second and thus end before the end of the masker.

As the aim of this experiment is to observe the effect of spectral variance on masking ability, we undertake the task of dynamically manipulating the masker's spectrum over the course of a stimulus. To do this, we alter the spectrum of each 125ms window (the LTAS of targets, described above) by adding a gaussian curve to its spectrum. The curves have a bandwidth of 250Hz and peak power of 10dB. To select the centers of each gaussian, we sample a normal distribution centered at 1,000Hz with a standard

deviation of 250Hz for the low-variance condition and 1,000Hz for the high-variance condition. We then generate the window's noise using this new spectrum and insert a silence according to its window type (as described above).

3.2.2 Results

3.2.2.1 Hypothesis Testing

Given the binary nature of whether participants' responses were correct, we use a logistic regression model. To account for variable difficulty between target items, we include a random effect of item. We nest this within subjects whose differing baseline ability we account for with an additional random effect.

We first include a fixed effect of condition yielding our initial multilevel mixed-effects logistic regression model. This model reveals no significant effect of condition on task performance, $\chi^2(3, N = 60) = 1.17, p = .76$. We then substitute a fixed effect of repeated or random (i.e. XXX vs. XYZ) and find that this model also reveals no significant effect on performance, $\chi^2(1, N = 60) = .18, p = .67$. Finally, we use the degree of spectral variance (i.e. high or low) as a fixed effect and do not find a significant effect of this either, $\chi^2(1, N = 60) = .99, p = .32$.

3.2.2.2 Exploratory Analyses

Following several unexpected experimental results presented here and in Courtland et al. (2019), we performed additional analyses to help explain what we had observed. These additional post-hoc analyses are reported in the "Exploratory Analyses" sections of this paper.

Variance Differences To test whether individual variability might cause condition variances to differ significantly (as in Courtland et al. (2019)), we use Bartlett's test (Bartlett, 1937). Bartlett's test reveals no significant variance differences between any conditions, $T(1) \leq .40, p \geq .53$ (see Figure 3.1 for visual).

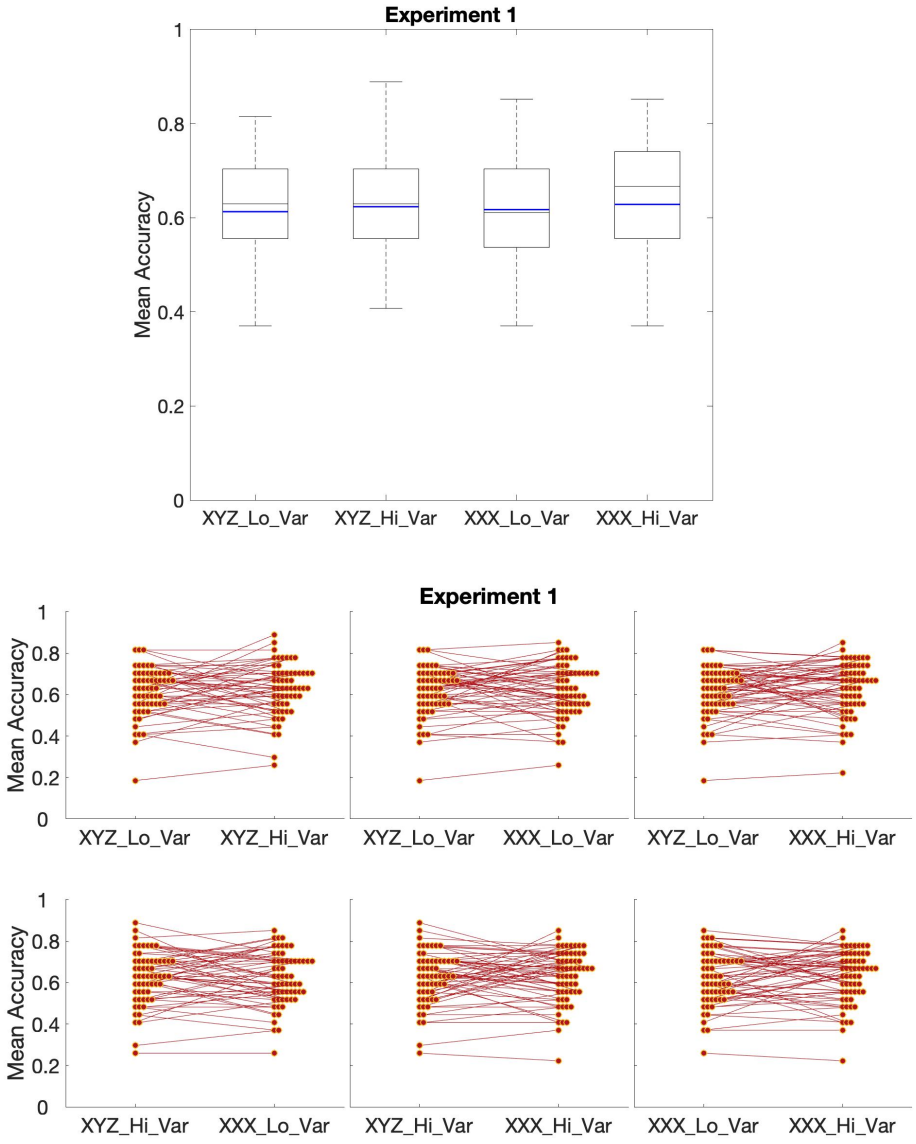

Figure 3.1: The results of the word recognition task reveal no significant difference in means across conditions (blue, above). Performance distributions (vertical histograms, below) show no significant difference in variance across conditions. Lines connect participant performances in each condition.

Condition performance correlations As in the experiments in Courtland et al. (2019), we measure correlations between conditions to test whether performance is related across any conditions (see Figure 3.2 for visual). We find significant correlations across all conditions, $r(58) \geq .5, p < .0001$. Given the

significant correlations between conditions, we test whether any significant differences exist between the strengths of the correlations (Lowry, 2019; Fisher, 1915), which would imply that certain conditions are more closely related than others. We find no significant difference between the strengths of any correlations, $|z| < .73, p > .23$ for all comparisons.

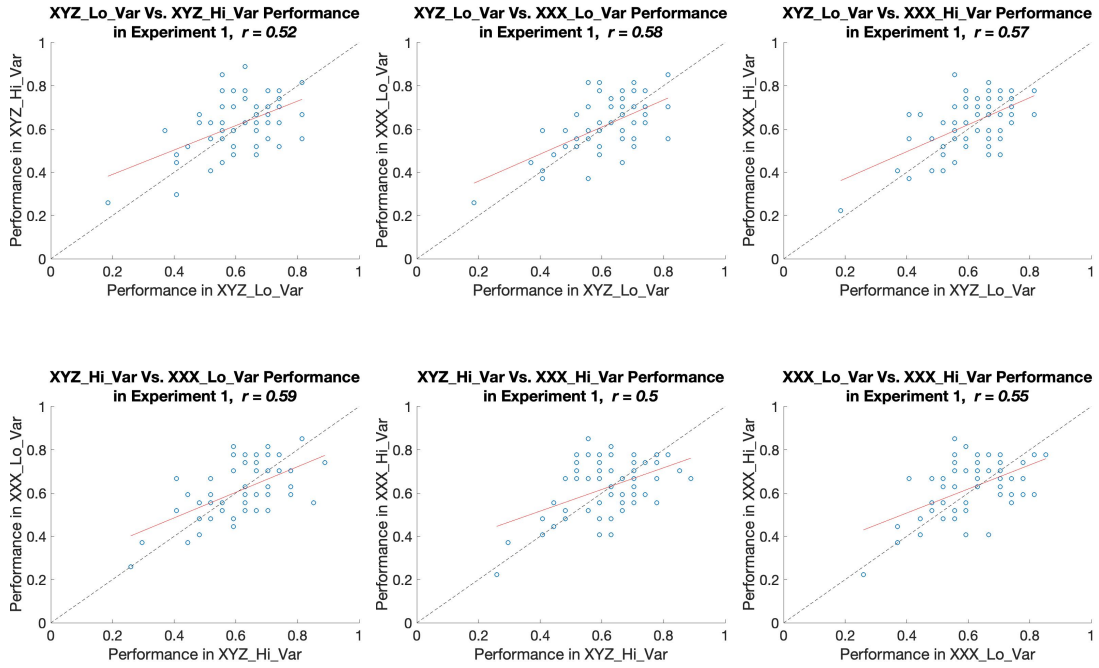

Figure 3.2: We find significant moderate correlations in performance between each pair of conditions in experiment 1. We find no evidence for differences in the strength of these correlations between comparisons, however. A least mean squares linear regression line is drawn on the plots in red and the identity line in black for reference.

3.2.3 Discussion

Despite our initial hypothesis of a complex interaction between spectral variance and temporally repeated structure, we find no experimental evidence supporting this hypothesis. Additionally, we find no evidence that either manipulation alone has a significant effect on task performance. While it is possible that our manipulation of spectral profile was not extreme enough to elicit an effect, it is worth noting that the difference in spectral quality between windows is highly perceptible.

The significant correlations we observed imply that performance is related across all conditions. The lack of significant *difference* in correlation strengths, on the other hand, implies that this is likely due to

participants' baseline task ability, rather than individual variability in response to experimental manipulations. The strengths of the correlations observed here are strikingly similar to that observed between conditions XXX and XYZ of the subsequent experiment 3 in Courtland et al. (2019). Those conditions lack any spectral manipulations – all windows contain unmodified LTAS noise – which suggests that spectral manipulations do not alter the relationship between conditions XXX and XYZ.

One possible account for the absence of any effects in this experiment – neither mean effects, nor variance effects, nor differences in correlation strengths – is the lack of recurrence of the repeating masking pattern across trials. In the first 2 experiments of Courtland et al. (2019), the threefold repeated pattern recurs across trials, yielding condition AAA (A representing a constant recurring pattern). This account is supported by the lack of effects in experiment 3 of Courtland et al. (2019), which is equivalent to the initial experiment save for a lack of recurrence. It is also possible, however, that the increase from 3 conditions used previously to 4 conditions did not provide enough observations per condition to detect any effects of the experimental manipulations.

3.3 Experiment 2

Given the lack of significance in experiment 1, we simplify the previous experimental manipulation to provide greater statistical power from more observations per condition. To do this, we abandon the novel spectral variance manipulation and test only whether our original manipulation of temporal regularity remains under spectrally dynamic maskers. Because we aim to test the original manipulation, we again use a recurrent as well as repeating masker. We therefore manipulate only whether or not the masker repeats and include just the narrow spectral variance conditions from experiment 1. This narrow variance should allow only the repeated masker to bind into an auditory object providing release from masking compared with the non-repeated masker.

3.3.1 Methods

Participants and Procedure

30 new participants meeting the same criteria as in §3.2.1.1 were recruited in the same manner from the USC undergraduate population. Testing proceeded in the same manner as well.

3.3.1.1 Stimuli

Given the derivative nature of this experiment, stimuli were quite similar to those described in §3.2.1.2, with the notable elimination of the high variance condition and recurrence of the repeating pattern across trials. This limits stimuli to two conditions – repeated recurrent patterns, AAA, or novel random patterns, XYZ – both with 10dB gaussians added to each window's spectrum. In contrast to the previous experiment, all added gaussians were centered at locations sampled from a normal distribution centered at 1,000Hz with a 250Hz standard deviation.

3.3.2 Results

3.3.2.1 Hypothesis Testing

We use the same multilevel mixed-effects logistic regression model as in experiment 1 with condition as a fixed effect. We observe a marginally significant effect of condition on performance, $\chi^2(1, N = 30) = 2.72, p < .10$, which manifests as a marginally significant pairwise comparison between AAA and XYZ, $\beta = .14, z = 1.65, p < .10$.

3.3.2.2 Exploratory Analyses

Variance Differences We test whether the conditions have significantly different variances using Bartlett's test (see Figure 3.3 for visual). We observe a marginally significant difference between the variances of AAA ($s = .144$) and XYZ ($s = .106$), $T(1) = 2.59, p = 0.108$, which equates to a small effect size, $d = .18$ using Cohen's d (Cohen, 1977).

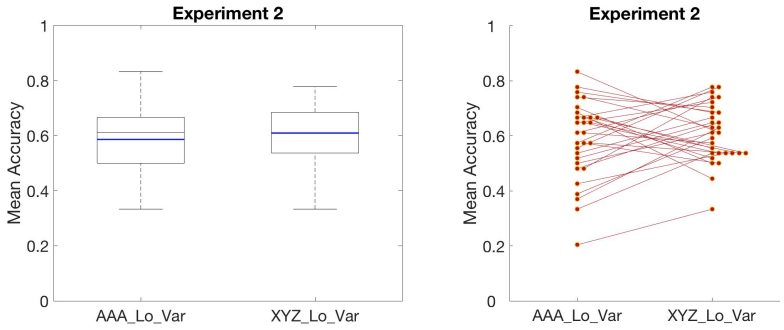

Figure 3.3: We observe a marginally significant difference in performance means (blue, left) across experiment 2's conditions despite the spectral manipulations we introduce. We also observe a marginally significant difference in variance across conditions (vertical distributions, right) supporting the previous results of Courtland et al. (2019).

Condition performance correlations We test for a correlation between conditions and observe only a marginally significant correlation between the two, $r(28) = 0.27, p = 0.077$.

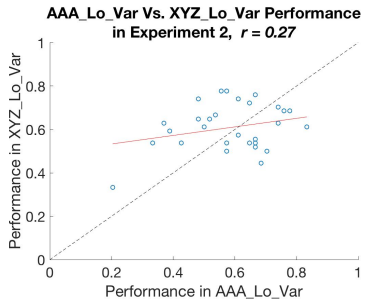

Figure 3.4: We find a weak marginally significant correlation in performance between the conditions of experiment 2. The correlation is weaker than that between the equivalent conditions in experiment 1. A red LMS line and black identity line are included for reference.

3.3.3 Discussion

The results suggest that the increase in statistical power and use of recurrence achieved their aim, resulting in a marginally significant effect of temporal regularity on performance. The direction of the finding, however, is surprising and opposite the highly significant finding initially observed in Courtland et al. (2019). Given that mean effects were not robust across experiments in that paper and that the results here are only marginally significant, it is possible that these findings are spurious. However, it is also possible that the spectral variance present in the maskers here defeats any benefit gleaned from recurrence

and repetition. The marginality of the effect here may also be due to the decrease in sample size from 60 participants to 30. The weakening of the significant correlation between this experiment's conditions compared to what we observed in the same conditions of experiment 1 is also likely due to this decreased sample size. It remains unclear though, why participants would perform better in XYZ in this experiment when in all previous experiments they have performed equivalently to or worse than AAA.

It is noteworthy that the variance effects reach marginal significance here. This finding supports our account in Courtland et al. (2019) that pattern recurrence is a driving factor in determining individual variability. The low correlation we observe is further evidence of the difference between these two conditions. Given the equivalence of the variance and correlation results between this experiment and those in Courtland et al. (2019), it seems that offline effects are invariant under small spectral manipulations.

3.4 Experiment 3

The results of experiments 1 and 2 bring into question our initial attribution of masking release in Courtland et al. (2019) to Auditory Scene Analysis. Here, we test the hypothesis that our repeated maskers behave more like informational maskers than energetic maskers. To do this, we leverage the finding that energetic maskers are highly sensitive to SNR while informational maskers are only loosely related to SNR (Brungart, 2001). Given this, we present the non-repeating and repeating maskers at both high and low SNRs to test whether the non-repeating maskers – the supposed energetic maskers – are more sensitive to SNR variation than the repeating maskers – the supposed informational maskers.

3.4.1 Methods

Participants and Procedure

60 new USC undergraduates meeting the same criteria as in §3.2.1.1 were recruited in the same manner. Testing proceeded in the same manner as well.

3.4.1.1 Stimuli

Stimuli were created using the method described in Courtland et al. (2019) (summarized in §3.2.1.2 before the spectral manipulations). In this experiment, in addition to the manipulation of whether patterns repeated, XXX, or were random, XYZ, maskers were normalized to a high-SNR, -5dB, or low-SNR, -15dB. This yields 4 conditions in total: high-SNR XXX, low-SNR XXX, high-SNR XYZ, and low-SNR XYZ.

3.4.2 Results

3.4.2.1 Hypothesis Testing

Given the parallel nature of the data, we use a similar multilevel mixed-effects logistic regression model to those in previous experiments. The model with a fixed effect of condition reveals a significant effect of condition on task performance, $\chi^2(3, N = 60) = 337.95, p < .0001$. Pairwise comparisons between conditions reveal significant differences between XXX_hi_SNR and XXX_lo_SNR, $\beta = -.97, z = -11.89, p < .001$, between XXX_hi_SNR and XYZ_hi_SNR, $\beta = .25, z = 3.11, p < .005$, and between XXX_hi_SNR and XYZ_lo_SNR, $\beta = -.91, z = -11.15, p < .001$. The model also reveals a significant difference between XXX_lo_SNR and XYZ_hi_SNR, $\beta = 1.22, z = 14.77, p < .001$, but not between XXX_lo_SNR and XYZ_lo_SNR. Lastly, the model reveals a significant difference between XYZ_hi_SNR and XYZ_lo_SNR, $\beta = -1.16, z = -14.06, p < .001$.

The difference observed above between SNR levels is supported by a model using a fixed effect of SNR level. This model reveals a significant effect of SNR level, $\chi^2(1, N = 60) = 329.54, p < .0001$, which corresponds to a significant difference between the high SNR and low SNR conditions, $\beta = -1.06, z = -18.15, p < .001$.

The observed difference between temporally regular XXX and random XYZ is also significant in a model using temporal regularity as a fixed effect, $\chi^2(1, N = 60) = 7.12, p < .01$, corresponding to a significant difference between XXX conditions and XYZ conditions, $\beta = .15, z = 2.67, p < .01$.

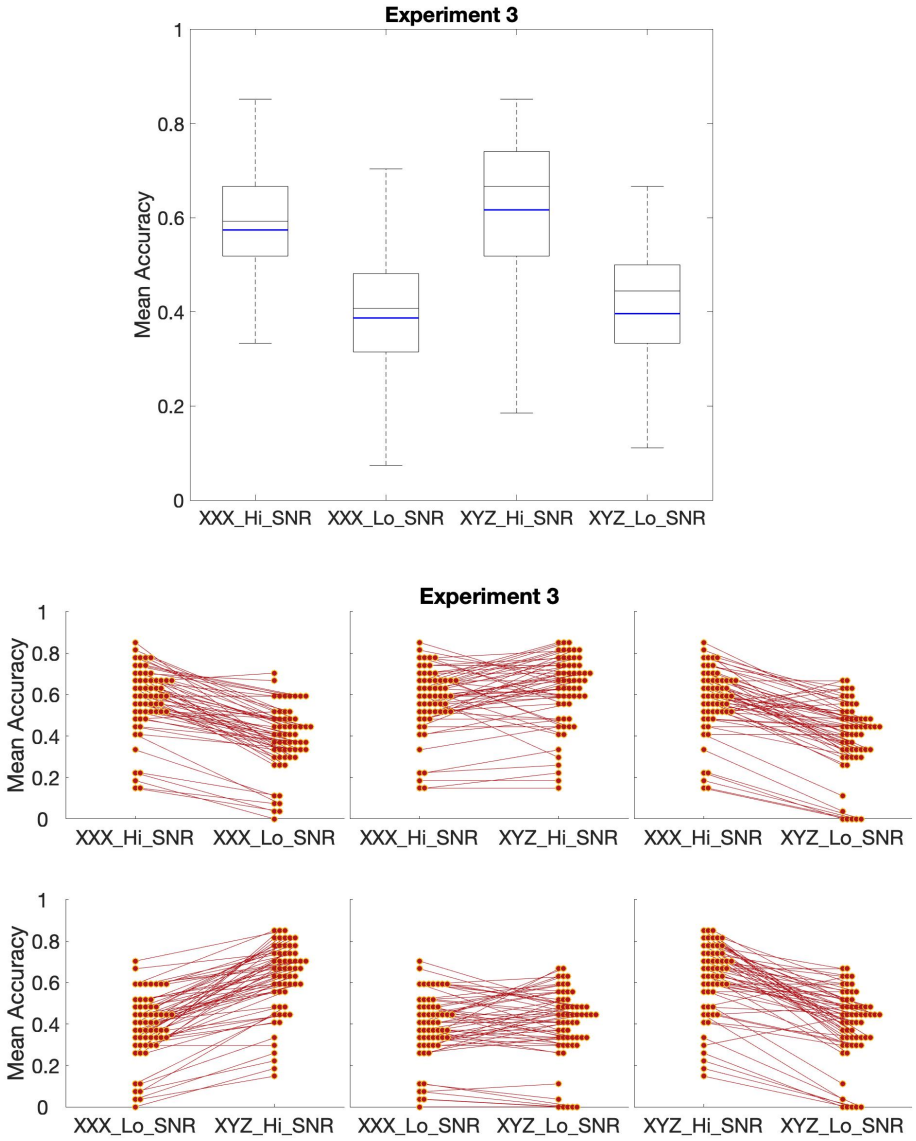

Figure 3.5: We observe significant differences in performance across the board between high SNR and low SNR conditions. We also observe a significant difference between performance in the two high SNR conditions, but not their low SNR counterparts. We find no significant differences in variance across the conditions.

