Efficacy of Multichannel Audio Versus Stereo in Word Recall

An Interactive Qualifying Project
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
In partial fulfillment of the requirements for the
Degree of Bachelor of Science

By

Thomas M. Tawadros
Harrison M. Hertlein
Peter L. Griffiths
Thai H. Dao
George W. Horta

Date:
7/21/2018

Report Submitted to:
Professor Frederick Bianchi
Worcester Polytechnic Institute

Abstract

This study reports on an experiment testing the efficacy of multichannel audio compared to stereo, or binaural, audio in terms of word recall. When asked to single out and recall words from multiple others, Subjects can focus on and recall no more than one at a time, and perform much worse when more than two words are played at once. Subjects recalled words with an accuracy of about 70%, and displayed increased caution and less confidence when presented with a complicated test prior to an easier one.

Acknowledgements

We would like to thank Professor Frederick Bianchi of the WPI Humanities department for providing us the foundation for this study. His feedback was invaluable to the development of our process. We would also like to thank Professor Vincent Manzo for making our study known to his summer class, and to all the participants that took time out of their summer to take part in our study.

Finally, we would also like to thank Worcester Polytechnic Institute for providing the opportunity and facilities to research audition for our Interactive Qualifying Project.

Executive Summary

Introduction

Humans have a unique relationship with the sense of hearing. It forms the foundation for communication, language, and cognitive function. The importance of hearing abilities in modern society is unparalleled as humans deal with multiple streams of information constantly, utilizing every sense we have to process this information. Scientists have long been researching many aspects of our hearing ability, but little research had been done on how sound localization impacts information intake and processing. Should there be a means to speed up human processing capability, it would be extremely beneficial.

This brings us to our study: We designed an experiment to test and compare the impact of stereo and 4-directional multichannel audio on short-term memory. With the findings of this experiment we hope to highlight useful data that can help in the development of methods to expand short-term memory capacity, and ultimately enhance human information intake.

Methods

The experiment was designed to be carried out in a 10’ 8” by 15’ 10” audio lab. In this space, 4 speakers were set up and placed at the 4 corners of an 8’ 4” square. The test Subject was seated in the center of the square, facing one of the speakers, having one speaker directly at his/her back, one to the left ear and one to the right ear. The test Subject then listened to sets of words played from the speakers and was asked to recall them. Words being played are from the left and right speakers in the stereo test and from all four speakers in the 4-directional multichannel test.

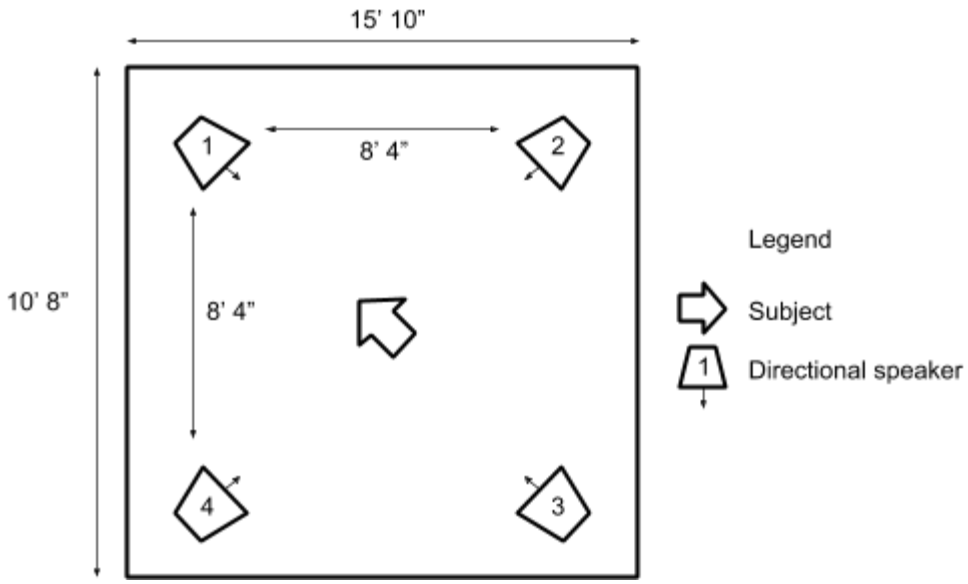


Figure 1: Lab Setup

Findings

Eleven Subjects took part in this study. Participants recalled an average of 25.78 words in the stereo test, and 9.28 words in the multichannel test. Participants correctly recalled an average of 19.07 words in the stereo test while correctly recalling an average of 6.4 words in the multichannel test. The percentage of correct words recalled out of the total words recalled for the total group was 73.95% in the stereo test, while the correct percent recalled was 68.94% in the multichannel test. The number of words recalled incorrectly for the total group averaged 6.72 in the stereo test, while the number recalled incorrectly averaged 2.88 in the multichannel test. The percent of words recalled incorrectly compared to the total number of words recalled for the total