3.4.2.2 Exploratory Analyses

Variance Differences We again use Bartlett's to test for a significant difference in variances across conditions (see Figure 3.5 for visual). We observe no significant difference between any conditions' variances, $T(1) \leq .44, p \geq .50$ for all comparisons.

Condition performance correlations To probe the relationships between conditions, we test correlations in performance across each pair of conditions. We observe significant correlations between performances in all conditions, $r(58) \geq .58, p < .0001$ for all comparisons (see Figure 3.6 for individual pairwise comparisons). As in §3.2.2.2, we also test whether any of the correlations found differ significantly in their strength. While most of the correlation strength comparisons are insignificant, two emerge as marginally significant. The first is the difference between XXX conditions and XYZ conditions: XXX_Hi_SNR/XXX_Lo_SNR ($r = .71$) Vs. XYZ_Hi_SNR/XYZ_Lo_SNR ($r = .58$), $z = 1.27, p = 0.10$. The second is between the low SNR conditions compared to the XYZ conditions: XXX_Lo_SNR/XYZ_Lo_SNR, $r = .74$, Vs. XYZ_Hi_SNR/XYZ_Lo_SNR, $z = 1.6, p = 0.06$.

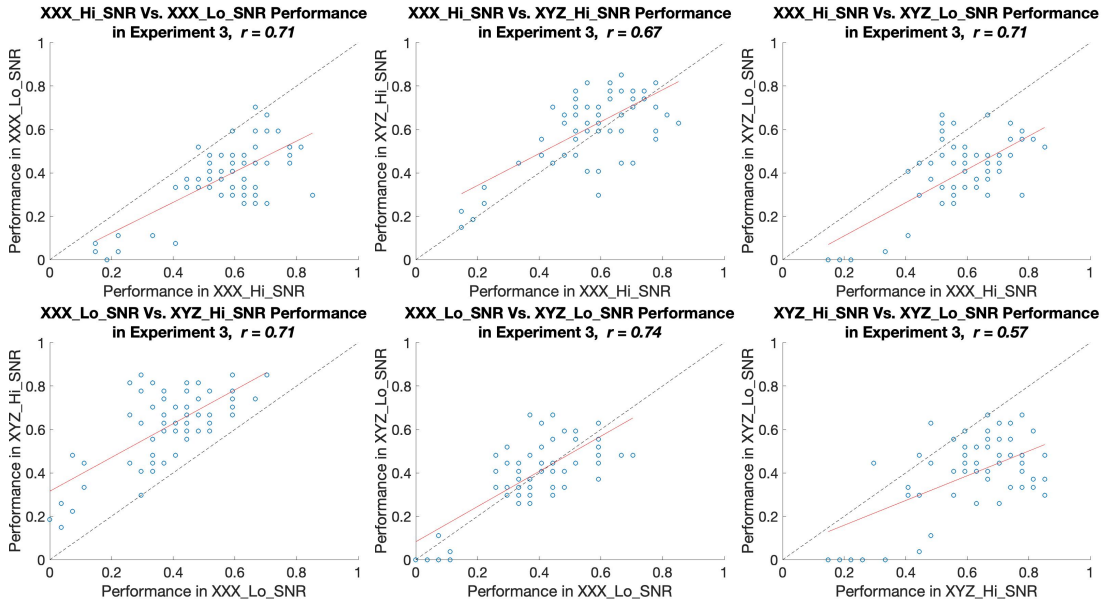

Figure 3.6: We find strong significant correlations between all conditions of Experiment 3. We find a marginal difference in the strength of correlation between the low SNR conditions and the baseline conditions, as well as between the XXX conditions and the baseline conditions. An LMS line is drawn in red and the identity line in black for reference.

3.4.3 Discussion

The significant main effect of SNR we observe here implies that all our maskers are behaving as energetic maskers despite the glimpsing windows they provide. The presence of the effect across both the XXX and XYZ conditions implies that our predicted interaction of temporal regularity and SNR did not occur.

In light of this evidence, it appears that temporal regularity does not alter the dependence on SNR that is characteristic of energetic masking.

One surprising finding is that in the high SNR conditions, performance significantly improves from XXX to XYZ. This is not predicted by our hypothesis and is in contrast to previous findings from Courtland et al. (2019) which found no evidence of a significant difference between the two at -10dB SNR. It is additionally surprising as the effect is not present in the low SNR conditions. Given that the low SNR condition means are .39 (XXX) and .40 (XYZ), this absence is not likely to be due to floor effects. It is possible the effect occurs due to something akin to Load theory (Lavie, 1995). In this account, given the relative ease of perception in the high SNR conditions, additional temporal information is unnecessary to overcome masking. This unnecessary information then captures cognitive resources away from the main task, detracting from word identification performance.

The strong correlations that occur with equivalent strength across most conditions likely capture baseline task ability rather than individual variability in how participants are affected by experimental manipulations. The exceptions to this equivalence, however, are noteworthy. The marginally significant difference in the strengths of XXX conditions' correlation and XYZ conditions' correlation implies that the ability of participants to detect temporal regularity is stable across SNR levels. In light of the findings above that XXX performance is worse in high SNR conditions than XYZ, this is perhaps best framed as those who automatically process the regularity do so across SNRs regardless of the fact that this may inhibit their performance on the task. Additionally, the marginally significant difference between the low SNR conditions compared with the XYZ conditions implies that individual susceptibility to SNR may moderate task performance above simple baseline ability.

3.5 Experiment 4

After all the experiments and analyses of Courtland et al. (2019) had been performed, it became clear that while the effect of online information on overall task performance had only reached significance in

the initial experiment, all the experiments showed mean effects in the right direction. Given that auditory streaming and pattern recognition unfold over time with increased exposure to the stimulus (Cusack et al., 2004), we hypothesized that longer exposure to maskers prior to target presentation might provide additional benefit. We therefore provide participants with 4 preamble patterns – providing 4 pre-target repetitions of the masking pattern in the repeated condition – to test whether significant mean effects will reemerge.

3.5.1 Methods

Participants and Procedure

The participants and procedure are the same as in previous experiments with 60 participants in this study.

3.5.1.1 Stimuli

The stimuli are created in the same manner as in Courtland et al. (2019) (§3.2.1.2 before spectral manipulations), with the exception that we present 4 masking patterns prior to target presentation giving us 5 masking patterns in all. Given the fivefold repetition, the repeating condition is now condition AAAAA, the glimpsing control condition is now condition VWXYA, and the baseline condition is now condition VWXYZ.

3.5.2 Results

3.5.2.1 Hypothesis Testing

The same multilevel mixed-effects logistic regression model as in previous experiments reveals a significant main effect of condition, $\chi^2(2, N = 60) = 17.63, p = .0001$. Pairwise comparison between conditions shows a marginally significant difference between AAAAA and VWXYA, $\beta = -.11, z = -1.63, p = .10$. The model also reveals significant differences between AAAAA and VWXYZ, $\beta = -.29, z = -4.16, p < .001$, and VWXYA and VWXYZ, $\beta = -.17, z = -2.53, p = .01$.

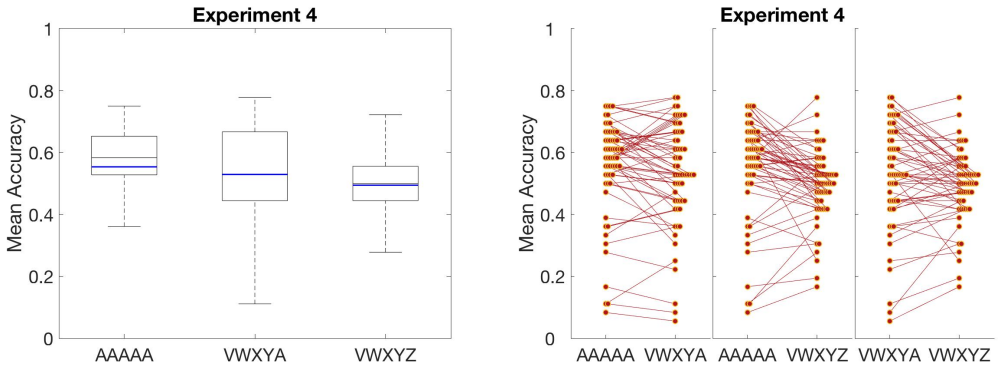

Figure 3.7: We observe significant mean differences (blue, left) between the baseline condition and both conditions containing pattern A. We also observe a marginally significant difference between the repeated and recurring condition over just the recurring condition. Additionally, we observe significant variance differences (see vertical distributions, right) between the conditions containing pattern A and the baseline condition.

3.5.2.2 Exploratory Analyses

Variance Differences We again test for unequal variances using Bartlett's, which reveals a significant variance difference between AAAAA ($s = .159$) and VWXYZ ($s = .115$), $T(1) = 5.98, p = .01$, a small effect size, $d = .43$ (Cohen, 1977). We also observe a significant variance difference between VWXYA ($s = .168$) and VWXYZ, $T(1) = 8.05, p < .005$, a small effect size, $d = .24$. We find no significant difference between the variances of AAAAA and VWXYA.

Condition performance correlations As before, we test for correlations between conditions. We find significant strong correlations between all conditions, $r(58) \geq .64, p < .0001$ (see Figure 3.8 for individual correlations). We also test for differences in correlations strengths and find a significant difference between the conditions with pattern A as the masker, AAAAA/VWXYA ($r = .83$), and between the repeated condition and baseline, AAAAA/VWXYZ ($r = .64$), $z = 2.31, p = .01$. We also find a marginally significant difference between the conditions with pattern A as the masker and between the recurring condition and baseline, VWXYA/VWXYZ ($r = .71$), $z = 1.6, p = .055$. We find no significant difference between correlation strengths of either pattern A condition with baseline (i.e. AAAAA/VWXYZ Vs. VWXYA/VWXYZ), $z = -0.71, p = 0.76$.

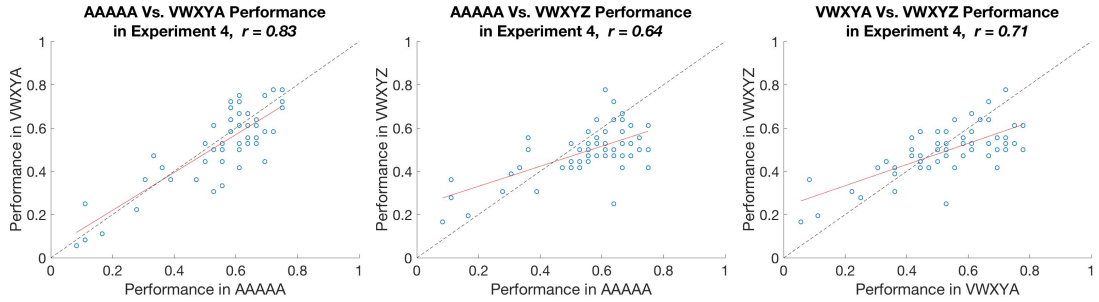

Figure 3.8: We observe a very strong correlation between conditions with pattern A as a masker. The correlations observed here are stronger than others in this paper as well as those in Courtland et al. (2019). The strength of the correlation between pattern A conditions is significantly stronger than between either and baseline. A red LMS line and black identity line are included for reference.

3.5.3 Discussion

The significant mean difference effects we observe here replicate the effects initially observed in Courtland et al. (2019). The differences between conditions with pattern A as the masker and the baseline condition corroborate the account presented in Courtland et al. (2019) that offline information – the recurrence of pattern A across trials – plays a crucial role in determining task performance. This is further supported by the variance differences we observe here, replicating the robust variance effects found in Courtland et al. (2019). The variance differences imply that individual variability plays a role in the conditions containing pattern A beyond simply baseline task ability. This account is supported by the difference in correlation strength between those conditions with pattern A as their masker and either of them with the baseline condition.

Additionally, the resurfacing of a significant difference between AAAAA and VWXYA provides additional evidence that online information plays a role in task performance above and beyond offline information. Of particular note is that the fourfold preamble repetition of the masking pattern used here replicated the initial mean-effects finding that two replications of twofold repeated maskers failed to produce in Courtland et al. (2019). This strengthening of effect with increased number of repetitions implies that the process unfolds and strengthens over time given increased exposure to masker patterns (both within and

across trials). This again hints at auditory object formation and stream segregation (Cusack et al., 2004; Bregman, 1990).

3.6 General Discussion

In this work, we examine various properties of the novel maskers created in Courtland et al. (2019) with a particular focus towards situating how they pattern with respect to canonical energetic masking and informational masking. To this end, we leverage several previous findings regarding differences between the two canonical maskers including restriction to a "critical band", dependence on SNR, and build-up over time. While the results of the previous paper suggested that the new maskers might pattern with informational masking along several dimensions, it seems that the maskers mostly behave as energetic maskers. The results of experiments 1 and 2 suggest that spectral variance does not play a moderating role in task performance and does not interact with temporal regularity. Experiment 3 suggests that the maskers are highly dependent on SNR which also does not interact with temporal regularity. The notable exception to our maskers behaving as energetic maskers is listeners' deriving benefit from a temporally regular structure which increases with exposure to a repeating pattern. Experiment 4 replicates our initial findings that temporal regularity plays an important role in task performance even beyond temporal recurrence. In experiment 4, we also observe the first evidence of mean effects – supporting previous variance and correlation effects – showing the significant and robust role temporal recurrence plays in task performance.

Several findings of this work support our previous account of individual variability in response to pattern A's recurrence across trials. The first of these is experiment 1's equivalent correlations across all conditions. All the correlations from this experiment are of the same magnitude as those between XXX and XYZ in experiment 3 of Courtland et al. (2019). While the previous paper provided a noticeable difference between the XXX/XYZ correlation and AAA/XYA correlations, there was only one instance of an XXX/XYZ comparison on which to draw conclusions. The additional correlations observed here provide

more indirect evidence that the relationship between these conditions is weaker than the relationship between conditions containing pattern A. This implies that response to pattern A moderates performance in a common manner across these conditions providing more information than baseline performance alone. This is further supported by the weakness of correlation between AAA and XYZ in experiment 2, one of the weakest observed between these two conditions. Lastly, experiment 4's mean effects provide strong support for the difference between the pattern A conditions and baseline. This comes from both a replication of the previously observed AAA vs. XYZ difference as well as the first evidence of mean effects from just the recurrent condition. Experiment 4's variance differences further display the effect of pattern A's recurrence on variability – replicating our previous findings. The difference in strength between experiment 4's AAAAA/VWXYA correlation and either condition with VWXYZ provides further evidence in favor of this account.

A detailed account of how pattern A may moderate variability is presented in Courtland et al. (2019), but a synopsis will be provided here. We attribute the difference in individual response to additional information contained in the masker to differences in cognitive resources. Under this account, all participants perceive the additional information yet only some have the additional resources to devote to processing this information (à la Load Theory: Lavie (1995)). Participants who have these surplus resources are able to make use of the additional information contained in the masker to provide release from masking relative to baseline. Those who do not are hurt by the additional distracting information present in the masker relative to baseline. While this accounts for many of the findings we observe here and in the previous paper, it still cannot account for why XYZ performance is better in experiment 2 than AAA. Considering the smaller sample size, the marginality of the effect, and its conflict with the numerous other findings that AAA performance is higher than or equivalent to XYZ performance, we believe this finding to be spurious. Further replications are needed to confirm this, however.

In addition to evidence for the role of offline recurrent information, we also find evidence here for the importance of online repeated information. The first piece of evidence comes from the difference between XXX and XYZ in experiment 3's high SNR conditions. The difference between these conditions hints at a

difference of purely online repeated information absent of any interaction with recurrent information. This is the first evidence of this difference, as parallel conditions of experiment 1 here and experiment 3 of our previous work show no difference between XXX and XYZ conditions (albeit at -10dB SNR). Additional evidence from experiment 3 comes from correlations, which reveal the XXX conditions to be more closely related than the XYZ conditions. This provides further evidence that purely online information moderates task performance beyond simply baseline ability. The most notable piece of evidence for the role of online information is the mean effect observed between AAAAA and VWXYA in experiment 4. This difference replicates our initial finding in Courtland et al. (2019) and implies that online repeated information confers benefit beyond the recurrent information present in both conditions. At the conclusion of our previous work, it was unclear whether the initial online effects we observed were genuine or spurious (despite their high degree of significance). The replication of the initial effect here, however, lends evidence to the effects being genuine especially given the different subjects, experimental design, and stimuli.

We are then left with the question of why we observe no significant mean difference in the previous work between conditions AAA and XYZ in experiments 3 and 4, between conditions AAA and XYA of experiment 4, and between conditions XXX and XYZ of experiment 3. The first possibility is that our previous studies were underpowered for the size of the effect occurring. As noted in the previous work, the numerical comparisons between these conditions are in the expected direction, but never reach significance. The re-emergence of a significant effect here may thus be attributed to an increased effect size due to increasing the number of pattern A repetitions, although this cannot explain the XXX vs. XYZ difference observed in experiment 3 here. An alternative explanation is that all our studies are underpowered, and given the stochastic nature of type 2 errors, we simply do not make this error in experiments 3 and 4 when in previous experiments we did. Lastly, we may simply have fallen twice into the type 2 error rate despite adequate power simply given the nature of chance. Given the issue of detection power, we aim to increase our sample sizes in future experiments.

With regards to explaining what is at play for online repetition benefit, experiments 1 and 2 imply that the online benefit is likely not due to auditory scene analysis. In light of this, the benefit may stem from

simple beat induction (Honing, 2012; Povel and Essens, 1985) or rhythmic entrainment (Merker et al., 2009). These accounts would both be in agreement with an increased effect size given increased exposure to patterns because both processes necessarily unfold over time. Given the likely reliance on lower-level beat abilities, our future work will include covariates of word recognition task performance that examine individual variability in rhythmic processes, such as the beat alignment test (Iversen and Patel, 2008).

Given the findings of our exploration through parameter space, several follow-up experiments are motivated to further delineate the characteristics of our novel maskers. Given the resurfacing of effects with increased repetitions of pattern A in experiment 4, an experiment testing performance across conditions with variable numbers of preamble repetitions is warranted (e.g. AAA vs. AAAA vs. AAAAA). Given the robust findings of recurrence effects, an experiment testing the effects of no preamble patterns on recurrence – just a 1 second trial containing a word masked by either a recurring or random pattern – should also be undertaken. Lastly, given their situation between the two canonical maskers, our maskers' recruitment of auditory attention must be probed using attentional capture or distractor tasks (Southwell et al., 2017).

Chapter 4

Measuring and modeling how language experience moderates performance on language-based cognitive tests using reported media consumption and neural language models.[1]

4.1 Introduction

Cognitive tests are increasingly used in research on individual differences. For example, a number of recent studies reported correlations between speech perception in noise and working memory (for meta-analysis, see: Dryden et al. 2017). Widely used tests for both (Kalikow et al., 1977; Daneman and Carpenter, 1980) were developed without much regard for potential individual differences in language experience, however. This raises the possibility that at least some of the variability in these tasks is related to differences in participants' language experience, as demonstrated in studies of higher-level language processing (Moore and Gordon, 2015; Wells et al., 2009). Currently, it remains unclear how much this robust correlation between the two tasks – found in 26 of the 30 studies surveyed by Akeroyd (2008) – reveals a correlation between the target constructs or a latent variable of language experience.

[1]This chapter draws on work from: Courtland, M., Davani, A., Reyes, M., Yeh, L., Leung, J., Kennedy, B., Dehghani, M., & Zevin, J.D. (2019). Subtle differences in language experience moderate performance on language-based cognitive tests. In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), Proceedings of the 41st Annual Conference of the Cognitive Science Society (pp. 1559-1565). Montreal, QB: Cognitive Science Society. **and** Courtland, M., Davani, A., Reyes, M., Yeh, L., Leung, J., Kennedy, B., Dehghani, M., Zevin, J. (2019). Modeling performance differences on cognitive tests using LSTMs and skip-thought vectors trained on reported media consumption. Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, 47–53.

Generalization of experimental results crucially relies on the validity and representativeness of the experiment to study the phenomenon of interest. Researchers therefore invest considerable resources in experimental design, particularly in controlling for systematic confounds. When experiments rely on language samples for stimuli, this issue is further complicated because participants bring their complex and diverse language histories into the lab. When participants' language experiences differ systematically and the experiment does not control for this, a confound arises that compromises experimental validity and leads to systematic bias. This is the case for many cognitive tests that standardize language materials in the name of equality, whereas a more equitable approach would be to normalize test difficulty for individuals based on their experience.

One of the primary reasons for the traditional standardization approach over a normalization approach is that creating stimuli that are natural and free from confounds is a difficult laborious undertaking (e.g. as attested by Cutler (1981); Kalikow et al. (1977); Calandruccio and Smiljanic (2012)). The time required to create language stimuli is made worse by the fact that experiments can typically only use each target word or phrase once over the course of the experiment, meaning each stimulus must be uniquely created. In addition to the effort required, experimenter bias and error possibly significantly affect results (Forster, 2000). While previous automation attempts have reduced experimenter bias, error, and workload (e.g. Lahl and Pietrowsky (2006); van Casteren and Davis (2007), vs. Hauk and Pulvermüller (2004)'s manual selection) the process still relies on language statistics calculated from corpora unrepresentative of many participants' language experiences (e.g. Coltheart (1981); Baayen et al. (1996); Kučera and Francis (1967); Thorndike (1944), etc.). This mismatch between the language statistics used to generate test items and participants' actual language experiences represents a persistent confound detracting from experimental validity and perpetuating testing bias.

Linguists have long considered the communicative capacities of every language to be equal and equally expressive (Joseph and Newmeyer, 2012; Pellegrino et al., 2011). Guidelines from the American Speech-Language-Hearing association on cultural competence encourage clinicians to take cultural variables into account in assessing and treating language disorders and differences (American Speech-Language-Hearing

Association, 2013). Despite these commitments in allied fields, and the demonstrable existence of multiple American Englishes (e.g. see for review: Labov et al. 2006; Schneider and Kortmann 2004), most cognitive tests assume "Mainstream" American English (MAE) as a default in the construction of stimuli, potentially confounding cognitive test performance with experience and fluency in MAE. Conversely, language experience is not deterministically related to the usual features that define distinct "dialects" – region, ethnicity, class, etc. People are cosmopolitan and idiosyncratic in the language experiences they seek out, and as a result, may be familiar with multiple language varieties, with potential consequences for their performance on cognitive tests.

Statistical learning, hypothesized to underlie much of language development (Seidenberg and MacDonald, 1999; Elman, 2001), is driven by patterns in language input. Given different input, then, language learners will necessarily construct different distributional models to generate and process speech and language. Online speech and language processing relies heavily on learned statistical regularities to facilitate top-down anticipatory processes. This is evidenced by the effects of surprisal observed when these anticipations are violated (Federmeier et al., 2005; Kutas and Hillyard, 1984, 1980). Given the highly demanding nature of online speech and language processing, anticipatory mechanisms help lessen the cognitive effort needed to accomplish the task. The greater the difference between the listener or reader's language model and the statistics of the language material they are processing, the greater the cognitive burden on the listener. For example, intelligibility levels in noise are better for one's own dialect than for a familiar, but less commonly encountered dialect (Clopper and Bradlow, 2008). In children who prefer a non-mainstream English, familiarity with "school English" is associated with performance on literacy tests (Charity et al., 2004).

Here, we examine the effect of variability in language experience on cognitive tests. We hypothesized that measuring people's language experience indirectly, by having them complete a "media diet" survey, would allow us to identify distinct clusters of individuals based on their viewing, listening, and reading habits. We expect these clusters to only loosely covary with the demographic factors that commonly define distinct "dialect" groups. This new measure of language differences between participants thus provides a

novel aspect of individual variability that we expect to moderate performance on language-based cognitive tasks. As this measure probes the role of language directly, it may be more informative in predicting task performance variability than standard demographic information. To test this we recruit from two populations that differ along traditional demographic lines: USC undergraduates – typically high-SES students pursuing higher education (USC Communications, 2018) – and members of the downtown Los Angeles community – mostly African American and Latinx lower-SES individuals, many of whom not pursuing education beyond high school (e.g. the zip code 90062: US Census Bureau 2018). We administer the aforementioned functional hearing and working memory tasks and expect survey responses to at least partly predict variability in task performance. As we expect this effect to be linguistic, we also predict that language models trained on the media sources will predict participants' behavioral performances.

Our method allows participants to report for themselves the language they are comfortable with and regularly consume. Allowing participants to define their own language experiences ensures stimulus representativeness, increases fairness, and captures individual variability. This moves away from a model that gives researchers the power to define which language materials are representative across all participants (e.g. *Black Beauty* and *Little Women*: Thorndike (1944)) and moves towards a model that empowers participants to define their own language variety. To this end, we develop a method for evaluating language experience's effect on cognitive test performance. In this work, we examine the relationship between the language that participants report consuming in media and their performance on two language-based cognitive tasks. We predict that participants' greater familiarity with the particular language variety of test items (as measured by semantic similarity and statistical predictability) will decrease test difficulty, resulting in higher scores.

We determined media consumption habits by administering a self-report survey, asking participants what media content they currently consume in a variety of categories (Movies, Books, TV, etc.) as well as what they consumed in their formative years. We then cluster participants using k-means clustering and test performance differences across clusters. We also test the distribution of traditional demographic variables across clusters. We then pursue a linguistic explanation for task performance by modeling the

language comprising the sources participants reported consuming and examining its relationship to their performance on the behavioral tests.

To accomplish this, we use neural network language models to learn the joint probability function of word appearances in a corpus. Learning the probability of a word appearing at a certain position in a sentence can be difficult due to sparse representation in the training corpus. However, we choose these models based on their ability to capture long-distance statistical dependencies within a sentence: an advantage they enjoy over n-grams (Bengio et al., 2003). We examine a vanilla long short-term memory (LSTM) model and an attention-based model (Bahdanau et al., 2014). Both are based on recurrent neural networks and are designed to exploit semantic information distributed throughout a sentence to model the probability distribution of vocabulary words appearing as the sentence-final word (Sundermeyer et al., 2012). In addition to modeling the predictability of sentence-final words, we also use a recurrent neural network based encoder to capture sentence-level semantics (Kiros et al., 2015). We use this model to examine whether semantic familiarity affects participants' performances. We model semantics by embedding test items and corpus sentences in a high dimensional vector space and observing the distances between each item and its neighbors from the corpus. We predict that greater semantic similarity and greater sentence-final word predictability as captured by these models will correlate with participants' performance on our cognitive tasks.

4.2 Methods

4.2.1 Participants

We recruited participants from the USC undergraduate population (N=70) and on a local community college campus (Los Angeles Trade-Technical College, N=25). USC students participated in exchange for course credit and community participants were compensated for their time at $15 per hour, pro-rated at 20 minute intervals. No requirements were placed on age, but due to recruitment populations, 80% of participants were between the ages of 19 and 26 (mean=22, std=6.25).

4.2.2 Cognitive Tests

To test language ability, participants complete the reading span task developed to assess verbal working memory (Daneman and Carpenter, 1980) and the speech perception in noise task (SPiN) developed to assess functional hearing (Kalikow et al., 1977). In the reading span task, participants read sets of sentences aloud while remembering the last word of each sentence. At the end of a set, they report the full sequence of sentence-final words in the set (with no partial credit). Set size increases (from 2 to 7) every three sets until participants cannot correctly recall any set at that length, at which point the task is terminated. The SPiN task presents spoken sentences over headphones masked with 12 talker babble (a combination of 6 male and 6 female voices speaking continuously). At the end of the sentence, participants are asked to report the final word of the sentence. We used recordings from the Nationwide Speech Project (Clopper and Pisoni, 2006) to create the stimuli and present trials at +6dB SNR which produced large individual differences in accuracy in pilot results. We choose these tests due to their importance as widely used individual difference measures in clinical populations to diagnose age-related decline (Byrne, 1998), aphasia (Caspari et al., 1998), Alzheimer's (Kempler et al., 1998), and schizophrenia (Stone et al., 1998). We also choose these tests for the important, yet often unacknowledged, role language processing is likely to play in both.

4.2.3 Survey

To capture participants' diverse language experiences, we use a proxy measure: the language materials they choose to consume regularly. To this end, we constructed an online survey (approx. 20 minutes long) that probes participants' current and formative media consumption habits (TV, Movies, Books, News, etc.), elicits short language production passages, and collects basic demographic information. We use this tool to glean each participant's media diet, which forms the basis for later linguistic grouping and analysis. We use the language obtained from the sources participants report as a model for participants' actual language input and a measurable substitute for language experience.

4.2.4 Equipment

Subjects sat in a noise attenuating booth and participated in the survey and behavioral tasks on a desktop PC computer. USC participants were allowed to complete the survey online prior to their lab session. Participants first completed the reading span task, followed by the SPiN, and finally the survey. The reading span task was administered and scored by a researcher to ensure subjects read aloud continuously. Upon completion of each sentence, the researcher advanced the display to the next sentence in the set and solicited verbal responses at the end of each set. After a brief training phase, participants were not given feedback on their performance and were not told their failure had caused the end of the test, simply that it had ended. The SPiN test was administered using Paradigm experiment software; participants typed their responses into a free-response text box. Trials began after a 500ms delay once participants had submitted their response. Stimuli were presented at a comfortable level, standard across participants.

4.2.5 Clustering

We create a media source space in which each dimension represents a reported source (e.g. movie) collected in our survey. Each participant is thus represented as a binary vector in this space, with 1s in dimensions corresponding to sources they consume, and 0s in those they do not. To ensure each dimension is informative (and reduce the dimensionality), we only represent sources reported 10 or more times – thus avoiding dimensions that would only differentiate a few participants (i.e. the rest would all receive 0s in that dimension). This leaves 314 dimensions along which participants were clustered using the k-means algorithm (Lloyd, 1982). Figure 4.1 shows the distortion values for different numbers of clusters, revealing 3 clusters to be the inflection point at which more clusters provide only marginal returns. The algorithm takes this point as the true number of clusters because increasing the number of clusters beyond this simply subdivides the true clusters, thus over-fitting.

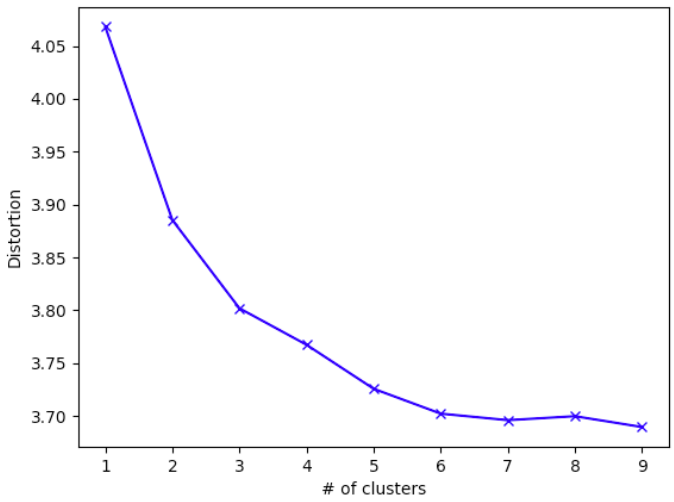

Figure 4.1: K-means clustering reveals 3 clusters of participants in our media consumption space. This is evidenced by the inflection in distortion decrease that occurs at $k = 3$.

4.2.6 Corpora Construction

We aggregate language data from the sources participants reported in our survey for further linguistic analysis. This produces two corpora (one for each cluster) that allow us to model their language differences. We fully acknowledge the difference between consuming sources as text, as our models do, and speech, as our participants do. Despite this, however, text fully captures the regularities of lexical and supra-lexical features we expect to influence performance on our behavioral tasks. It is possible that analyzing the speech from these sources would provide further insights into their influence on our tasks, but this is beyond the scope of this paper.

We collect the sources for each corpus by scraping repositories of television scripts (Springfield! Springfield!, 2019) and movie subtitles (YIFYSubtitles, 2019). In total, we collect 1027 scripts of complete series (e.g. all episodes of *Futurama*) and 194 movie subtitles. We then clean the sources by removing information that does not reach viewers (e.g. stage directions, parenthetical notes, etc.). Each corpus is then tokenized into sentences for model training.

4.2.7 Language Modeling

To model the language statistics of each cluster's corpus, we use 5-gram language models with backoff (Katz, 1987). These models estimate the likelihood of a sentence as the product of the conditional probabilities of its words given the words that precede them. Thus for a sentence of length L, the likelihood is:

$$\prod_{l=1}^{L} P(w_l|w_{l-(n-1)}...w_{l-1}) \tag{4.1}$$

where n is a hyperparameter set to control the number of preceding words considered for context ($n = 1$ is simply the marginal probability). Because the probability of encountering the preceding string of words in training decreases as the length of the string increases, backoff allows the algorithm to decrease n until the preceding string *has* been seen in training (thus allowing the conditional probability to be estimated). Therefore, while we initially set our $n = 5$, probabilities may be calculated given less prior context.

4.2.8 Cloze Modeling

Cloze probability refers to the probability of encountering the last word of a sentence given the sequence of words that precede it (i.e. all non-final words of that sentence). That is, given a sentence of words w_1 through w_n, the cloze probability is expressed by: $P(w_n|w_1...w_{n-1})$. This conditional probability is a particularly important metric for our purposes because of the privileged position sentence-final words enjoy in scoring both of our behavioral tasks (cf. Duffy and Giolas (1974)'s effect of predictability on task performance). Both our behavioral tasks place participants in a condition of increased cognitive burden (either using adverse listening conditions or simultaneous verbal storage and processing demands) and then ask them to identify or remember the last word of a sentence (Daneman and Carpenter, 1980; Kalikow et al., 1977). If these words are predictable for a given participant, top-down processing can alleviate

the cognitive burden of online language processing, making the task easier (Winn, 2016). If participants systematically differ in their ability to predict these sentence-final words, as might be caused by different language experiences, the task would effectively be easier for one group of participants, leading to higher scores.

To test this, in addition to the 5-gram model which proceeds from the beginning of a sentence seeking to model its probability, we also model the surprisal associated with encountering the final word of the sentence. While in theory the model aligns with the concept of cloze probability, this rarely occurs in practice given the sparseness of a training corpus. To model this, we adopt a similar method to n-gram models with backoff. We calculate the conditional probability of the last word given the $n-1$ preceding terms:

$$P(w_L|w_{L-(n-1)}...w_{L-1}) \tag{4.2}$$

where we initialize $n = 5$ and reduce its value until the preceding string has been encountered in the training corpus ($n = 1$ is simply the marginal probability of the word occurring sentence-finally).

To better approximate the nature of the task and increase the power of our models, we also train a vanilla LSTM and LSTM with attention on each cluster's corpus to predict the last word of a sentence given all the previous words. The attention-based LSTM model is composed of a layer of LSTM cells that capture the hidden representation of the sequence of words from the beginning of the sentence up to the last word. The final representation for sentence i is shown by H_i (eq. 4.5, below) and is generated by applying attention weights (α_{ij}, eq. 4.4) to the LSTM's hidden states, h_{ij}, corresponding to each word j in sentence i of length n. W_s, W_t, u_s, b_s and b_t are learned simultaneously during back propagation (Wang et al., 2016).

$$u_{ij} = tanh(W_s h_{ij} + b_s) \tag{4.3}$$

$$\alpha_{ij} = \frac{exp(u_s u_{ij})}{\sum_{k=0}^{n-1} exp(u_s u_{ik})} \tag{4.4}$$

$$H_i = \sum_{j=0}^{n-1} (\alpha_{ij} * h_{ij}) \tag{4.5}$$

Using a fully connected and a softmax layer, we then calculate the probability of each word w in the vocabulary appearing immediately after the sequence as p_w(i.e. at the end of that sentence).

$$v_{iw} = W_t H_i + b_t \tag{4.6}$$

$$p_w = \frac{exp(v_{iw})}{\sum_{k=0}^{|vocabulary|} exp(v_{ik})} \tag{4.7}$$

For the experiment, we use a vocabulary consisting of the 10k most frequent words in the corpus. The hidden size of the LSTM and attention vectors are set to 100. We use 300-dimensional GloVe word embeddings as the semantic representation of the words (Pennington et al., 2014).

4.2.9 Skip-thought Vectors

To obtain a quantitative measure of semantic similarity, we embed test items and sentences from each cluster's corpus in a high dimensional vector space and measure the distance of each test item to neighboring items from the corpus. To encode target and corpus items into vectors, we use combine-skip-thought vectors as detailed in Kiros et al. (2015). These encode sentences using RNNs with GRU into a 4800-dimensional vector which is the concatenation of a 2400-dimensional uni-directional encoder and a 2400-dimensional bi-directional encoder (1200 dimensions for backwards and forwards each). Results from the original paper show that these vectors capture a high degree of sentence-level semantics, particularly as it relates to encoding similarity as vector-space distance: the closer two sentences are in the embedded vector space, the more semantically related they are. We therefore take the distances in this embedded vector space to be indicative of how typical a test item's semantics are given the corpus of a participant's cluster.

We measure each test item's mean distance from all corpora items using the Taxicab distance (L^1 norm, eq. 4.8) and standardized Euclidean distance (eq. 4.9):

$$\sum_{i=1}^{n} |u_i - v_i| \tag{4.8}$$

$$\sqrt{\sum_{i=1}^{n} (u_i - v_i)^2 / V[x_i]} \tag{4.9}$$

where $V[x_i]$ is the variance vector over the components of all vectors.