group was 26.05% in the stereo test, while the percent number of words recalled incorrectly was 31.06% in the multichannel test.

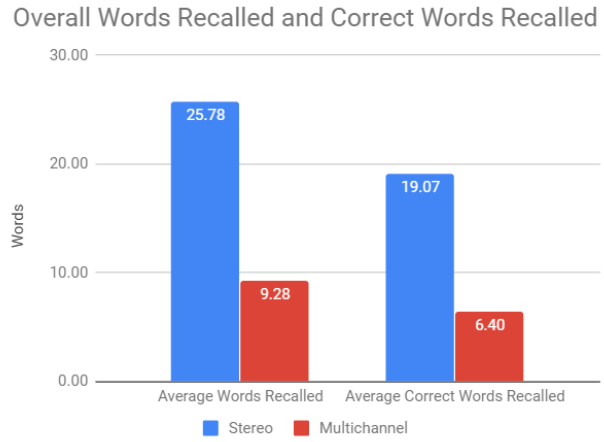


Figure 2: Overall Words Recalled and Correct Words Recalled

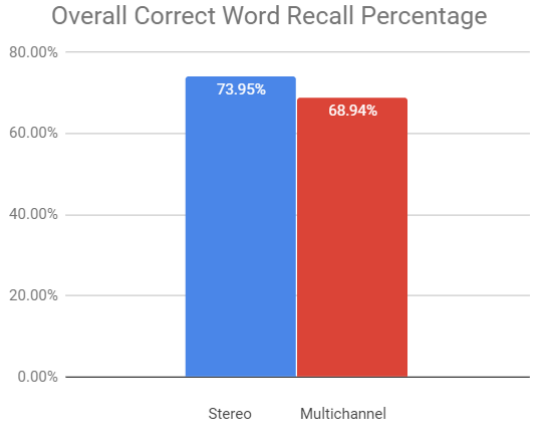


Figure 3: Overall Correct Word Recall Percentage

It is possible that the difference in the words remembered between the experiments is due to the "cocktail party" effect. This is the ability to selectively "tune in" to certain sounds when presented with more than one audio source at once, and may cause an individual to only listen to one source of sound selectively. In the stereo test, it is possible that the cocktail party effect single-handedly decided the number of words recalled by test Subjects; since there were only two words being played at once, the Subject could focus on only one source and recall half the number of words. This theory matches with actual results with a 4.65% difference in total number of words recalled correctly. However, in the 4-directional multichannel audio test this theory doesn't hold, as test Subjects only recalled an average of 6.40 words out of 40 words being played, where theoretically they should be able to recall 10 out 40 words. This is a 36% difference; therefore, the cocktail party effect only could have influenced the result partially. The remaining difference can be attributed to the overwhelming or disorienting nature of the multichannel test, as the sounds may have interfered with one another. It is also possible that the stress of having to distinguish words out of the grouping impeded their ability to remember and distinguish the words.

The results could also indicate possible imbalance in volume of sounds being played from different channels.

Conclusion

Our intention in this experiment was to see if there was a difference in hearing and remembering stereo audio compared to multi-channel audio, the finding could suggest new methods of using multichannel audio to increase learning capabilities.

Based on our collected data, we conclude that multichannel is not as effective as stereo in the simultaneous perception of voices. In the experiment, our test Subjects recalled a significant number more words, and more correct words, during the stereo experiment than the multichannel experiment. With multichannel, our Subjects were more likely to remember an incorrect word than a correct word. We also tried to detect a correlation between musical training and the ability to distinguish and recall words, however we found no convincing pattern.

Recommendations

- Determine limits of human hearing and memory as to what the maximum number of words can be perceived at once is.
- Define and operate under the optimal conditions for hearing multiple voices and perceiving them.
- Repeat this study with a larger sample size to provide more definitive data.
- Develop a better system for the Subjects to convey the words they remembered.