We also measure the mean distance to the closest 100 corpus neighbors in the event that similarity to all corpus items proves less informative than similarity to the closest matches from the corpus.

4.3 Results

4.3.1 Clustering

The clustering included all reported media sources and revealed three clusters based on participants' consumption habits. Despite a substantial drop in distortion from 2 clusters to 3 (see Fig. 4.1 for distortions), cluster 0 proved too small to analyze: it contains just 2 participants. Its size precludes both behavioral analysis, which requires an adequate number of samples to be statistically feasible, and computational modeling, which requires a corpus built from an adequate number of reported sources (aggregated across a cluster). Given these limitations, the following analyses will only use clusters 1 and 2 as the sample population (still 98% of the original sample). This clustering, far from an artifact of random seed, proved stable across random restarts. Over 1000 iterations, on average 75% of participants were re-clustered in the same groups (see **Behavioral Data** for the effects on statistical tests).

Regarding cluster membership, we expected USC students and community members to be unevenly distributed between clusters, and this was true, although not categorically. As seen in Table 4.1, the two are

relatively balanced across clusters. Thus, cluster membership and *a priori* group membership are treated as orthogonal in the following analyses.

In addition to the a priori population, we examined the distribution of traditionally considered covariates across the clusters. We wanted to test whether self-reported media consumption provided new information beyond existing measures (i.e. we were not just capturing an existing highly correlated dimension of variance). As seen in Table 4.1, typical demographic variables were fairly evenly distributed across the clusters. One-way chi-square tests revealed that none of the demographic variables significantly differed from an even split across clusters (i.e. the expected values if cluster and variable were independent).

Variable	Level	Cluster Ns		Cluster %	
		1	2	1	2
Population	USC	34	24	59%	41%
	LATTC	7	13	35%	65%
Gender[2]	Female	33	21	61%	39%
	Male	7	16	30%	70%
Schooling	High School	10	10	50%	50%
	Associate	4	4	50%	50%
	Some College	19	15	56%	44%
	Bachelor's	7	6	54%	46%
	Master's	1	2	33%	66%
Mono-lingual	True	10	11	48%	52%
	False	31	26	54%	46%
SES Self-Report[3]	High	13	13	50%	50%
	Medium	16	10	62%	38%
	Low	12	14	46%	54%

Table 4.1: The distribution of traditionally considered covariates across clusters is fairly even. We observe no obvious imbalance between clusters along any demographic dimensions our survey measured. One-way chi-square tests support this.

Given the orthogonality of self-reported media consumption to traditional demographic variables, we hereafter focus on the observed dimension of variance: media diet. We probe how the clusters differ in their media habits in order to delineate their makeup. We examine the clusters' centroids to calculate which dimensions (i.e. sources) they differ maximally along. This provides a measure of which media sources are most distinct between clusters. We find the following sources to be the 5 most different between clusters 1 and 2 and provide the difference in mean consumption between the two (i.e. $\bar{x}_1 - \bar{x}_2$) in parentheses: Star

[2]One participant in cluster 1 chose not to report gender.

[3]Participants reported their SES on a continuous scale. Here, we bin responses into 3 quantiles to report distribution across clusters.

Wars (.64, specific films reported in the series were less powerful, on the order of .11-.17), Yes! (-.47), CNN (-.26), People (-.12), and Harry Potter (-.12). We hesitate to draw any conclusive generalities on the two clusters' media diets, but at a glance it appears that cluster 1 consumes lots of high fantasy (Star Wars, Lord of the Rings, The Chronicles of Narnia, etc.) while cluster 2 consumes more nonfiction (Yes!, CNN, People, etc.).

4.3.2 Behavioral Data

As seen in Figure 4.2, the SPiN task revealed a main effect of cluster, $F(1, 76) = 7.30$, $p < .01$, but no main effect of population and no interaction between the two. In the reading span data, we again find a main effect of cluster, $F(1, 76) = 4.05$, $p < .05$, and a main effect of population $F(1, 76) = 13.57$, $p < .001$, and no interaction between the two. In our 1000 clustering iterations, 63% of iterations revealed statistically significant effects of cluster on the SPiN task (at $\alpha = .05$). This was not replicated with the span task, however: only 4% of our iterations found statistically significant effects.

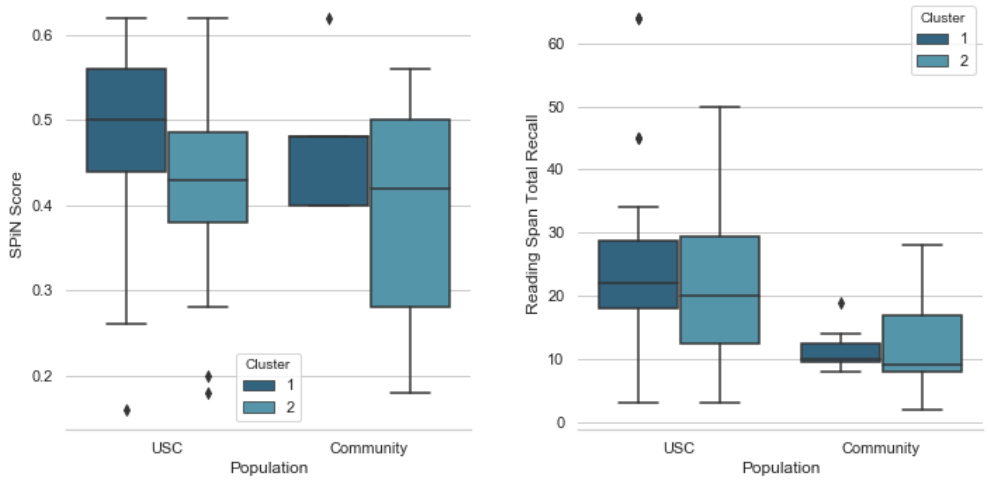

Figure 4.2: Results from the SPiN test reveal a significant difference between clusters but not populations. Reading span also shows an effect of cluster, but a larger effect of population. We observe no significant interactions.

The span test will play a minor role in further analyses, due to the difficulty in handling test result data and its scoring. Because the span task is terminated whenever participants fail to recall a set, participants provide unequal numbers of observations. The analyses are additionally constrained by the small number

of items a typical participant completes. While observations exist for items later in the test, they are for a few extraordinary participants. This presents a problem not only in the paucity of observations, but also in the fact that these participants are unrepresentative of the general sample in their task abilities. As such, both item-level statistics and graphical representations are challenging.

Our survey obtains several pieces of demographic information that are traditionally considered relevant covariates of performance on our cognitive tasks, such as socioeconomic status (SES, self-reported), age, education level, and monolingual status. None of these correlated significantly with performance on either task.

4.3.3 Language Media Input Modality

The above findings of differences between cluster performances motivated us to explore differences between clusters' survey behavior (other than the categorical responses which were used in clustering) to explain their performance data. In particular, we wondered whether the stronger task performances of cluster 1 might be due to increased experience with the tasks of speech perception and reading.

To probe this, we tested whether cluster 1 reported significantly more speech sources (TV, Movies, Music, and News shows) and significantly more text sources (Books, Newspapers, Magazines, Online News, and other online reading) than cluster 2. Indeed, we find that cluster 1 participants report significantly more listening on average than cluster 2: $t(42.82) = 3.09, p =< .005, d = 0.67$ (a medium effect). We also find that cluster 1 participants report significantly more reading on average than cluster 2: $t(72.6) = 5.10, p < .001, d = 1.13$ (a large effect). This may indicate an effect of modality-specific training on task performance. To probe this, we test the correlation between the number of speech sources a participant reports and their SPiN task performance. We test rank correlation rather than linear correlation as we are unsure of the linearity of the relationship between number of sources and modality-specific benefit, as well as to control for the effect of outliers in both performance and reporting volume. We observe a significant correlation between the two: $\rho(76) = .31, p = .005$. We do not, however, observe a significant correlation between number of text sources and span performance.

We also tested whether past modality preference (solicited with "when you were growing up...") would relate to current modality preference. We find a strong correlation between the amount of spoken language items reported growing up and amount of current items reported: $r(76) = .91, p < .001$. This correlation extends to the number of written language items, although not as strongly: $r(76) = .49, p < .001$.

4.3.4 Statistical Models

To evaluate the claim that our language models were capturing meaningful statistical regularities in the language of each cluster's corpus, we tested whether the log-likelihood produced by a model for each of the test items would correlate with mean performance on those items for the cluster. We do not observe a significant correlation between cluster 1's 5-gram model and performance on either the SPiN ($r(48) < 0.01, p > .1$) or span ($r(25) = -0.01, p > .1$). We also observe no significant correlation between cluster 2's 5-gram model and its performances on SPiN ($r(48) = -0.02, p > .1$) or reading span ($r(25) < 0.01, p > .1$). Additionally, we tested the correlation between cluster 1's 5$^{\text{th}}$-order surprisal model and its performance and found no correlations with SPiN ($r(48) = 0.07, p > .1$) or span ($r(25) = -0.09, p > .1$). Similar results were obtained for cluster 2 (SPiN: $r(48) = 0.21, p > .1$, reading span: $r(25) = 0.02, p > .1$).

In addition to modeling statistical properties of particular items, we also tested whether the difference between the language and surprisal models might capture the significant differences we see on our behavioral tasks. This method avoids any idiosyncrasies of particular items (as comparisons are within item) and instead captures any language differences of media sources. We again find a lack of significant correlation between 5-gram likelihood differences and task performance for both the SPiN ($r(48) = 0.09, p > .1$) and reading span ($r(25) = .25, p > .1$). Similar results are observed for the 5$^{\text{th}}$-order surprisal model (SPiN: $r(48) < 0.01, p > .1$, reading span: $r(25) = 0.08, p > .1$). As shown in Fig. 4.3, differences between the model likelihoods are close to zero for most items, with a few outliers.

To examine the non-correlations and clustering around 0 on Fig. 4.3's x-axis, we tested the correlation between models and found strong correlations for both the SPiN ($r(48) = 0.94, p < .001$) and span

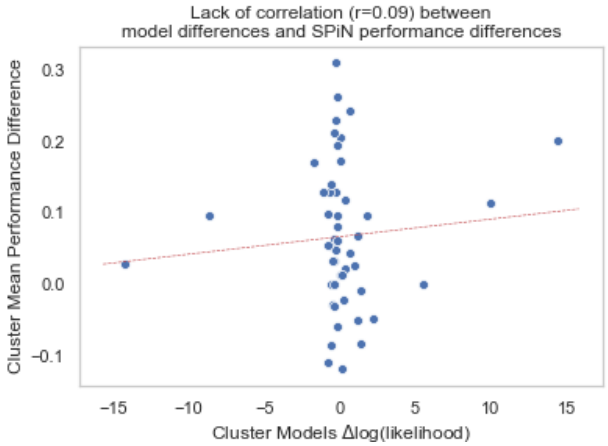

Figure 4.3: The non-correlation of cluster performance differences with model likelihood differences indicates the statistical information captured by the models is a poor predictor of behavioral performance. The significant cluster performance difference can be seen here by the majority of items occurring above 0-difference on the y-axis. A LMS-Regression line is drawn in red for reference.

($r(86) = 0.97, p < .001$) test. These strong correlations, coupled with linear regression slopes of $\beta_1 = 0.96$ (SPiN) and $\beta_1 = 0.94$ (span) imply nearly identical log-likelihood scores between models despite training on categorically different sources. While the results reported here are specific to 5-gram language models and 5$^{\text{th}}$-order surprisal models, other lower-ordered models of both yielded similar results.

4.3.5 Neural Models

For each test item, we correlate each cluster's LSTM activation of the sentence-final word with that cluster's mean behavioral performance (i.e. the percent of the cluster's participants who answered that item correctly). We use rank correlation as we are uncertain of how linear the mapping between predictability and performance benefit will be.

We observe significant rank correlations between the activation of both clusters' vanilla LSTMs and their respective mean performances on the SPiN items ($\rho(48) = .39, p < .01$ for cluster 1, $\rho(48) = .46, p < .005$ for cluster 2). We observe weaker but still significant correlations between the attention-based LSTM activations and mean performances on SPiN items ($\rho(48) = .31, p < .05$ for cluster 1,

	Cluster 1		Cluster 2	
	SPiN	span	SPiN	span
Vanilla LSTM	.39	-.03	.46	-.15
Attn. LSTM	.31	.02	.29	-.03
Taxicab	.486	.075	.519	-.022
Std. Euclid.	.408	-.048	.440	.092

Table 4.2: Mean behavioral performance on SPiN target items is significantly rank correlated to both LSTM activations and skip-thought distances for both clusters. We find no significant correlations with the span test for either cluster.

$\rho(48) = .29, p = .05$ for cluster 2). This poorer performance of the more complex model is noteworthy. We observe no significant rank correlations between any model's activations and performance on the corresponding span task item (see Table 4.2).

To test the ability of skip-thought vectors to predict performance, for each cluster we test for a correlation between the distance from all its corpus items' vectors to a given test item's vector and the mean performance of its participants on that item. Given uncertainty of whether the distance-performance relationship will be linear, we use rank correlation. Using the distance metrics in eqs. (4.8) and (4.9), we observe significant rank correlations between vector-space distances and performances on the SPiN task (see Table 4.2 for test statistics, all $\rho(48), p < .005$) but not the span task. In addition to the mean distance of all items, we calculated the distance to the closest 100 neighboring corpus items and obtained similar results.

4.4 Discussion

We observed significant performance differences on a speech perception in noise task and a working memory task between clusters of participants derived from self-reported media consumption. These differences were above and beyond differences driven by *a priori* participant groups – students at a university vs. participants from the surrounding community. This clustering was robust to randomness and orthogonal to any traditionally considered demographic variables. As we have no reason to believe that the tests' target

constructs systematically vary between our clusters, we conclude that media diet represents an uncorrelated latent variable moderating task performance. To our knowledge, our identifying media consumption as a significant orthogonal predictor of cognitive task performance is a novel contribution of this work.

This novel predictor is surprisingly powerful at explaining language test performance considering its complete lack of explicit linguistic information. In pursuing a linguistic explanation for our finding, we used statistical language models trained on sources participants reported consuming to analyze test items. These models did not identify particular stimuli driving performance differences, and we found no obvious differences in how well stimuli fit our models. However, our more complex recurrent neural models did in fact reveal a correlation between models trained on our media corpora and behavioral performance. This implies the statistics used in the initial models were not sophisticated enough. Cloze probability, for example, is computed as a simple ratio of the tokens of a word in context to all tokens in that exact context in the corpus.

Our neural language models tailored to the media consumption of different "clusters" of English speakers *do significantly* predict performance at the item level on a test of functional hearing (SPiN). In particular, LSTM models, which are perhaps the most natural way to model a task in which the predictability of the final word in a sentence has a strong influence on performance, correctly predict accuracy for each cluster. For the reading span task, in contrast, neither type of model correctly predicted performance. It is possible that the models are not capturing the relevant linguistic information for reading span or that reading span simply depends less on language (and language experience) overall than SPiN. An alternative explanation, however, comes from the difficulty in handling span performance data and its scoring. In the span task, items are presented in a fixed order, and difficulty increases from trial to trial as participants are required to maintain more items in working memory. This makes scoring at the item level difficult to interpret. Given these complications with the scoring procedure, it is possible that item-level analysis of the reading span is uninformative and invalid compared to the straight-forward scoring procedure of the SPiN.

Regarding the SPiN task, the robustness of the correlation between skip-thought vector mean-neighbor distances and participant performance is curious, however. The interesting aspect of this relationship is the direction of the correlation: that as the distance from corpus neighbors increases, performance on the item *increases*. This implies that unusual items are scored better on than familiar ones. This finding is not necessarily at odds with the finding of the neural cloze models: that increased predictability of the last word positively correlates with performance on that sentence. The two models differ in several key aspects which may explain their differences. Firstly, skipthought distances do not capture statistical predictability but rather semantic similarity, so while the last word (or in fact the sentence as a whole) may be semantically odd, it also may be relatively easy to predict the last word from the rest of the sentence. Secondly, skipthoughts operate at the level of the entire sentence rather than at the level of just the last word, which means that all of the words contributing to their embedding except the sentence-final one do not directly factor into the scoring of behavioral performance. This means that the majority of the linguistic information they encode is uninformative for capturing predictability of the last word, which is a direct correlate to how the task is scored. Lastly, skipthoughts are capturing the semantic novelty of a sentence. It is possible that the increased attentional resources these items demand over semantically typical items actually causes participants to perform better on these items rather than worse. This must be tested further before concrete conclusions can be drawn, but it represents an interesting future direction for study.

An additional finding of note is the highly significant difference in the number of sources reported by cluster 1 compared with cluster 2. It is possible that the greater number of sources indicates that cluster 1 contains more voracious consumers of media than cluster 2. This increased media consumption in the modalities of our tests may be providing cluster 1 members with modality-specific training they leverage at test time. Indeed, the correlation we observed between number of speech sources reported and SPiN performance supports this explanation. This is especially plausible given that watching a TV show or movie involves perceiving character dialog often obscured by various sources of noise (soundtrack, sound effects, etc.). It is also possible, however, that the increased responses and performance from cluster 1 is

indicative not of their increased modality-specific training but rather a latent variable such as attentiveness or enthusiasm at participating in all aspects of the study.

It should be mentioned that participants' responses may reflect a (possibly implicit) choice to make specific habits known in the context of the survey. Given the importance of shared experience in forming relationships, what pieces and types of information people share and what they keep private often acts as a type of signaling that forms the basis of social cohesion. Thus, media diet survey responses may be more appropriately interpreted as signaling membership in a language community than literally reflecting the language practices of that community. Indeed, the vast majority of items in the corpora are professionally produced texts, which are likely to differ less than spontaneous spoken and written communication. In future work, we plan to obtain rich, naturalistic language samples in addition to the media corpora included so far to strengthen the evidence found here.

The identification of a dimension (other than the target construct) that test performance differs significantly along brings into question not only specific test validity probed here, but also the validity of the entire practice of test item standardization. This is true whether the dimension is categorical media consumption or the linguistic content of that media. Tests that use language to probe target constructs must take the language of their test into account – not as a static entity to be standardized, but as the diverse and dynamic communication medium that it is. Test validity relies on the ability to generalize a test's result to participants' everyday behavior. This is only valid if the test is representative of the language they encounter in their daily lives (Coleman, 1964). Thus, tests employing standardized language not only contain inherent inequity for those less familiar with the test language, they are also less valid.

We believe the results obtained here are an initial step toward taking participants' self-reported language experience into account in interpreting their performance on cognitive tests. In light of the evidence that a connection likely exists, we support the approach of normalizing, rather than standardizing, the language of cognitive tests. We predict normalization will produce tests that are simultaneously more fair and more valid. Regarding increased validity, the use of dynamically generated corpora would afford a

significant benefit over static corpora by reducing sampling error. Every corpus necessarily contains idiosyncratic sampling error affecting results (Clark, 1973). The repeated use of norms generated from a single corpus (e.g. as was traditionally taken from Kučera and Francis (1967) or Thorndike (1944)) amplifies this noise and its role in experimental results. The construction of dynamic corpora we are planning will mitigate this effect by providing multiple samples across which real statistical regularities are likely to replicate, while sample noise is not (like bootstrapping: Efron (1979)).