Authorship

	Section	*Author(s)*	*Editor(s)*
	Abstract	George	
	Acknowledgments	Thomas	
	Executive Summary	Thai	All
1	**Introduction**	Thai	All
2	**Background**	Thomas	Peter/Thomas/Harrison/Thai
3	**Methods**	Thomas	Peter/Thomas/Harrison
4	**Findings**	Harrison/George	George/Harrison/Thomas/Peter
5	**Conclusion & Recommendations**	Peter/Harrison	Peter/Harrison/ Thomas /Thai
	Appendix	Thomas	

Table of Contents

List of Figures

Appendix A

Appendix B

Appendix C

List of Tables

1. Introduction

Hearing forms the foundation for communication and spoken language. The effectiveness of our hearing depends greatly on the environment we are in. Our surroundings affect what we hear in a variety of ways: The number and placement of sound sources, the volumes of sounds, and more. All can either improve or impair hearing abilities. The modern society we live in demands a highly specialized auditory system. We constantly deal with many sources of information coming at us: emails, conversations, traffic lights, etc. The human body utilizes all its senses to take in information. It stands to reason, then, that if there is a way to speed up the intake or the processing of information, it would be a tremendous advantage to whomever possesses this ability. Principal in the processing of information is the short-term or working memory. Short-term memory is an integral part of day-to-day cognitive function. Without the luxury of time to commit information to long-term memory, short-term memory briefly allows us to store and access a small amount of important information, with older studies estimating its capacity around 7 items (Miller, 1956).

With human hearing ability and short-term memory as the primary focus, we shifted our attention towards past researches on this Subject. One such influential study was Jens Blauert's "Sound Localization in the Median Plane" (1969) dealing with sound localization in the median plane. However, we noticed that in this previous research the sound system setups were either exclusively in mono, or stereo, or a combination of both; there was no emphasis on the impact or differences they made on the tests. This opened up a new opportunity for further researching into the matter: an experiment designed to test and compare the impacts of stereo and multichannel audio on short-term memory.

This brings us to our study: Our study tested the efficacy of stereo vs. 4-directional multichannel audio in the short-term memory retention of information by using two-syllable, English nouns. Our hypothesis is that multichannel sound is more effective than stereo sound in terms of word recall. With the findings of this study we hope to possibly discover data that would ultimately help developing new ways to speed up human information intake.

In this experiment, we wanted to put a person in a controlled environment with stereo and multichannel audio. Information would be output from these two audio systems and responses from human Subjects would be recorded. This would determine the impact of each setup on

short-term memory. The potential of this study is in its new perspective on the capacity of short-term memory. Many applications could be derived from the findings of this research, such as new methods to expand short-term memory capacity, new techniques to enhance information intake, and more.

2. Background

Our research was inspired by the studies of sound localization and organization, as well as the study of memory. It can be easy to think of these fields as unrelated, but this belies the complex ways in which the aforementioned phenomena interact to produce our sense of hearing. In fact, it's difficult to conceive of the human auditory system missing any one of these components. Our experiment of listening and recalling words in stereo and multichannel configurations necessitated the involvement of all three systems on the part of the Subject. However, it is not the purpose of this study to delineate the path that a sound takes from being heard to being recalled on a multiple-choice questionnaire. Our goal was to explore whether or not the use of multichannel audio improves our ability to recall words.

In order to understand our findings, it is necessary to present some of the theoretical basis for this study, as well as explain in relatively basic terms how hearing works. Therefore, this background in divided into four sections. First is a primer on the physical and mathematical description of sound and relevant human anatomy. Second, we explore the process of sound localization. The third section is dedicated to sound organization and psychoacoustics. Finally, we delve into our modern understanding of short-term memory. In this way, a series of curtains is lifted, one behind the other, presenting successively deeper looks into our ability to both hear and remember what we have heard.

2.1. What is Sound?

From Impairments, National Research Council (US) Committee on Disability Determination for Individuals with Hearing Impairments (2004)

2.1.1. Qualitative Description

Our sense of hearing can be described as the subjective experience of sound. Qualitatively, sound is any pressure variation (or vibration) that propagates through a medium. This medium can be anything, so long it has a physical density that can change. The human auditory system has evolved to best perceive sound in the gaseous medium we call air, but anyone who has spent time underwater will tell you that sound certainly does not stop below sea-level. A few parallels can be drawn between the senses of sight and hearing, vision and audition respectively. The light that humans can see is but a small sliver of the electromagnetic spectrum. Likewise, we can only perceive a small range of vibrations as sound. This range is commonly

cited as 20-20000 Hertz (defined below) but varies between people and degrades over time. Additionally, just as the quality of light changes depending on the material it travels through (glass, water, fog) so too does sound; the different acoustic properties of air and water are manifest when listening to someone yell above versus underwater.

2.1.2. Quantitative Description

According to the National Research Council, mathematical descriptions of sound refer to two domains: time and frequency. In the time domain, because sound is the variation of pressure, it can be modelled as an oscillation of pressure over time. In the frequency domain, a sound is described in terms of the tonal components that make up that sound. A tonal sound is one that can be modelled as a sinusoidal function of pressure over time. Tonal sounds are the building blocks that make up the vastly more complex sounds of daily life. As such they are commonly used as stimulus when studying human audition (Bregman, 1990).

There are three basic attributes of sound that arise from its mathematical representation: frequency, amplitude, and temporal variation. Frequency is the number of times an oscillation occurs per second ($\frac{1}{s} = s^{-1}$), measured in Hertz (Hz). Amplitude is the amount of pressure being exerted and can be represented in a number of ways: physical pressure is given by $p = F/Ar$ where F = force divided by the Ar = area it is applied over. The intensity of a sound is proportional to pressure and is given by $I = p^2/p_o c$ where the square of pressure is divided by the product of c = speed of sound and p_o = density of the sound-carrying medium. (The presence of medium specific density in this equation is one reason we perceive sound differently underwater.) The most common measure of amplitude is in units of Decibels (dB), which can be expressed in terms of either pressure or intensity by $dB = 20\log_{10}(p/p_{ref})$ and $dB = 10\log_{10}(I/I_{ref})$ respectively. A Decibel is a general comparative measure used in a variety of fields besides acoustic. P_{ref} and

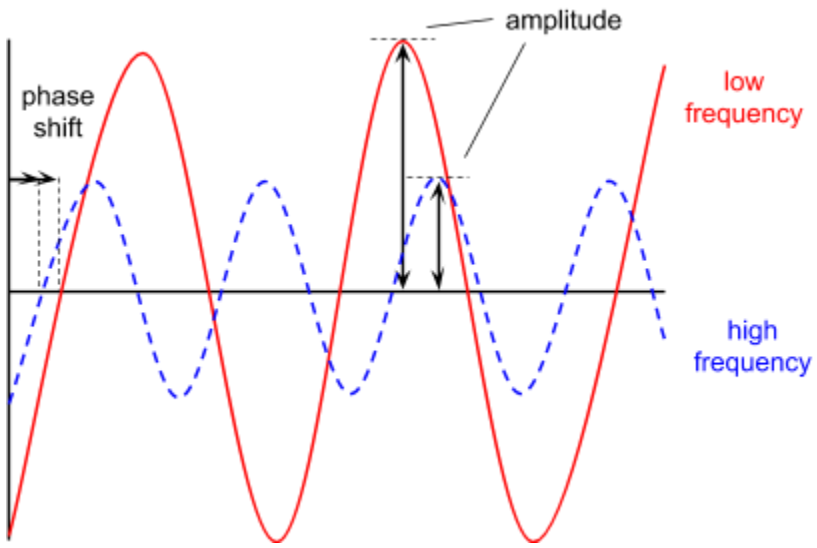


Figure 4: Amplitude, Frequency, and Phase

I_{ref} are reference values typically set to the threshold of human hearing (around 20 micro Pascals or $p = 2 \times 10^{-5}$ Pa). Finally, temporal variation is a catch-all phrase that includes a variety of properties including:

- Duration (t): how long a sound lasts (typically in seconds)
- Phase (angular degrees °): the cycle of periodic change in pressure can be expressed in terms of completing traversal around 360 degrees of a circle. $\theta = 360°(t)(f)$ where t = time and f = frequency
- Tone: introduced above; a simple sound that can be described be a regular sinusoidal oscillation: $A = \sin(2\pi ft + \theta)$ where A = amplitude, f = frequency, t = time, and θ = phase shift

2.1.3. Hearing Anatomy

A transducer is a system that converts energy from one form to another. Thus, the human ear is fundamentally a transducer that converts energy in the form of sound vibrations into neural-electrical energy. When we listen to music or someone speaking, we take for granted the complex machinery that enables our sense of hearing.