Generating equitable stimuli is a difficult or possibly infeasible task for human researchers, but could potentially be automated using generative models. If such models were driven by statistics that are highly representative of participants' language experience, they may do a better job of capturing cognitive constructs without smuggling in variability resulting from differences in language experience. Perhaps the most exciting future direction of this research will be to facilitate using more representative language statistics in designing stimuli for cognitive tests. The study of how language experience influences test performances that we take here represents a first step to understanding and mitigating this test inequity.

Here we aim to show that participants' diverse language experiences must be taken into account when diagnostic tools like those tested here are designed. Ideally, given the unique nature of language experience, test creators should strive to create tests that present equal difficulty to each participant by using personalized test language. This step to ensure equity is especially important given that test scores cannot simply be adjusted for using traditionally defined dialectal boundaries – as demonstrated here by the uninformativeness of the demographic variables that define these boundaries.

While the eventual goal of this work is to generate valid and fair stimuli *ex nihilo* given people's language models, the evaluation of existing stimuli materials represents a necessary first step taken here. The development of models capturing linguistic features that predict behavioral performance provides the possibility for using these models to identify or synthesize fair test items. Modeling the relationship between language experience and task performance allows rapid prototyping and evaluation of stimuli sets with previously infeasible speed. This allows a much larger set of candidate stimuli to be evaluated affording new levels of rigor to the test creation process. This speed also opens the door for individual

personalization of test items, a task far too labor-intensive to perform manually. Our future work will test our models' ability to create test stimuli equitable across diverse language communities.

These methods for promoting equity are likely relevant to education where equality vs. equity is debated as the difference between equal access to educational resources vs. access to resources leading to equal outcomes (e.g. Green (1983); Stromquist (2005); Espinoza (2007)). Language-based cognitive testing and access to education share several features in common. Both are moderated by the complex individual variability of personal experience. Those with the worst outcomes in both are underrepresented among those setting policy and creating tests (National Science Foundation, 2013; Thaler and Jones-Forrester, 2013; Thaler et al., 2015). And most importantly, both also determine relevant real-world outcomes for test takers. Many cognitive tests use linguistic stimuli to assess other cognitive functions; by identifying specific ways that language variety influences test performance, we can start to tease apart educationally and clinically meaningful deficits from normal social and cultural differences.

Chapter 5

Discussion

5.1 The Findings of this Dissertation

In chapter 2, we examined the effect of temporal regularity on listeners' ability to overcome energetic masking. We found that temporal regularity confers a benefit, but not in the manner we initially expected. We expected participants to be able to recognize a repeated noise pattern within a trial to overcome its masking effects. What we observed was participants using the recurrent nature of a particular pattern across trials to overcome this pattern's masking effects. Additionally, while we initially observed a large effect of differences in central tendency supporting our expectations, this effect proved unstable. Despite this, we identified and successfully replicated robust effects of variance and correlation across conditions. We obtain repeated evidence that participants are utilizing offline informational content in cross-trial learning. This finding that release from energetic masking uses a higher level cognitive mechanism begs a reexamination of relegating energetic masking solely to the periphery.

In chapter 3, we began to tease out what mechanism is responsible for the advantage observed in chapter 2 and what its limitations are. We observed evidence that the effect is not robust to spectral manipulations of either small or large magnitudes. We also observed that the mechanism cannot overcome the canonical sensitivity to signal-to-noise ratio of energetic masking. Lastly, and perhaps most importantly,

we simultaneously replicate both the initial elusive central tendency effect of chapter 2 and the robust variance and correlation effects of chapter 2.

One aspect of the findings from chapters 2 and 3 should be discussed before we move on to discuss chapter 4. The findings of higher variability in conditions containing the recurrent pattern and higher correlation between these conditions seem at first glance to be separate phenomena. However, a deeper examination reveals that the increased variability may actually influence the strength of the correlations. The correlation data supporting our findings exhibit a notable asymmetry: the correlations between conditions containing the recurrent pattern are stronger than those between conditions lacking the recurrent pattern. This finding is effected by the significantly decreased variance we observe in conditions lacking recurrent information. This restricted range can actually cause weaker correlations (Nikolić et al., 2012; Alexander et al., 1984; Aguinis and Whitehead, 1997), which may at least partly explain this asymmetry. That being said, this likely does not explain the entirety of the decrease between conditions lacking recurrent information, which on balance are still higher than those across recurrent informational lines (i.e. between conditions lacking recurrent information and those containing it).

In chapter 4, we examine how language background moderates performance on standardized cognitive tests relying on language, such as those in chapters 2 and 3. We observe a significant categorical effect of people's media consumption habits on their performance on both the reading span task (Daneman and Carpenter, 1980) and the speech perception in noise (SPiN) task (Kalikow et al., 1977). Most importantly, the dimension of media habits is orthogonal to commonly considered demographic dimensions. This means it represents a previously unexamined dimension of individual variability which is a significant predictor of test scores. Pursuing a linguistic explanation for this difference, we then obtain a proxy for people's implicit language models by constructing a neural language model learned from the linguistic content they report consuming. We find that these language models correlate with people's performance on the SPiN task but not the reading span task.

The lack of effect we observe with the reading span task may be due at least in part to difficulty scoring and uniformly analyzing participant performance. The first issue with scoring is that there exist multiple

methods to score the reading span task (Whitney et al., 2001). Secondly, contrary to the "tidy" data format of SPiN, span often produces unequal numbers of observations across participants, and trials occur in sets thus hampering analysis of individual items on the basis of their language fit to a particular subject. While the SPiN task provides tidy data, the task divides observations into two treatments: high probability and low probability. The high probability sentences represent plausible, albeit antiquated, propositions. However, the low probability sentences all achieve their aim of a low cloze probability by employing reported speech (e.g. "Mr. Smith spoke about the aid."). As any persistent confound within a test threatens its validity, it is desirable to avoid reusing this syntactic structure by finding other low probability sentences satisfying the task constraints. This represents an exciting generative use of the language models we are developing.

While the word recognition tasks of chapters 2 and 3 avoid this confound of reported speech constructions, they exhibit a different shortcoming: their lack of sentential context makes them questionable targets for functional hearing. If the linguistic mechanisms that underlie functional hearing do not have a chance to operate until target presentation, their ability to aid in lexical identification may be minimal. This jeopardizes the ecological validity of these tasks, as words are not often spoken in isolation in a naturalistic context. Words occur embedded in speech streams that provide myriad contextual cues including prosody and pragmatic context. These additional cues can be utilized to vastly different degrees by different individuals leading to disparate functional outcomes that are undetectable by isolated word recognition tasks. The word recognition tasks do, on the other hand, capture participants' abilities to overcome patterned noise – a dimension along which we observe significant variability. This dimension is not captured by the SPiN task or other similar tests and is crucial in determining functional outcomes. Thus the tests seem complementary in which aspects of functional hearing they capture.

5.2 Individual Variability

While our word recognition task and the SPiN task seem to address different aspects of functional hearing, they share a common moderating factor: language experience. While the SPiN task's inclusion of higher linguistic structure may amplify this factor, a participant's experience in the language variety of the test is at play in all our tests. The role of language experience in task performance is particularly hazardous given its latent nature and how much it can vary between participants. Given these facts, it has traditionally been impossible to control for. However, advances in computing technology have enabled a previously infeasible levels of personalization to be applied to the problem. The work we undertake in chapter 4 represents a first step towards solving the problem of bias against non-"mainstream" Englishes.

The dichotomy of "dialect" and "mainstream" is not only simplistic in its categorical classification, but also results in serious issues regarding equity and adequate representation in society's expectations of communicative behavior (Cazden, 1988; Christian, 1997). The education literature has recently addressed a similar issue by replacing the deficit model, focusing on students' alleged shortcomings, with an asset model, which regards differences as "human variation rather than pathology" (Reid and Valle 2004; see also Harry and Klingner 2007). The confounding of cognitive test performance with experience and fluency in Mainstream American English means that subjects whose native English is not "mainstream" are fundamentally disadvantaged.

This language disadvantage can take a toll on people when they are expected to constantly operate in a language variety which is not their native variety. This is particularly true because they cannot take full advantage of predictive top-down mechanisms that alleviate the bottom-up processing burden of speech and online language comprehension. These mechanisms are crucial in allowing listeners to converse comfortably. Semantic context can be particularly powerful for reducing listening effort (Winn, 2016). Given the power of semantic context, it is unsurprising that those with lessened ability to utilize this cue must expend greater listening effort (e.g. non-native listeners: Borghini and Hazan 2018). Having a different internal language model can thus be thought of as a lessened version of this disadvantage. The greater the

difference between the internal language model and the input's language statistics, the more cognitive burden the listener must bear. A similar effect across dialects is suggested by decreased intelligibility levels in noise relative to one's own dialect but listening effort itself has not been measured (Clopper and Bradlow, 2008). An analogous case can be argued for reading effort, an activity that is increasingly important in, and necessary to secure, gainful employment.

Individual differences in language experience have not been studied in this way because native speakers of the same broader language were assumed to have analogous implicit language models. However, recent findings reveal this may not be true, and divisions exist within language communities previously considered uniform (Dehghani et al., 2014). Given the underlying difference in statistical language models, communicative burden may arise even between interlocutors of the "same language community". When this burden is borne by a native speaker of non-mainstream English, the increased listening effort required compounds social inequity. A better method of estimating implicit language models can thus aid in the testing and understanding of listening effort, the ensuing mental fatigue, and its effect on communicative ability.

Another process crucial to individual variability is inhibitory control. The ability to inhibit distracting information, such as masking noise or simultaneous language processing tasks, and focus on the task goal can be a large source of individual differences. An additional manner in which inhibitory control might play a crucial role in both tasks is in lexical activation, where it serves to narrow down the list of possible competitors to facilitate lexical selection (Hasher et al., 2008; Dey and Sommers, 2015).

Two other major sources of individual variability in online speech and language processing are lexical access and word decoding ability (Frederiksen, 1981; Jackson and McClelland, 1979; Perfetti, 1985; Perfetti and Lesgold, 1977). Variability in lexical access is a factor in the auditory tasks and variability in word decoding is relevant to the reading span task. Both processes interact closely and complexly with the predictive processes described above. These processes can help explain significant contributors to the variation we observe in all our experiments.

5.3 Construct Validity

Experimental linguistics endeavors to perform controlled manipulations to illuminate how human beings produce and process speech and language. The ability to generalize experimental results is predicated upon the validity and representativeness of the experimental paradigm used. As in the case of our tests capturing different aspects of functional hearing (discussed above), scientific experimentation involves the restriction of a complex, dynamic, multifaceted phenomenon to a controlled manipulation along an isolated dimension of interest. Each experiment, therefore, represents only a glimpse of the entire phenomenon in its natural state. Given that the dimensions of variability of a phenomenon often interact with each other, this reduction can only provide a limited understanding of a phenomenon's true complexity.

The validity of scientific reduction is of particular concern for something as complex and fluid as language. Not only is language itself incredibly variable and complex, but it is also embedded in a matrix of interpersonal relationships, societal expectations, personal identity, and many more contextual variables moderating its use. In addition to scientific inquiry, these factors have a significant effect on behavior including the functional hearing outcomes considered here (e.g. familiar voices being more resistant to masking: Holmes and Johnsrude 2019; for review see Smith and Kampfe 1997). While naturalistic observation may at first appear to be preferable, it presents several problems of its own. The first is the effect the observer's presence has on the behavior of the observed. Secondly, while the data collected may be more valid and natural, subsequent analysis typically requires constraining variability to consider only a targeted question. Lastly, the contextual variability weakens the ability to compare experimental results across participants. While laboratory science may remedy the last issue, we must remember that its solution is to create an equally unnatural environment for all participants.

5.3.1 Issues of Validity Concerning the Tests Used Here

The first relevant aspect of validity is test validity, that is, that the test measures what it claims to measure. While this has been an issue of debate in the field, what is of principal importance here is not the absolute

measure of its test validity, but its test validity compared to existing tests. As the goal of this research is to improve upon existing measures of the same construct, improved validity and specificity for measuring that construct alone will provide strong evidence of increased test validity. Serving as a baseline, Waters and Caplan (2003) found the canonical reading span task to be a poor classifier of working memory ability across its various incarnations. When divided broadly into high-, medium-, and low-span groups (a common division scheme, e.g. see Just and Carpenter 1992), only 6.5% of participants were assigned to the same group across all tests. Thus, much room for improvement exists in future personalization of the tests.

One aspect of language-based testing that is of particular concern is parallel-forms reliability. Given that linguistic test items can only be used once, test creators must generate several forms of the test whose equivalence must be experimentally verified (Kalikow et al., 1977; Morgan et al., 1981). The personalized testing considered here necessitates test items differing for each individual. Because each individual's language model differs, it is plausible that varying the assessment tool of a latent construct should adjust test specifics to each subject. However, the creation of individualized forms introduces additional concerns about external validity (i.e. the degree to which the test will generalize to the broader population). In this respect, it amplifies the need to consider parallel forms reliability. I discuss below how its parallel forms reliability might be quantitatively assessed.

With personalization, the content validity (i.e. that the test captures a representative sample of its testing domain) also becomes more difficult to ensure across parallel versions. While we can ensure properly representing the proxy measure of internal language models by design of the sampling method, no such guarantee exists for individuals' actual language models. Fundamentally, however, this is really an issue of proxy measure representativeness for an individuals' language model. To this end, I again claim that no absolute validity is needed here, simply an improvement compared to current methods. Current tests are normed against standard corpora (e.g. Marcus 1993; Baayen et al. 1996) which obtain their language content from general sources. However, by the very nature of standardization, the corpus is necessarily far from every individual's language experience. Our method of personalization, by contrast, attempts to approximate individuals' language experiences as closely as possible, creating more representative proxies.

Given this increased proxy representativeness, samples drawn from our proxies should in turn increase representativeness of an individual's language experience compared with standard methods, resulting in higher content validity.

The last issue of validity to discuss here concerns the generalizability of results obtained from our study participants to other samples drawn from the general population. Given the limited time and population access of the current work (particularly chapters 2 and 3), determining the validity across all subsets of the population is left to further work. However, a concerted effort was made in chapter 4 to adequately sample along dimensions that have traditionally confounded standard versions of the tests (race, age, SES). Despite this effort, we acknowledge that our samples represent only 2 subsets of a very diverse general population and our results should be received accordingly.

5.3.2 Linguistic Issues Jeopardizing Validity

Several previous attempts to build systems aiding in stimuli selection (e.g. Lahl and Pietrowsky 2006; van Casteren and Davis 2007) address issues raised about experiment design (Cutler, 1981; Forster, 2000). These systems rely on norms from machine-readable databases (e.g. Baayen et al. 1996; Coltheart 1981). The databases contain quantitative linguistic information such as corpus word frequency (e.g. from Thorndike (1944); Kučera and Francis (1967)), word length, phonemic and syllabic lengths, age of acquisition, and part of speech. They also contain experimentally obtained ratings of qualitative concepts such as familiarity, concreteness, and imageability (Clark and Paivio, 2004; Toglia and Battig, 1978; Gilhooly and Logie, 1980). Quantitative alternatives to these ratings exist using distributional analyses to model behavioral judgments for semantic features (e.g. Ursino et al. (2018); Devlin et al. (1998)), but are not currently used. The systems produce lists of lexical items optimally matched along dimensions such as those above that the experimenter deems relevant.

These previous approaches have several shortcomings, however. Firstly, if the experiment necessitates more than isolated word presentation, the researcher must create sentential contexts for the target items on their own. This lack of rigor is particularly troublesome in tests such as SPiN, where non-target

words play a critical role in the task. Secondly, the use of a single static corpus (or several static corpora) amplifies the sampling error due to repeated use of a single language sample. Lastly, the mismatch between corpus materials and everyday language experience means the test lacks ecological validity: a single corpus cannot simultaneously reflect every participant's language experience. To address these issues, in chapter 4 we undertake the first steps toward a reexamination and modernization of the method and frequency of corpus construction and subsequent test item creation. Because our items are selected or generated as whole sentences, we avoid the problem of human-created sentential frames. Additionally, more frequent construction of corpora will address the issue of noisy data amplification because repeated aggregated sampling will underweight idiosyncrasies that might be present in any given sample (Efron, 1979). Lastly, the mismatch between language experience and corpus makeup is addressed by ensuring that only representative language samples are included in constructing a participant's corpus.

These improvements to traditional testing methods can increase test equity, subsequently improving both clinical and experimental validity. This validity relies on the central tenet of research that stimuli represent an accurate sample of participants' language. If this is true, the results of a study on how participants processed experimental stimuli can be validly generalized to how they process everyday language (Coleman, 1964). This generalization is not a foregone conclusion, and significant efforts must be taken to ensure the experiment's construct validity (Clark, 1973). Accomplishment of the proposed method will provide a means of alleviating the burden of controlled stimuli construction and allow previously infeasible levels of exhaustive rigor to be applied to the process.

Several issues exist in the old approach that our technique does not solve, however. The first is that familiarity might be a more accurate metric than lexical frequency (Forster, 2000). This discrepancy may arise from differences between participants' actual language experiences and the frequentist statistics of the norming corpora (Zevin and Seidenberg, 2002). While our models do not explicitly model frequency as such, they are statistical in nature and do not account for any dimension of familiarity. Secondly, given the contrived and highly edited (and censored) writing process that media sources must undertake before being

published, it is possible that any reflection of participants' underlying language model has been bleached out by prescriptive formality.

5.3.3 Non-linguistic Issues Jeopardizing Validity

In addition to the content of the test, the manner in which it is administered can leave significant artifacts in the test results. The procedures of test administration here are equivalent with existing methods, so any change in validity should stem from the test's content. Despite this, the existing methods and our methods share confounds that jeopardize the absolute validity of all these laboratory methods.

Firstly, there are effects simply due to participants' knowledge of being observed (Hawthorne effects: McCarney et al. 2007). This alone can cause participants to behave differently than they do naturally. Secondly, the environment of a research lab on a university or hospital campus can have a significant effect on participants' state of mind. This is particularly true when the surroundings produce different associations for members of the research community (e.g. students) and community members from the surrounding area. If participants expected to do better or worse on the tasks, we may have unknowingly captured Golem or Pygmalion effects (Babad et al., 1982; Rosenthal and Jacobson, 1968). Lastly, in addition to differences evoked by the surroundings, a related factor of stereotype threat is at play when participants are recruited from markedly different populations (Steele and Aronson, 1995; Shih et al., 1999). If participants are aware that they were recruited as part of a particular population, their behavior can be informed by stereotypes about that population. While we did not make it explicitly known to participants that they belonged to an a priori population, students likely know their membership to a group whose cognitive abilities are presumed to be high. Additionally, community members may have been acutely aware of their non-membership to the university community as they traveled through the university gates, passed and entered academic buildings, and crossed paths with university members. They may also have been aware of the historical test inequity that has been propagated against their specific demographic community (Thaler et al., 2015; Thaler and Jones-Forrester, 2013).

These confounds are certainly undesirable in test administration. If future testing, when put into practice, must occur in environments similar to our lab, developing the test in our lab may not necessarily be harmful. Cronbach and Meehl (1955) suggest examining test reliability in its robustness to manipulations it might encounter in future implementation. Given the high plausibility that the test will be administered to a diverse population in a laboratory environment, observing its robustness in this context may have virtues in establishing its future reliability.

5.4 Future Directions

5.4.1 Exposure Context of Recurrent Masking Pattern

The repeated grouping along offline information we observe in chapters 2 and 3 was initially unexpected and leads to 2 possible explanations. The first is that because the recurrent masking pattern's presence (rather than repetition) is the primary driver of difference, participants may be leveraging pattern A's appearance as 2/3 of masking patterns (in most experiments) to assume every trial's masking pattern will be pattern A. Even in experiments where pattern A comprised only 1/3 of trials, participants would still consider pattern A to be the most frequent masker by far. This would explain why participants only perform differently when they unexpectedly encounter the random masking pattern in the baseline condition. An alternative explanation is that participants, through their exposure to pattern A over the course of the experiment learn to better recognize pattern A when it occurs as the masking pattern and thus overcome its masking effects .

If the first explanation is correct, only when pattern A occurs in the trial-final position should participants learn to better overcome its masking effects. In contrast, the second explanation allows pattern A to occur anywhere in the trial as long as participants gain exposure over the course of the experiment. We therefore propose an experiment in which the first half contains pattern A as often as in the experiments performed here but not as the masking pattern. In the second half, we will test their performance

as in previous experiments. If the exposure theory is correct, participants' performance should not significantly differ from previous experiments. If, however, the assumption theory is correct, the lack of pattern A's occurrence as a masker during the first half of the experiment should negatively impact participants' performance in the second half.

5.4.2 Beat Tracking and Attention

Given the reprised finding of online pattern recognition in chapters 2 and 3, an examination of the phenomenon's reliance on rhythmic mechanisms is warranted. In order for the online information to be extracted from the patterns and used to provide top-down information, it is likely that a pattern recognition system is recruited. In addition to this pattern recognition system, the auditory attention system most likely plays a part while participants are focused on the task.

In future work, participants' beat tracking abilities should be measured using a task such as the beat alignment task (Iversen and Patel, 2008). This may help explain significant individual variability in participant performance and likely represents a better target than the musical training we previously pursued (Honing et al., 2009; Slater and Kraus, 2016; Slater et al., 2018).

We can also incorporate attentional measures into our tasks by having participants perform them under attentional load and comparing performance with how they are currently performed. This attentional manipulation is likely to have an effect on the recognition of our repeated maskers (like the repeated maskers of Chait et al. 2012). Additionally, the repetitive structure of our maskers hints that susceptibility to attentional capture may represent another component of the individual variability we observe here (for review see Jones et al. 2010).

5.4.3 Individual Variability Covariates

Our future work will aim to further probe our novel masking paradigm's relationship with pure energetic and pure informational masking. One aspect we have not yet examined is the correlation between participants' performance on pure masking tasks and our maskers. It is plausible that participants' ability to

overcome canonical maskers will be predictive of their performance on our tasks, although it is not clear which canonical masker will be a more informative predictor and the degree to which the two canonical types will correlate with each other.

Lastly, while we were unable to identify covariates predicting task performance here, we still anticipate the existence of informative covariates. These include involvement with rhythmic activities (e.g. dance), exposure to regular noise in development (e.g. neighborhood sirens), a self-report of sensitivity to noise, and working memory and inhibitory control tasks in the auditory modality.

5.4.4 Creation and Validation of Personalized Test Stimuli

The results from chapter 4 illuminate the necessity to personalize language test stimuli. Luckily, the results also hint at a possible path for accomplishing this personalization: the models found to correlate with behavioral performance can be used to select sentences that are acceptable fits to a participant's language model. If personalization is considered too extreme, a first step may be using the models to evaluate a set of candidate stimuli from extant behavioral test stimuli and select the ones that are minimally biased across participants.

The canonical test can be administered after the personalized version to allow for the comparison of participant performance on current test materials versus those generated from their language models. The hypothesis of personalization is that participants should perform equally well or better on a personalized version of a test than the standard version. If participants are found to perform significantly better on personalized versions than standardized versions (e.g. by a paired t-test), this would serve as strong evidence of the virtues of personalization.

5.4.5 Establish Repeatability (Test-Retest Reliability)

In addition to increasing the validity of the test procedure, we hypothesize that personalizing stimuli will increase test-retest reliability (Cronbach, 1972). This is due to the fact that a participant's score more accurately reflects the underlying construct of interest rather than being an artifact of the particular test

items. This reduction of sampling noise should reduce variability between test applications as the underlying construct is unlikely to change in such a short span. To study test-retest reliability, participants of the personalized stimuli study above can be invited back 6-8 weeks after initial participation and tested on a newly personalized set of test items.

Previous work by Waters and Caplan (2003) assessed the test-retest reliability of various working memory measurements, including the reading span and its derivatives. They calculated reliability as the product-moment correlation (Pearson's r) between performance on tasks taken approximately two months apart. They note that most correlations are significant, but below the desirable range of $.8 - .9$ (Anastasi, 1982), and that only the sentence span met the de facto minimum acceptable cutoff for retest reliability of .7 (Nunnally, 1978). Although test-retest reliability is traditionally measured using the correlation between participants' scores across instances, this metric does not properly capture the desired property of reliability (Altman and Bland, 1983; Webb et al., 2006).

Altman and Bland (1983) (see also Bland and Altman 1986) correctly note that the Pearson product-moment correlation reveals only the strength of the linear relationship between two random variables, and not in fact the agreement of their values. This is a proper measure in correlation analyses, whose aim is to observe whether a relationship exists between two different variables often without specificity of the proportions by which they must be related. This is strictly not the case, however, for test-retest reliability analysis. In test-retest analysis we require that the values agree with one another, and thus values *must* be in a 1:1 relationship. Therefore, product-moment correlation, which captures only the strength of the relationship between sessions' values, is not an adequate measurement for the equality of their values. While evidence of low correlation between two sessions certainly casts doubt on the validity of the test, Bland and Altman (1986) note that with regards to high correlation, it would be remarkable if two applications of the same method of measurement were not related.

5.5 Conclusion

5.5.1 Contributions to Perception

All the tests used in this work involve heavy influences of top-down predictive mechanisms to actively perceive the target stimuli. This top-down active perception is central to the experimental manipulations in chapters 2 and 3 in that the observed effects depend on active prediction of upcoming auditory stimuli. This active prediction is present in chapter 4 in the lexical domain as well, in the form of predictability of subsequent words given a participant's language model. Most of the experiments involve systematic manipulations (i.e. through regularity or personalization) of what are often conceptualized of as the distractor stimuli in their paradigm: masking noise in our word recognition task and the language processing (rather than storage) demands of the reading span task. Affording participants the ability to overcome distractors in these ways can tell us about the degree to which distractor items are perceived and processed, and provides insight into the automaticity of perception. Given their role as distractors, they are often not studied with as much intensity as their corresponding target constructs. This de-emphasizes the crucial role they play in their respective tasks and has significant effects on performance, as shown here. As naturalistic perception never occurs in ideal laboratory conditions, the ability to assess the role of interfering demands on perceptual processing can provide a more ecologically valid understanding of people's perceptual abilities in their everyday sensory environment.

Results from chapters 2 and 3 provide insight into the question of perceptual automaticity raised here. It provides a possible additional reason why speech perception in noise is difficult. While traditional conceptualizations of speech perception in noise have attributed the increased difficulty in adverse listening conditions to either obstruction of the speech signal (i.e. energetic masking) or confounding of attentional filtering and perceptual grouping processes (i.e. informational masking), if the distracting stimulus also recruits automatic cognitive resources to attempt to perceive and make sense of it, it is diverting these resources away from the perceptual processes that require them to operate.

This work also has deep parallels in perceiving noisy signals. While the noise in the SPiN and word recognition tasks is quite obvious, the slight mismatch between a comprehender's language model and a tranmitter's language model may also be conceptualized of as noise (see also MacDonald and Christiansen 2002). The task of decoding a signal with which the receiver is mostly familiar but occasionally has trouble receiving, can also be thought of as establishing a communication channel in which encoding and decoding schemes (representing interlocutors' language models) differ only slightly. The extent to which slight differences in the encoding and decoding schemes effect the fidelity of the signal's transmission quality and bandwidth is an interesting question and provides a novel framework for examining communicative burden. Conceptualized in this way, communicative burden takes the form of increased processing load to change one's default coding scheme. Should the sender take up the burden, they accommodate their encoding scheme (language model) to the receiver's decoding scheme, thus increasing their overall processing load. The same logic applies in reverse to the receiver's accommodation. Additionally, if the bandwidth of the channel is constrained relative to normal communication, information can take longer to transmit, taking additional effort and resources from both sender and receiver.

5.5.2 Contributions to Memory

Chapter 4's tests and language processing demands (including lexical retrieval) have fairly obvious applications to memory. Additionally, in a less obvious fashion, the word recognition tasks in the first two chapters also rely heavily on different components of memory. Because the presentation modality is auditory, it necessarily involves echoic memory. Given the time span of the stimulus, coupled with its need for temporal and pattern recognition computations to be carried out, it will also necessarily enter into the phonological loop of working memory. One interesting thing to note is that under the conceptualization of the phonological loop, it is only possible to rehearse linguistic information to maintain it in working memory. Given that the masking patterns are not linguistic, they should not be able to be rehearsed in the phonological loop. Yet, they have a greater duration than echoic memory and we have found evidence that

participants utilize offline information across trials. This implies that in order to treat the pattern as a coherent item and process its information, it must be stored somewhere. This means that either the information has traveled out of working memory into long-term memory and is being recalled on subsequent trials or that a reexamination is needed of limiting working memory rehearsal to strictly linguistic information. If the long-term memory account is at play here, participants are likely relying on similar mechanisms as those in recognition memory which classify stimuli as familiar or not (for review, see Wixted 2007).

It remains unclear, however, given Baddeley and Hitch (1974)'s working memory model, how the effects of linguistic experience observed here would fit into a symbolic account of language processing and a capacity account of working memory. Perhaps increased stimuli familiarity implies that the strength of memory traces to the lexicon should be greater for familiar items than for items encountered more rarely. However, a simpler and less ad hoc explanation would be that the connectionist view of memory and language processing (e.g. McClelland and Rumelhart (1989)) is indeed what underlies the observed effect. Given that connectionism appears to be the most parsimonious account, this work contributes evidence in its favor. Additionally, this work sheds light on the retrieval of lexical items from long-term memory through partial cue matching of glimpses (in the word recognition tasks and SPiN task) or concurrent processing demands (as in the span task).

5.5.3 Contributions to Verbal Working Memory

The work contained in chapter 4 provides an observational, ecologically valid extension of the results of Wells et al. (2009). Rather than changing subjects' experience to match language comprehension materials, the degree their experience matched comprehension materials correlated with their performance. The next step in this research represents the complementary approach to Wells et al. (2009): instead of tailoring experience to stimuli, we will tailor stimuli to experience. This would allow increased ecological validity (and the plausibility of its translational implementation) while still affording the opportunity for experimental manipulation. The manipulation lies in observing the performance of participants whose language

materials have been matched to their personal experience, and comparing this to a control group of individuals matched along relevant dimensions (e.g. age, socio-economic status, etc.) who received materials tailored to their matched partner and *not* to them individually (propensity score matching: Rosenbaum and Rubin 1983). Should participants in the target group perform significantly better than their matched counterparts, an additional piece of experimental evidence will be gained for MacDonald and Christiansen 2002's experience account of language comprehension.

Further, participants should also perform better on personalized test items than on standardized language materials, as these do not reflect their personal experience as closely. Considerable effort has gone into the development of population-specific norm performance scores for standardized language comprehension tests (e.g. Hispanics and African-Americans: Norman et al. 2011, see Thaler and Jones-Forrester 2013; Thaler et al. 2015 for reviews). These norms are necessary because performance on standardized tests varies systematically with population, and valid norms are needed for comparison. However, the use of a standardized test created for one population and applied to another is fraught with complications. It should be stressed that the effects of these language differences on task performance cannot simply be remedied by transforming score distributions from the original population to the population of interest, particularly if the latter represents a broadly defined heterogeneous population (e.g. African-Americans and Hispanics: Thaler et al. 2015; Benuto and Leany 2015; Thaler and Jones-Forrester 2013; Llorente 2008). Doing so ignores important dimensions of variance not originally accounted for as well as the non-linear relationship between the dimensions of test performance and language experience. The test may not measure the same constructs in the second population as the first, and even if the constructs are the same, the dynamic range of natural variability may be artificially constrained by suboptimal test items. Additionally, quality-control measures such as parallel forms reliability testing (e.g. Morgan et al. 1981) must be reassessed with respect to the new population.

An alternative to this model lies in creating a personalization system for cognitive testing and verifying its validity, particularly in generalizing to new participants. The methodology undertaken in chapter 4 illuminates the feasibility of this new method. Because of the individualized and dynamic nature of

participants' corpora, the ability to correctly and quickly model language experience allows for the test to be matched to a particular individual at a particular time. After the system has been created and tested, it can then be used to generate stimuli of a uniform difficulty (and difficulty range) rather than uniform content. As this sidesteps the issue of population norming, large-scale studies deriving new norms for each population can be avoided. Additionally, this new approach acknowledges that participants do not fit neatly into a single population and not all members of a particular population have the same language experience. Given the fairness and fitness this methodology promises, we expect the test creation method to not only decrease the onus of test creation but also increase the fairness and validity of the resulting test.

5.5.4 Refining Cognitive Testing

Creation of language norms began as an onerous undertaking, expensive in both time and labor costs needed for construction. As a result, norming was performed infrequently and by just a few individuals. Furthermore, whether reflecting selection bias given researchers' demographics or the availability of printed materials, the norms were constructed from samples (e.g. the Bible, Black Beauty, and Little Women: Thorndike 1944) drawn from a narrow sub-type of the broad English language community: "Mainstream" literary English. To the extent that any literary material is an acceptable approximation of everyday language, these norms are heavily biased towards those who operate in the English represented in the corpus. Given the paucity of available norming materials, later clinical tests (e.g. Daneman and Carpenter 1980; Kalikow et al. 1977) and psycholinguistic experiments seeking to normalize their test items necessarily relied on these earlier materials despite their outdatedness and bias. These tests may subsequently be introducing bias with a significant effect on clinical assessments and research results despite best efforts to neutralize testing bias and ensure test equity.

One well-documented example of this is the mini-mental state exam (MMSE; Folstein et al. 1975). The MMSE is the most frequently utilized basic screening measure for screening cognitive impairment in older adults. However, the test is not without criticism (for review, see: Tombaugh and McIntyre 1992; Mitchell 2009; Folstein et al. 1975), with some calling for its retirement (Carnero-Pardo, 2014) citing, among other

concerns, its high confounding with SES variables (see also Strauss et al. 2006). Additionally, several studies have found the test to have high verbal bias (Lancu and Olmer, 2006; Starr, 2010; Tombaugh and McIntyre, 1992). In fact, of the test's 30 possible points, 29 require language processing or production, with 20 requiring explicit language processing or production (the remaining 9 simply relying on language to probe the construct). Evidence of the test's reliance on patients' language models can be found in the numerous instances of bilinguals scoring significantly lower than monolinguals (Anderson et al., 2017; Ní Chaoimh et al., 2015; Bialystok and Luk, 2012; Ivanova and Costa, 2008), despite that bilingualism may in fact delay the onset of dementia (Alladi et al., 2013; Craik et al., 2010) by providing additional cognitive reserve (Stern, 2002). Additionally, systematic variance has been found along the lines of demographic variables such as race and SES, which is highly confounded with language model differences (e.g. for review see Benuto and Leany 2015; Clark et al. 2005; Chiodo et al. 1994; Mejia et al. 2004). The MMSE is far from being unique: of the 28 tests evaluated in a comprehensive overview of cognitive battery tests to diagnose dementia and mild cognitive impairment, 16 contained strong verbal components (Mitchell and Malladi, 2010).

In addition to the personalization of test items, the testing procedure or experience can also be greatly improved by incorporating each patient's language model. One example is in clinician-patient interaction. In addition to subtle power dynamics inherent in this interaction, special care must be taken on the part of the clinician to adequately understand the patient and adequately be understood by the patient (Thaler et al., 2015). For example, having test instructions that are mismatched to the patient's language model may cause Golem effects, in which the patient performs worse on the measure because they anticipate that the test will poorly assess them (Rosenthal and Jacobson 1968; see also Steele and Aronson 1995). Considering that systematic language differences may well exist between the clinician and patient, the ability to accommodate test instructions, environment, and interactions can mitigate some of the systematic biases inherent to the testing procedure.

Bibliography

Aguinis, H. and Whitehead, R. (1997). Sampling variance in the correlation coefficient under indirect range restriction: Implications for validity generalization. *Journal of Applied Psychology*, 82(4):528–538.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, 98(23):13367–13372.

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.

Alexander, R. A., Carson, K. P., Alliger, G. M., and Barrett, G. V. (1984). Correction for Restriction of Range when Both X and Y are Truncated. *Applied Psychological Measurement*, 8(2):231–241.

Alladi, S., Bak, T. H., Duggirala, V., Surampudi, B., Shailaja, M., Shukla, A. K., Chaudhuri, J. R., and Kaul, S. (2013). Bilingualism delays age at onset of dementia, independent of education and immigration status. *Neurology*.

Altman, D. G. and Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *The Statistician*, 32(3):307.

American Speech-Language-Hearing Association (2013). Issues in Ethics: Cultural and Linguistic Competence.

Anastasi, A. (1982). *Psychological testing*. Macmillan ; Collier Macmillan, New York; London.

Anderson, J. A. E., Saleemi, S., and Bialystok, E. (2017). Neuropsychological assessments of cognitive aging in monolingual and bilingual older adults. *Journal of Neurolinguistics*, 43:17–27.

Anderson, J. R. (2013). *The Architecture of Cognition*. Cognitive science series. Psychology Press, Hillsdale, NJ, US.

Andreou, L.-V., Kashino, M., and Chait, M. (2011). The role of temporal regularity in auditory segregation. *Hearing Research*, 280(1):228–235.

Arbogast, T. L., Mason, C. R., and Kidd, G. J. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, 112(5):2086–2098.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In Spence, K. W. and Spence, J. T., editors, *Psychology of Learning and Motivation*, volume 2, pages 89–195. Academic Press.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1996). CELEX2.

Babad, E. Y., Inbar, J., and Rosenthal, R. (1982). Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology*, 74(4):459–474.

Baddeley, A. (1986). *Working memory.* Oxford psychology series. Clarendon Press/Oxford University Press, New York, NY, US.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Baddeley, A., Lewis, V., Eldridge, M., and Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General*, 113(4):518.

Baddeley, A. D. and Hitch, G. (1974). Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*.

Bartlett, M. S. (1937). Properties of Sufficiency and Statistical Tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 160(901):268–282.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bentin, S., McCarthy, G., and Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, 60(4):343–355.

Benuto, L. T. and Leany, B. D., editors (2015). *Guide to Psychological Assessment with African Americans*. Springer New York, New York, NY.

Bernabei, R., Bonuccelli, U., Maggi, S., Marengoni, A., Martini, A., Memo, M., Pecorelli, S., Peracino, A. P., Quaranta, N., Stella, R., Lin, F. R., and participants in the Workshop on Hearing Loss and Cognitive Decline in Older Adults (2014). Hearing loss and cognitive decline in older adults: questions and answers. *Aging Clinical and Experimental Research*, 26(6):567–573.

Best, V., Ozmeral, E., Gallun, F. J., Sen, K., and Shinn-Cunningham, B. G. (2005). Spatial unmasking of birdsong in human listeners: Energetic and informational factors. *The Journal of the Acoustical Society of America*, 118(6):3766–3773.

Bialystok, E. and Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual adults. *Bilingualism: Language and Cognition*, 15(2):397–401.

Bizley, J. K. and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews. Neuroscience*, 14(10):693–707.

Bland, J. M. and Altman, D. G. (1986). Statistical Methods For Assessing Agreement Between Two Methods Of Clinical Measurement. *Lancet*, pages 307–310.

Boersma, P. and Weenink, D. (2012). Praat: doing phonetics by computer.