Audition, the process of hearing, begins at the external ear, the fleshy protrusions on either side of the head that capture sound from the environment. The structure of the external ear, known as the pinna, has evolved to effectively syphon sound to the middle ear. This is where signal processing begins. The tympanic membrane, colloquially known as the eardrum, is the divider between the external and middle ear. In the middle ear the eardrum, along with three miniscule bones called the ossicles, receives sound vibrations from the ear canal and transfers them to the fluid and soft tissue of the inner ear.

The process of transferring sound waves from one medium to another (air to tissue) is referred to as immittance. Immittance is a combination of two factors: impedance (the reflection of waves) and admittance (the transferal of waves). There is always a level of impedance when transferring sound from a less dense medium to a denser one. Impressively, the “35 dB impedance loss” (2004) when transferring sound from air to ear is almost entirely overcome by structural focusing of the sound waves. Once sound vibrations are captured in the inner ear, the transduction process is performed by the structures of the cochlea. This spiral-shaped organ is divided into multiple compartments and contains sensorineural hair cells that respond to

vibrations in the surrounding fluid. Shearing, or bending, of these hairs produces neuro-electrical potentials, which propagate down the auditory fibers of the cranial nerve.

The type of neural response generated by the cochlea is a function of the frequency, intensity, and time interval of the vibrations that stimulate it. In this way, frequency, intensity, and temporal variation information (the three basic properties of sound) are encoded into the neural response, which is then sent to the central auditory nervous system. This system is responsible for our subjective experience of hearing, a result of signal processing done by many parts of the brain and brainstem. As such, it is also responsible for the processes of sound localization and organization, which are centrally relevant to our study.

2.2. Sound Localization

How do we tell where a sound is coming from? To the average person the answer is intuitive, not a confounding problem. Consider, however, what is known about the auditory system from the previous section. The signals carried by the auditory nerve to the brain contain information about frequency, intensity, and temporal variation. None of these properties tell you where in a three-dimensional environment a sound is emanating from (Brainard, 1992, p. 1). Furthermore, if we consider each ear as a point in space, there is no way to unambiguously determine where a sound is coming from (Blauert, 1969). This problem is known as the cone of confusion, a result of the fact that multiple points in space will produce the same interaural time difference. However, this is assuming that both ears receive and transmit the same information; this is rarely the case.

The basic dimensions used when discussing sound localization are azimuth, elevation and range. Azimuth and elevation refer to the horizontal and vertical angle of the sound source from the listener respectively. Range refers to the distance between the source and the listener. The auditory system uses a different set of cues to determine the location of a sound source in each of these dimensions. These cues include interaural time and level differences (ITD and ILD), and head-related transfer functions. In addition, source

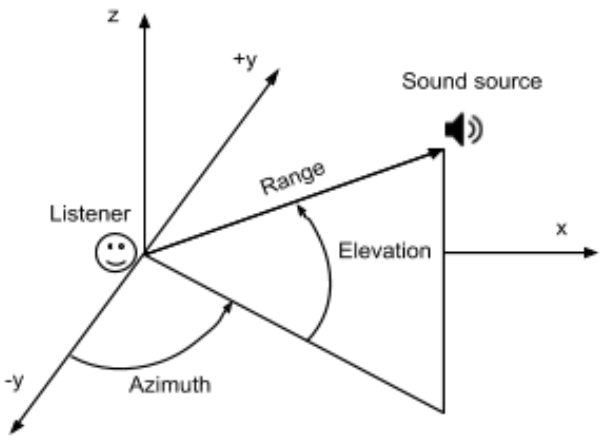


Figure 5: Azimuth, Elevation, and Range

level assumptions, spatial masking-level differences, and audio scene analysis come into play when considering more complex situations (Impairments, 2004).