Borghini, G. and Hazan, V. (2018). Listening Effort During Sentence Processing Is Increased for Non-native Listeners: A Pupillometry Study. *Frontiers in Neuroscience*, 12.

Bowen, R. W., Pola, J., and Matin, L. (1974). Visual persistence: Effects of flash luminance, duration and energy. *Vision Research*, 14(4):295–303.

Boxtel, M. v., Beijsterveldt, C. v., Houx, P., Anteunis, L., Metsemakers, J., and Jolles, J. (2000). Mild Hearing Impairment Can Reduce Verbal Memory Performance in a Healthy Adult Population. *Journal of Clinical and Experimental Neuropsychology*, 22(1):147–154.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. MIT Press, Cambridge, Mass.

Brown, J. (1958). Some Tests of the Decay Theory of Immediate Memory. *Quarterly Journal of Experimental Psychology*, 10(1):12–21.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120(6):4007.

Burk, M. H. and Humes, L. E. (2007). Effects of Training on Speech Recognition Performance in Noise Using Lexically Hard Words. *Journal of Speech Language and Hearing Research*, 50(1):25.

Burk, M. H. and Humes, L. E. (2008). Effects of Long-Term Training on Aided Speech-Recognition Performance in Noise in Older Adults. *Journal of Speech Language and Hearing Research*, 51(3):759.

Byrne, M. D. (1998). Taking a computational approach to aging: The SPAN theory of working memory. *Psychology and Aging*, 13(2):309–322.

Calandruccio, L. and Smiljanic, R. (2012). New Sentence Recognition Materials Developed Using a Basic Non-Native English Lexicon. *Journal of Speech, Language, and Hearing Research*, 55(5):1342–1355.

Calcus, A., Colin, C., Deltenre, P., and Kolinsky, R. (2015). Informational masking of speech in dyslexic children. *The Journal of the Acoustical Society of America*, 137(6):EL496–EL502.

Capaldi, E. J. and Neath, I. (1995). Remembering and forgetting as context discrimination. *Learning & Memory*, 2(3-4):107–132.

Caplan, D. and Walters, G. S. (1996). Syntactic Processing in Sentence Comprehension Under Dual-Task Conditions in Aphasic Patients. *Language and Cognitive Processes*, 11(5):525–51.

Caplan, D. and Waters, G. S. (1990). Short-term memory and language comprehension: A critical review of the neuropsychological literature. In *Neuropsychological impairments of short-term memory.*, pages 337–389. Cambridge University Press, New York, NY, US.

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology*, 27(1):115–127.

Carnero-Pardo, C. (2014). Should the Mini-Mental State Examination be retired? *Neurología (English Edition)*, 29(8):473–481.

Carpenter, P. A. (1994). Working memory constraints in comprehension: Evidence from individual differences, aphasia, and aging. *Handbook of psycholinguistics*, pages 1075–1122.

Carpenter, P. A. and Just, M. A. (1983). What your eyes do while your mind is reading. In *Eye movements in reading*, pages 275–307. Elsevier.

Carroll, J. B. (1971). Defining Language Comprehension: Some Speculations. In *Proceedings of the Research Workshop on Language Comprehension and the Acquisition of Knowledge*, Durham, North Carolina.

Caspari, I., Parkinson, S. R., LaPointe, L. L., and Katz, R. C. (1998). Working Memory and Aphasia. *Brain and Cognition*, 37(2):205–223.

Cazden, C. B. (1988). *Classroom Discourse: The Language of Teaching and Learning*. Heinemann, Portsmouth, NH.

Chait, M., Ruff, C. C., Griffiths, T. D., and McAlpine, D. (2012). Cortical responses to changes in acoustic regularity are differentially modulated by attentional load. *NeuroImage*, 59(2):1932–1941.

Chang, F., Dell, G. S., Bock, K., and Griffin, Z. M. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of psycholinguistic research*, 29(2):217–230.

Charity, A. H., Scarborough, H. S., and Griffin, D. M. (2004). Familiarity with school English in African American children and its relation to early reading achievement. *Child development*, 75(5):1340–1356.

Chen, J., Li, H., Li, L., Wu, X., and Moore, B. C. J. (2012). Informational masking of speech produced by speech-like sounds without linguistic content. *The Journal of the Acoustical Society of America*, 131(4):2914–2926.

Cherry, C. (1957). *On human communication: a review, a survey, and a criticism*. Technology Press of Massachusetts Institute of Technology, United States.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5):975.

Chiodo, L. K., Kanten, D. N., Gerety, M. B., Mulrow, C. D., and Cornell, J. E. (1994). Functional status of Mexican American nursing home residents. *Journal of the American Geriatrics Society*, 42(3):293–296.

Chomsky, N. (1965). Aspects of the theory of syntax. *Cambridge, MA: MITPress*.

Christensen, C. B., Christensen-Dalsgaard, J., and Madsen, P. T. (2015). Hearing of the African lungfish (Protopterus annectens) suggests underwater pressure detection and rudimentary aerial hearing in early tetrapods. *Journal of Experimental Biology*, 218(3):381–387.

Christian, D. (1997). *Vernacular Dialects in U.S. Schools.* ERIC/CLL, 1118 22nd Street N.

Clark, C. M., DeCarli, C., Mungas, D., Chui, H. I., Higdon, R., Nuñez, J., Fernandez, H., Negrón, M., Manly, J., Ferris, S., Perez, A., Torres, M., Ewbank, D., Glosser, G., and Belle, G. v. (2005). Earlier Onset of Alzheimer Disease Symptoms in Latino Individuals Compared With Anglo Individuals. *Archives of Neurology*, 62(5):774–778.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359.

Clark, J. M. and Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):371–383.

Clopper, C. G. and Bradlow, A. R. (2008). Perception of Dialect Variation in Noise: Intelligibility and Classification. *Language and Speech*, 51(3):175–198.

Clopper, C. G. and Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, 48:633–644.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Elsevier, Hoboken.

Cohen, M. A., Evans, K. K., Horowitz, T. S., and Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychonomic bulletin & review*, 18(3):586–591.

Cohen, M. A., Horowitz, T. S., and Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):6008–6010.

Coleman, E. B. (1964). Generalizing to a Language Population. *Psychological Reports*, 14(1):219–226.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

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

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562.

Cord, M. T., Leek, M. R., and Walden, B. E. (2000). Speech recognition ability in noise and its relationship to perceived hearing aid benefit. *Journal of the American Academy of Audiology*, 11(9):475–483.

Courtland, M., Goldstein, L., and Zevin, J. D. (2019). Speech perception with temporally patterned noise maskers. *In Prep*.

Craik, F. I., Bialystok, E., and Freedman, M. (2010). Delaying the onset of Alzheimer disease. *Neurology*, 75(19):1726.

Craik, F. I. M. and Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 12(6):599–607.

Cronbach, J. and Meehl, P. (1955). Construct Validity In Psychological Tests. *Psychological Bulletin*, 52:281–302.

Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, pages 1–33.

Cusack, R., Decks, J., Aikman, G., and Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):643–656.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10(1):65–70.

Daneman, M. (1988). Word knowledge and reading skill. In *Reading research: Advances in theory and practice, Vol. 6.*, pages 145–175. Academic Press, San Diego, CA, US.

Daneman, M. and Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466.

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika*, 9(3):185–197.

Davis, F. B. (1968). Research in Comprehension in Reading. *Reading Research Quarterly*, 3(4):499–545.

Dehghani, M., Sagae, K., Sachdeva, S., and Gratch, J. (2014). Analyzing Political Rhetoric in Conservative and Liberal Weblogs Related to the Construction of the "Ground Zero Mosque". *Journal of Information Technology & Politics*, 11(1):1–14.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological review*, 104(4):801.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8):1117–1121.

Demany, L., Clément, S., and Semal, C. (2001). Does auditory memory depend on attention? In D.J. Breebaart, A.J.M. Houtsma, A. Kohlrausch, V.F. Prijs, and R. Schoonhoven, editors, *Physiological and Psychophysical Bases of Auditory Function*, pages 461–467. Shaker Publishing BV, Maastricht (The Netherlands).

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., and Seidenberg, M. S. (1998). Category-Specific Semantic Deficits in Focal and Widespread Brain Damage: A Computational Account. *Journal of Cognitive Neuroscience*, 10(1):77–94.

Dey, A. and Sommers, M. (2015). Age-Related Differences in Inhibitory Control Predict Audiovisual Speech Perception. *Psychology and aging*, 30.

Dirks, D. D. and Bower, D. R. (1969). Masking Effects of Speech Competing Messages. *Journal of Speech and Hearing Research*, 12(2):229–245.

Divenyi, P. L. and Haupt, K. M. (1997). Audiological Correlates of Speech Understanding Deficits in Elderly Listeners with Mild-to-Moderate Hearing Loss. III. Factor Representation. *Ear and Hearing*, 18(3):189–201.

Dole, M., Hoen, M., and Meunier, F. (2012). Speech-in-noise perception deficit in adults with dyslexia: Effects of background type and listening configuration. *Neuropsychologia*, 50(7):1543–1552.

Dryden, A., Allen, H. A., Henshaw, H., and Heinrich, A. (2017). The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis. *Trends in Hearing*.

Duffy, J. R. and Giolas, T. G. (1974). Sentence Intelligibility as a Function of Key Word Selection. *Journal of Speech, Language, and Hearing Research*, 17(4):631–637.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). Note on informational masking (L). *The Journal of the Acoustical Society of America*, 113(6):2984.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Elman, J. L., editor (2001). *Rethinking innateness: a connectionist perspective on development*. A Bradford book. MIT Press, Cambridge, Mass., 1. mit press paperback ed., 5. print edition.

Erdman, S. A. and Demorest, M. E. (1998). Adjustment to Hearing Impairment II: Audiological and Demographic Correlates. *Journal of Speech, Language, and Hearing Research*, 41(1):123–136.

Ericsson, K. A. and Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2):211–245.

Ericsson, K. A. and Staszewski, J. J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. *Complex information processing: The impact of Herbert A. Simon*, 2:235–267.

Espinoza, O. (2007). Solving the equity–equality conceptual dilemma: a new model for analysis of the educational process. *Educational Research*, 49(4):343–363.

Farnham-Diggory, S. and Gregg, L. W. (1975). Short-term memory function in young readers. *Journal of Experimental Child Psychology*, 19(2):279–298.

Federmeier, K. D., Mai, H., and Kutas, M. (2005). Both sides get the point: Hemispheric sensitivities to sentential constraint. *Memory & Cognition*, 33(5):871–886.

Ferreira, F. and Clifton Jr, C. (1986). The independence of syntactic processing. *Journal of memory and language*, 25(3):348–368.

Festen, J. M. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736.

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4):507–521.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28(7):1109–1115.

Fowler, C. A. (1986). An event approach to a theory of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14:3–28.

Frazier, L. (1987). Sentence processing: A tutorial review. In *Attention and performance 12: The psychology of reading.*, pages 559–586. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Frederiksen, J. R. (1981). Sources of Process Interactions in Reading. Technical Report BBN-4459, Bolt Beranek and Newman Inc., Cambridge MA.

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007). Auditory attention - focusing the searchlight on sound. *Current Opinion in Neurobiology*, 17(4):437–455.

Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., and Pantev, C. (2005). Automatic Encoding of Polyphonic Melodies in Musicians and Nonmusicians. *Journal of Cognitive Neuroscience*, 17:1578–1592.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, 6(1):110–125.

Gilhooly, K. J. and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427.

Glanzer, M., Fischer, B., and Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior*, 23(4):467–486.

Glucksberg, S. and Cowen, G. N. (1970). Memory for nonattended auditory material. *Cognitive Psychology*, 1(2):149–156.

Goldstein, E. B. (2014). *Sensation and perception*. Wadsworth, Cengage Learning, Belmont, CA, ninth edition. edition.

Goldsworthy, R. L. (2015). Correlations Between Pitch and Phoneme Perception in Cochlear Implant Users and Their Normal Hearing Peers. *Journal of the Association for Research in Otolaryngology*, 16(6):797–809.

Green, T. F. (1983). Excellence, Equity, and Equality. In Shulman, L. S. and Sykes, G., editors, *Handbook of Teaching and Policy*, pages 318–341. Longman, Inc., New York.

Gregory, R. L. (2004). Perception. In *The Oxford Companion to the Mind*. Oxford University Press.

Guediche, S., Blumstein, S. E., Fiez, J. A., and Holt, L. L. (2014). Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience*, 7.

Guyer, B. L. and Friedman, M. P. (1975). Hemispheric processing and cognitive styles in learning-disabled and normal children. *Child Development*, pages 658–668.

Haesen, B., Boets, B., and Wagemans, J. (2011). A review of behavioural and electrophysiological studies on auditory processing and speech perception in autism spectrum disorders. *Research in Autism Spectrum Disorders*, 5(2):701–714.

Hall III, J. W., Grose, J. H., Buss, E., and Dev, M. B. (2002). Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children. *Ear and Hearing*, 23(2):159–165.

Harm, M. W. and Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, 106(3):491.

Harry, B. and Klingner, J. (2007). Discarding the deficit model. *Educational Leadership*, 64(5):16.

Hasher, L., Lustig, C., and Zacks, R. (2008). Inhibitory Mechanisms and the Control of Attention. In Conway, A., Jarrold, C., Kane, M., Miyake, A., and Towse, J., editors, *Variation in Working Memory*, pages 227–249. Oxford University Press.

Hauk, O. and Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5):1090–1103.

Hind, S. E., Haines-Bazrafshan, R., Benton, C. L., Brassington, W., Towle, B., and Moore, D. R. (2011). Prevalence of clinical referrals having hearing thresholds within normal limits. *International Journal of Audiology*, 50(10):708–716.

Hitch, G. J. and Baddeley, A. D. (1976). Verbal Reasoning and Working Memory. *Quarterly Journal of Experimental Psychology*, 28(4):603–621.

Holcomb, P. J. and Neville, H. J. (1990). Auditory and Visual Semantic Priming in Lexical Decision: A Comparison Using Event-related Brain Potentials. *Language and Cognitive Processes*, 5(4):281–312.

Hollingworth, A. (2005). The Relationship Between Online Visual Representation of a Scene and Long-Term Scene Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):396–411.

Holmes, E. and Johnsrude, I. (2019). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. preprint, PsyArXiv.

Honing, H. (2012). Without it no music: beat induction as a fundamental musical trait. *Annals of the New York Academy of Sciences*, 1252(1):85–91.

Honing, H., Ladinig, O., Háden, G. P., and Winkler, I. (2009). Is Beat Induction Innate or Learned?: Probing Emergent Meter Perception in Adults and Newborns using Event-related Brain Potentials. *Annals of the New York Academy of Sciences*, 1169(1):93–96.

Hosemann, J., Herrmann, A., Steinbach, M., Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2013). Lexical prediction via forward models: N400 evidence from German Sign Language. *Neuropsychologia*, 51(11):2224–2237.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. The Macmillan Company.

Hunt, E., Frost, N., and Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In *Psychology of learning and motivation*, volume 7, pages 87–122. Elsevier.

Ihlefeld, A. and Shinn-Cunningham, B. (2008a). Disentangling the effects of spatial cues on selection and formation of auditory objectsa). *The Journal of the Acoustical Society of America*, 124(4):2224–2235.

Ihlefeld, A. and Shinn-Cunningham, B. (2008b). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America*, 123(6):4369–4379.

Ivanova, I. and Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica*, 127(2):277–288.

Iversen, J. R. and Patel, A. D. (2008). The beat alignment test (BAT): Surveying beat processing abilities in the general population. In *in Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC10*, pages 465–468.

Iyer, N., Brungart, D. S., and Simpson, B. D. (2007). Effects of periodic masker interruption on the intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 122(3):1693–1701.

Jackson, M. D. and McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General*, 108(2):151–181.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10(4):409–416.

Johnson, K. and Mullennix, J. W. (1997). Speech Perception without Speaker Normalization. In *Talker variability in speech processing*, pages 145–165. Academic Press, San Diego.

Jones, M. R., Fay, R. R., and Popper, A. N., editors (2010). *Music perception*. Number v. 36 in Springer handbook of auditory research. Springer, New York.

Joseph, J. E. and Newmeyer, F. J. (2012). 'All languages are equally complex': The rise and fall of a consensus. *Historiographia Linguistica*, 39(2-3):341–368.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329.

Just, M. A. and Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1):122–149.

Just, M. A., Carpenter, P. A., and Keller, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review*, 103(4):773–780.

Kaernbach, C. (2004). The Memory of Noise. *Experimental Psychology*, 51(4):240–248.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5):1337–1351.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 400–401.

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., and MacDonald, M. C. (1998). Sentence Comprehension Deficits in Alzheimer's Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. *Brain and Language*, 64(3):297–316.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2008). Informational Masking. In Yost, W. A., Popper, A. N., and Fay, R. R., editors, *Auditory Perception of Sound Sources*, Springer Handbook of Auditory Research, pages 143–189. Springer US, Boston, MA.

Kidd, G. J. and Wright, B. A. (1994). Improving the detectability of a brief tone in noise using forward and backward masker fringes: Monotic and dichotic presentations. *The Journal of the Acoustical Society of America*, 95(2):962–967.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5):580–602.

Kintsch, W. and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-Thought Vectors. *arXiv:1506.06726 [cs]*.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., and Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE*, 8(1):e53398.

Kutas, M. and Federmeier, K. D. (2010). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1):621–647.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Labov, W., Ash, S., and Boberg, C. (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Mouton de Gruyter, Berlin ; New York.

Lahl, O. and Pietrowsky, R. (2006). EQUIWORD: A software application for the automatic creation of truly equivalent word lists. *Behavior Research Methods*, 38(1):146–152.

Lancu, I. and Olmer, A. (2006). The minimental state examination–an up-to-date review. *Harefuah*, 145(9):687–90, 701.

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, 10(4):477–493.

Lander-Portnoy, M. (2016). Release from Energetic Masking Caused by Repeated Patterns of Glimpsing Windows. In *Interspeech 2016*, pages 1672–1676.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):451–468.

Lavie, N. (2000). Selective attention and cognitive control: Dissociating attentional functions through different types of load. In Monsell, S. and Driver, J., editors, *Attention and Performance XVIII*, pages 175–194. M I T Press.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, 9(2):75–82.

Lavie, N. (2010). Attention, Distraction, and Cognitive Control Under Load. *Current Directions in Psychological Science*, 19(3):143–148.

Leek, M. R., Brown, M. E., and Dorman, M. F. (1991). Informational masking and auditory attention. *Perception & Psychophysics*, 50(3):205–214.

Leibold, L. J. (2012). Development of Auditory Scene Analysis and Auditory Attention. In Werner, L., Fay, R. R., and Popper, A. N., editors, *Human Auditory Development*, volume 42, pages 137–161. Springer New York, New York, NY.

Leibold, L. J., Yarnell Bonino, A., and Buss, E. (2016). Masked Speech Perception Thresholds in Infants, Children, and Adults. *Ear and Hearing*.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Liberman, M. C., Epstein, M. J., Cleveland, S. S., Wang, H., and Maison, S. F. (2016). Toward a Differential Diagnosis of Hidden Hearing Loss in Humans. *PLOS ONE*, 11(9):e0162726.

Licklider, J. C. R. and Miller, G. A. (1948). The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, 20(4):593–593.

Lim, S. and Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive science*, 35(7):1390–1405.

Llorente, A. M., editor (2008). *Principles of Neuropsychological Assessment with Hispanics*. Issues of Diversity in Clinical Neuropsychology. Springer New York, New York, NY.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Lowry, R. (2019). Two Correlation Coefficients.

MacDonald, M. C. and Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109(1):35–54.