In 1948 Lloyd Jeffress presented his place theory of sound localization. When listening to sound centered directly in front or behind us, each ear receives sound waves at roughly the same time. This prompts the auditory system to interpret the source as equidistant from both ears. That is, the sound source must be located on a plane that cuts through the center of the head, the medial plane. When a sound source is located off-center, sound will arrive at one ear before the other. Jeffress proposed a neural model for how this difference in arrival time between the ears, ITD, could be used to determine the horizontal angle of a sound source (1948). It has since been shown that humans can perceive a 1 to 3-degree angular change on the horizontal place (Impairments, 2004). In his study of sound localization on the median plane, Blauert (1969) notes that many points in space share the same ITD. This means unambiguous localization of a sound by ITD alone is impossible, a problem known as the cone of confusion. What other factors enable localization?

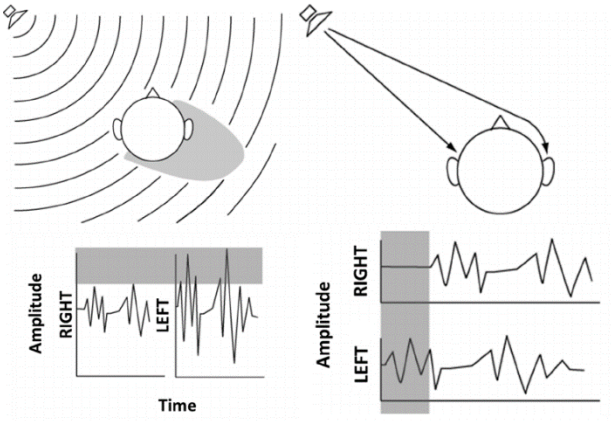


Figure 6: Interaural Time and Level Differences

The difference in level between the ears, ILD, is also used to determine azimuth. When a sound reaches the head from one side, it is effectively "shadowed" from the opposite ear by the head's physical shape. This creates a difference in level (colloquially volume) between the ears and is likewise used to ascertain horizontal angle. Traditionally, ITD is the primary cue for low frequency sounds (below 1500 Hz) and ILD the cue for higher frequency ones (Impairments, 2004). However, more recent studies have questioned the nature of this frequency dependence (Jonides, 2015; Goupell, 2018; Jones et al., 2015).

As a sound travels from its source to the ear canal of the listener, it is distorted in a frequency specific way. A head-related transfer function (HTRF) describes these changes to the frequency spectrum of a sound. Every listener has a HTRF that corresponds, unsurprisingly, to the shape of their head. The changes described by HTRFs are primarily used to locate high frequency sounds. While the lack of ILD or ITD could place a sound source on the medial plane,

the HTRF allows one to distinguish whether the sound source is in front or behind them, as well as the elevation.

Determining the range of a sound requires some knowledge or assumption about the sound at the source. If the listener knows the level at the source, a lower level at the ear will place the sound at a distance proportional to the difference between the two. In addition, a sound that is far away is likely to be reverberated, or reflected, by the ground and other obstacles. The brain uses the reflection-to-direct sound ratio as another perceptual distance cue. Highly reflective environments can pose a problem as they create a complex mixture of source and reflected sound at the ear. It is theorized that in this situation the brain only processes the first instance of the sound, ignoring reflections and thus eliminating confusion (Impairments, 2004).

When attempting to detect a target sound in the presence of another, we can think of the target sound as the "signal" and the other as a "mask" that inhibits detection. The farther apart the sources of signal and mask, the easier it is to detect and locate said signal. This increase of signal detectability with distance from a mask is referred to as spatial mask-level difference. This cue is not simply about locating a sound source. It is also about creating a distinction between two competing sounds. Conditions like this have been extensively recreated and studied in the lab as we are constantly met with similar conditions in real life. In a crowded room sounds emanate from all directions, yet somehow one is able to single out and *listen* to one voice, a clink of glass, or background music. Our ability to do so is bounded by several factors, but this phenomenon was first defined by Cherry (1953) as the cocktail party effect. More recently, Drullman and Bronkhorst (2000) used a 3D auditory display to study speaker recognition in a cocktail party-esque environment. A 3D auditory display presents sounds in a virtual 3D environment around the listener over headphones. This is accomplished by used HRTFs to modulate audio before it reaches the ears. Using this technique, they found that listeners performed better in recognizing two or more simultaneous speakers when they were located in different parts of the virtual environment. In this example (and in the cocktail party effect in general) localization allows a listener to tell where a sound is coming from, but it is not what allows us to segregate two sounds playing at the same time. Why is it that we can perceive multiple people talking at the same time, instead of a jumbled mass of words? These are the types of questions explored in the field of psychoacoustics.