MacDonald, M. C., Just, M. A., and Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, 24(1):56–98.

Marcus, M. (1993). Building a Large Annotated Corpus of English: The Penn Treebank:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

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

Martin, R. C. (1995). Working memory doesn't work: A critique of Miyake et al.'s capacity theory of aphasic comprehension deficits. *Cognitive neuropsychology*, 12(6):623–636.

Martin, R. C., Shelton, J. R., and Yaffee, L. S. (1994). Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. *Journal of Memory and Language*, 33(1):83–111.

Masutomi, K., Barascud, N., Kashino, M., McDermott, J. H., and Chait, M. (2016). Sound segregation via embedded repetition is robust to inattention. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):386–400.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7-8):953–978.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., and Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC Medical Research Methodology*, 7:30.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86.

McClelland, J. L. and Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press.

Mejia, S., Gutierrez, L. M., Villa, A. R., and Ostrosky-Solis, F. (2004). Cognition, Functional Status, Education, and the Diagnosis of Dementia and Mild Cognitive Impairment in Spanish-Speaking Elderly. *Applied Neuropsychology*, 11(4):194–201.

Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of verbal Learning and verbal Behavior*, 2(1):1–21.

Merker, B. H., Madison, G. S., and Eckerdal, P. (2009). On the role and origin of isochrony in human rhythmic entrainment. *Cortex*, 45(1):4–17.

Middlebrooks, J. C. and Green, D. M. (1991). Sound Localization by Human Listeners. *Annual Review of Psychology*, 42(1):135–159.

Miller, G. A. (1947). The masking of speech. *Psychological bulletin*, 44(2):105.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Miller, G. A. and Licklider, J. C. R. (1950). The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, 22(2):167–173.

Mitchell, A. J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research*, 43(4):411–431.

Mitchell, A. J. and Malladi, S. (2010). Screening and Case Finding Tools for the Detection of Dementia. Part I: Evidence-Based Meta-Analysis of Multidomain Tests. *The American Journal of Geriatric Psychiatry*, 18(9):759–782.

Moberly, A. C., Lowenstein, J. H., Tarr, E., Caldwell-Tarr, A., Welling, D. B., Shahin, A. J., and Nittrouer, S. (2014). Do Adults With Cochlear Implants Rely on Different Acoustic Cues for Phoneme Perception Than Adults With Normal Hearing? *Journal of Speech Language and Hearing Research*, 57(2):566.

Moore, M. and Gordon, P. C. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods*, 47(4):1095–1109.

Morgan, D. E., Kamm, C. A., and Velde, T. M. (1981). Form equivalence of the speech perception in noise (SPIN) test. *The Journal of the Acoustical Society of America*, 69(6):1791–1798.

Munakata, Y., McClelland, J. L., Johnson, M. H., and Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological review*, 104(4):686.

National Science Foundation (2013). Women, Minorities, and Persons with Disabilities in Science and Engineering: 2013: (558442013-001). Technical report, American Psychological Association.

Navon, D. (1984). Resources—A theoretical soup stone? *Psychological review*, 91(2):216.

Neath, I. (2000). Learning and memory: In humans. *Encyclopedia of psychology*, 5:16–19.

Neath, I. and Surprenant, A. M. (2005). Mechanisms of Memory. In *Handbook of Cognition*, pages 222–239. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom.

Ng, B. S. W., Schroeder, T., and Kayser, C. (2012). A Precluding But Not Ensuring Role of Entrained Low-Frequency Oscillations for Auditory Perception. *The Journal of Neuroscience*, 32(35):12268–12276.

Nikolić, D., Mureşan, R. C., Feng, W., and Singer, W. (2012). Scaled correlation analysis: a better way to compute a cross-correlogram. *European Journal of Neuroscience*, 35(5):742–762.

Norman, D. A. (1969). Memory While Shadowing. *Quarterly Journal of Experimental Psychology*, 21(1):85–93.

Norman, M. A., Moore, D. J., Taylor, M., Franklin, D., Cysique, L., Ake, C., Lazarretto, D., Vaida, F., Heaton, R. K., and the HNRC Group (2011). Demographically corrected norms for African Americans and Caucasians on the Hopkins Verbal Learning Test–Revised, Brief Visuospatial Memory Test–Revised, Stroop Color and Word Test, and Wisconsin Card Sorting Test 64-Card Version. *Journal of Clinical and Experimental Neuropsychology*, 33(7):793–804.

Norris, D., McQueen, J. M., and Cutler (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3):299–325.

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, 42(4):313–329.

Näätänen, R., Kujala, T., Kreegipuu, K., Carlson, S., Escera, C., Baldeweg, T., and Ponton, C. (2011). The mismatch negativity: an index of cognitive decline in neuropsychiatric and neurological diseases and in ageing. *Brain*, 134(12):3435–3453.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118(12):2544–2590.

Ní Chaoimh, D., De Bhaldraithe, S., O'Malley, G., Mac Aodh Bhuí, C., and O'Keeffe, S. T. (2015). Importance of different language versions of cognitive screening tests: Comparison of Irish and English versions of the MMSE in bilingual Irish patients. *European Geriatric Medicine*, 6(6):551–553.

Oden, G. C. and Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85(3):172–191.

Otten, M., Nieuwland, M. S., and Berkum, J. J. V. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, 8(1):89.

Paglialonga, A., Grandori, F., and Tognol, G. (2013). Using the Speech Understanding in Noise (SUN) Test for Adult Hearing Screening1. *American Journal of Audiology (Online)*, 22(1):171–4.

Patterson, K. E., Seidenberg, M. S., and McClelland, J. L. (1989). Dyslexia in a distributed, developmental model of word recognition. In Morris, R., editor, *Parallel distributed processing*. Oxford, Clarendon Press, Oxford, England.

Pearlmutter, N. J. and MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of memory and language*, 34(4):521–542.

Pekkarinen, E., Salmivalli, A., and Suonpää, J. (1990). Effect of Noise on Word Discrimination by Subjects with Impaired Hearing, Compared with Those with Normal Hearing. *Scandinavian Audiology*, 19(1):31–36.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). Across-Language Perspective on Speech Information Rate. *Language*, 87(3):539–558.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perfetti, C. A. (1985). *Reading ability.* Reading ability. Oxford University Press, New York, NY, US.

Perfetti, C. A. and Goldman, S. R. (1976). Discourse memory and reading comprehension skill. *Journal of Verbal Learning and Verbal Behavior*, 15(1):33–42.

Perfetti, C. A. and Lesgold, A. M. (1977). Discourse Comprehension and Sources of Individual Differences. Technical report, Learning Research and Development Center, Pittsburgh Univ., PA.

Peterson, L. and Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, 58(3):193.

Petsas, T., Harrison, J., Kashino, M., Furukawa, S., and Chait, M. (2016). The effect of distraction on change detection in crowded acoustic scenes. *Hearing Research*, 341:179–189.

Pollack, I. (1955). Masking by a Periodically Interrupted Noise. *The Journal of the Acoustical Society of America*, 27(2):353–355.

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, 57(S1):S5–S5.

Posner, M. I. (1966). Components of skilled performance. *Science*, 152(3730):1712–1718.

Povel, D.-J. and Essens, P. (1985). Perception of Temporal Patterns. *Music Perception: An Interdisciplinary Journal*, 2(4):411–440.

Prosser, S., Turrini, M., and Arslan, E. (1990). Effects of different noises on speech discrimination by the elderly. *Acta Oto-Laryngologica. Supplementum*, 476:136–142.

Reali, F. and Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 57(1):1–23.

Reid, D. K. and Valle, J. W. (2004). The Discursive Practice of Learning Disability: Implications for Instruction and Parent—School Relations. *Journal of Learning Disabilities*, 37(6):466–481.

Riecke, L., Formisano, E., Herrmann, C. S., and Sack, A. T. (2015). 4-Hz Transcranial Alternating Current Stimulation Phase Modulates Hearing. *Brain Stimulation*, 8(4):777–783.

Rizzo, N. D. (1939). Studies in Visual and Auditory Memory Span with Special Reference to Reading Disability. *The Journal of Experimental Education*, 8(2):208–244.

Roberts, L. E., Martin, W. H., and Bosnyak, D. J. (2011). The Prevention of Tinnitus and Noise-Induced Hearing Loss. In Møller, A. R., Langguth, B., De Ridder, D., and Kleinjung, T., editors, *Textbook of Tinnitus*, pages 527–534. Springer New York, New York, NY.

Rosen, S., Souza, P., Ekelund, C., and Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, 133(4):2431–2443.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenthal, R. and Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development*. Holt, Rinehart and Winston, New York.

Roth, F. P. (1984). Accelerating language learning in young children. *Journal of Child Language*, 11(1):89–107.

Rönnberg, J., Rudner, M., Lunner, T., and Zekveld, A. (2010). When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health*, 12(49):263–269.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, 2(9):437–442.

Sanford, A. J. and Garrod, S. C. (1981). *Understanding written language: Explorations of comprehension beyond the sentence*. John Wiley & Sons.

Schneider, E. W. and Kortmann, B., editors (2004). *A handbook of varieties of English: a multimedia reference tool*. Mouton de Gruyter, Berlin ; New York.

Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. S. (2004). A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *The Journal of the Acoustical Society of America*, 115(2):813–821.

Seidenberg, M. S. and MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23(4):569–588.

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3):114–123.

Sharkey, N. E. and Mitchell, D. C. (1985). Word recognition in a functional context: The use of scripts in reading. *Journal of Memory and Language*, 24(2):253–270.

Sheft, S. and Yost, W. A. (2008). Method-of-adjustment measures of informational masking between auditory streams. *The Journal of the Acoustical Society of America*, 124(1):EL1–EL7.

Shih, M., Pittinsky, T. L., and Ambady, N. (1999). Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance. *Psychological Science*, 10(1):80–83.

Slater, J. and Kraus, N. (2016). The role of rhythm in perceiving speech in noise: a comparison of percussionists, vocalists and non-musicians. *Cognitive Processing*, 17(1):79–87.

Slater, J., Kraus, N., Woodruff Carr, K., Tierney, A., Azem, A., and Ashley, R. (2018). Speech-in-noise perception is linked to rhythm production skills in adult percussionists and non-musicians. *Language, Cognition and Neuroscience*, 33(6):710–717.

Smiarowski, R. A. and Carhart, R. (1975). Relations among temporal resolution, forward masking, and simultaneous masking. *The Journal of the Acoustical Society of America*, 57(5):1169–1174.

Smith, S. M. and Kampfe, C. M. (1997). Interpersonal relationship implications of hearing loss in persons who are older. *Journal of Rehabilitation; Alexandria*, 63(2):15–21.

Snyder, J. S., Alain, C., and Picton, T. W. (2006). Effects of Attention on Neuroelectric Correlates of Auditory Stream Segregation. *Journal of Cognitive Neuroscience*, 18(1):1–13.

Sohoglu, E. and Chait, M. (2016a). Detecting and representing predictable structure during auditory scene analysis. *eLife*, 5.

Sohoglu, E. and Chait, M. (2016b). Neural dynamics of change detection in crowded acoustic scenes. *NeuroImage*, 126:164–172.

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., and Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160105.

Soveri, A., Tallus, J., Laine, M., Nyberg, L., Bäckman, L., Hugdahl, K., Tuomainen, J., Westerhausen, R., and Hämäläinen, H. (2013). Modulation of Auditory Attention by Training. *Experimental Psychology*, 60(1):44–52.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied*, 74(11):1.

Springfield! Springfield! (2019). https://www.springfieldspringfield.co.uk.

St John, M. and Gernsbacher, M. (2013). Learning and losing syntax: Practice makes perfect and frequency builds fortitude. *Foreign language learning: Psycholinguistic experiments on training and retention*, pages 231–255.

Starr, J. M. (2010). Cognitive impairment in older adults: A guide to assessment. *Clinical Medicine*, 10(6):579–581.

Steele, C. M. and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5):797–811.

Stenbäck, V., Hällgren, M., Lyxell, B., and Larsby, B. (2015). The Swedish Hayling task, and its relation to working memory, verbal ability, and speech-recognition-in-noise. *Scandinavian Journal of Psychology*, 56(3):264–272.

Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, 8(3):448–460.

Stevens, K. N. (1968). *The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data*. Unknown Publisher.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–1891.

Stone, M., Gabrieli, J. D. E., Stebbins, G. T., and Sullivan, E. V. (1998). Working and strategic memory deficits in schizophrenia. *Neuropsychology*, 12(2):278–288.

Stone, M. A., Füllgrabe, C., and Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, 132(1):317–326.

Stone, M. A. and Moore, B. C. J. (2014). On the near non-existence of "pure" energetic masking release for speech. *The Journal of the Acoustical Society of America*, 135(4):1967–1977.

Strauss, E., Sherman, E. M. S., Spreen, O., and Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. Oxford University Press, Oxford ; New York, 3rd ed edition.

Stromquist, N. (2005). Comparative and International Education: A Journey toward Equality and Equity. *Harvard Educational Review*, 75(1):89–111.

Sun, H.-M. and Gordon, R. D. (2010). The influence of location and visual features on visual object memory. *Memory & Cognition*, 38(8):1049–1057.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). LSTM Neural Networks for Language Modeling. In *INTERSPEECH*, page 4.

Sörqvist, P. (2010). The role of working memory capacity in auditory distraction: A review. *Noise and Health*, 12(49):217–224.

Tanner, W. P. J. (1958). What is Masking? *The Journal of the Acoustical Society of America*, 30(10):919–921.

Taylor, B. (2007). Predicting Real World Hearing Aid Benefit with Speech Audiometry: An Evidence-Based Review Brian Taylor.

Teki, S., Chait, M., Kumar, S., Kriegstein, K. v., and Griffiths, T. D. (2011). Brain Bases for Auditory Stimulus-Driven Figure–Ground Segregation. *The Journal of Neuroscience*, 31(1):164–171.

Teki, S., Chait, M., Kumar, S., Shamma, S., and Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *eLife*, 2.

Thaler, N. S. and Jones-Forrester, S. (2013). IQ Testing and the Hispanic Client. In *Guide to Psychological Assessment with Hispanics*, pages 81–98. Springer, Boston, MA.

Thaler, N. S., Thames, A. D., Cagigas, X. E., and Norman, M. A. (2015). IQ Testing and the African American Client. In Benuto, L. T. and Leany, B. D., editors, *Guide to Psychological Assessment with African Americans*, pages 63–77. Springer New York, New York, NY.

Thompson, E. C., White-Schwoch, T., Tierney, A., and Kraus, N. (2015). Beat Synchronization across the Lifespan: Intersection of Development and Musical Experience. *PLOS ONE*, 10(6):e0128839.

Thomson, D. M. and Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of experimental psychology*, 86(2):255.

Thorndike, E. L. (1913). *The psychology of learning*, volume 2. Teachers College, Columbia University.

Thorndike, E. L. (1944). *The teacher's word book of 30,000 words*. New York :.

Tierney, A. T. and Kraus, N. (2013). The ability to tap to a beat relates to cognitive, linguistic, and perceptual skills. *Brain and Language*, 124(3):225–231.

Toglia, M. P. and Battig, W. F. (1978). *Handbook of semantic word norms*. Lawrence Erlbaum Associates ; Distributed by the Halsted Press Division of John Wiley, Hillsdale, N.J.; New York.

Tombaugh, T. N. and McIntyre, N. J. (1992). The Mini-Mental State Examination: A Comprehensive Review. *Journal of the American Geriatrics Society*, 40(9):922–935.

Trainor, L. J., Marie, C., Bruce, I. C., and Bidelman, G. M. (2014). Explaining the high voice superiority effect in polyphonic music: Evidence from cortical evoked potentials and peripheral auditory models. *Hearing Research*, 308:60–70.

Treisman, A. (1964a). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, 3(6):449–459.

Treisman, A. M. (1964b). Selective Attention In Man. *British Medical Bulletin*, 20(1):12–16.

Ursino, M., Cuppini, C., Cappa, S. F., and Catricalà, E. (2018). A feature-based neurocomputational model of semantic memory. *Cognitive Neurodynamics*.

US Census Bureau (2018). American Community Survey Data.

USC Communications (2018). Facts and Figures | About USC.

Valtin, R. (1973). Report of Research on Dyslexia in Children. In *Proceedings of the Annual Meeting of the International Reading Association*, page 12, Denver, Colorado.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443–467.

van Casteren, M. and Davis, M. H. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, 39(4):973–978.

Van Dijk, T. A., Kintsch, W., and Van Dijk, T. A. (1983). *Strategies of discourse comprehension*. Academic Press, New York.

Wang, D. and Brown, G. S. (2006). *Computational auditory scene analysis: principles, algorithms, and applications*. Wiley interscience ;John Wiley [distributor], Hoboken, N.J. :Chichester.

Wang, Y., Huang, M., and Zhao, L. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Wanner, H. E. and Maratsos, M. P. (1978). An ATN approach to comprehension. *Linguistic theory and psychological reality*.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, 167(3917):392–393.

Waters, G. S. (1996). The Measurement of Verbal Working Memory Capacity and Its Relation to Reading Comprehension. *The Quarterly Journal of Experimental Psychology Section A*, 49(1):51–79.

Waters, G. S. and Caplan, D. (1996a). The Capacity Theory of Sentence Comprehension: Critique of Just and Carpenter (1992). *Psychological Review*, 103(4):761–772.

Waters, G. S. and Caplan, D. (1996b). Processing resource capacity and the comprehension of garden path sentences. *Memory & Cognition*, 24(3):342–355.

Waters, G. S. and Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, 35(4):550–564.

Waters, G. S., Caplan, D., and Rochon, E. (1995). Processing capacity and sentence comprehension in patients with Alzheimer's disease. *Cognitive neuropsychology*, 12(1):1–30.

Watkins, M. J. and Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General*, 104(1):5–29.

Webb, N. M., Shavelson, R. J., and Haertel, E. H. (2006). 4 Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics*, volume 26, pages 81–124. Elsevier.

Wells, J., Christiansen, M., Race, D., Acheson, D., and Macdonald, M. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58(2):250–271.

Whitney, P., Arnett, P. A., Driver, A., and Budd, D. (2001). Measuring Central Executive Functioning: What's in a Reading Span? *Brain and Cognition*, 45(1):1–14.

Wicha, N. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of cognitive neuroscience, Journal of Cognitive Neuroscience*, 16, 16(7, 7):1272, 1272–1288.

Widin, G. P. and Viemeister, N. F. (1979). Intensive and temporal effects in pure-tone forward masking. *The Journal of the Acoustical Society of America*, 66(2):388–395.

Wiley, T. L., Cruickshanks, K. J., Nondahl, D. M., Tweed, T. S., Klein, R., and Klein, B. E. (1998). Aging and word recognition in competing message. *American Academy Of Audiology*, 9:191–198.

Wilson, R. H., Carnell, C. S., and Cleghorn, A. L. (2007). The Words-in-Noise (WIN) test with multitalker babble and speech-spectrum noise maskers. *Journal of the American Academy of Audiology*, 18(6):522–529.

Winkler, I., Haden, G. P., Ladinig, O., Sziller, I., and Honing, H. (2009). Newborn infants detect the beat in music. *Proceedings of the National Academy of Sciences*, 106(7):2468–2471.

Winn, M. B. (2016). Rapid Release From Listening Effort Resulting From Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, 20:233121651666972.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, 114(1):152–176.

YIFYSubtitles (2019). http://www.yifysubtitles.com.

Zevin, J. D. and Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, 47(1):1–29.

Zhang, C., Lu, L., Wu, X., and Li, L. (2014). Attentional modulation of the early cortical representation of speech signals in informational or energetic masking. *Brain and Language*, 135:85–95.