2.3. Organization and Psychoacoustics

We have previously discussed how the auditory system transduces vibrations in the air into neural-electrical potentials, which are then sent to the central auditory nervous system (CANS). The CANS (the brain and brainstem) is ultimately responsible for our experience and perception of hearing. Psychoacoustics is the study of this perception, of how the CANS parses and organizes signals from the ears to form subjective experience. This introduction to psychoacoustics begins with the foundational work of Albert Bregman and the theory of auditory scene analysis. We then discuss the neural storage of sounds and related differences between speech and non-speech storage.

Auditory Scene Analysis (ASA), first published in 1990, is the preeminent text detailing the findings and theory of modern psychoacoustics. In it, author Albert Bregman describes the "auditory process of organization that has evolved, in our auditory systems, to solve a problem that [he referred] to as 'auditory scene analysis'" (p. 3). This was a novel perspective at the time, especially considering the relative youth of the field; "If you were to pick up a general textbook on perception written before 1965 and leaf through it, you would not find any great concern with the perceptual or ecological questions about audition" (1990, p. 1). This is not to say that the field hadn't existed before 1990; The term "psychophysics" was coined in 1860 by Gustav Fechner, and refers to the study of the "relationship between sensory perception (psychology) and physical variables (physics)" (Yost, 2015, p. 43). The physical characteristics of sound (amplitude, frequency) clearly make psychoacoustics a subset of Fechner's broader field. It is, however, only a label given to a field of inquiry dating back to Pythagoras and Aristotle.

Bregman introduces auditory scene analysis (ASA) in terms of representations. The process of perception, he claims, is two-fold. First, representations of the environment must be formed. These representations then inform the behavior used in response to the environment. An integral step in creating representations of the environment "is to decide which parts of the sensory stimulation are telling us about the same environmental object or event" (1990, p. 3). Scene analysis, then, is the process which extracts and perceptually links or separates parts of an environmental stimulus. Throughout his book, Bregman draws parallels between visual and auditory scene analysis. It is due to a variety of factors that the study of vision has received greater focus throughout human history relative to audition. However, the claim that this is because audition is *simpler* than vision is false (1990, p. 2).

As it turns out, auditory scene analysis is highly similar to visual counterpart, which provides an easy way of illustrating the fundamental problem of scene analysis in general. “In vision,” Bregman writes, “you can describe the problem of scene analysis in terms of the correct grouping of regions...but what about the sense of hearing?” (1990, p. 6). While the same organizational principle is at work, the process of “grouping regions” is less obvious in audition than in vision. To illustrate, we are constantly presented with situations similar to that of figure 7. One object hides, or occludes, another object behind it. But this description jumps the gun; the two-dimensional image presents no information about depth. Why does it seem obvious that one object is *behind* the other? How does one distinguish three-dimensional “objects” in this context? What turns regions A and B from disjoint polygons into the side of a rectangular prism?

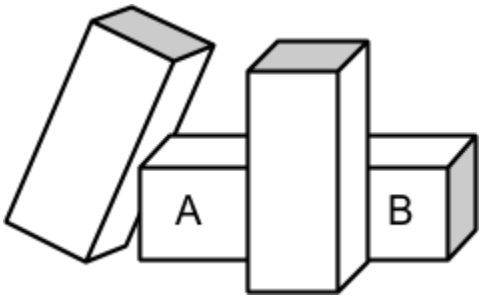


Figure 7: Visual Scene Analysis

The ease with which we assign faces A and B to the same side of the same shape is noteworthy. We are coerced into perceiving a two-dimensional image as a three-dimensional scene. Visual scene analysis can be thought of as this process, where one interprets a third dimension from two. However, this is not to say that the interpretation is necessarily correct. This basic example, while illustrative, only hints at the perceptual machinery that underlies our ability to represent our environment. It is obvious just how much we take for granted when Bregman asks how a computer would go about identifying A and B as parts of the same region (1990, p. 4).

Like the eyes that input photons to form a neural “image,” the ears input vibrations to form a neural “spectrogram” like that of the spoken words “nineteenth century” in figure 8. Both are the raw data which must be interpreted to form a useful representation of the environment.

Auditory scene analysis, then, is the process of deciding which parts of the spectrogram belong together. The result of this process, the equivalent to objects in visual scene analysis, is what Bregman terms "auditory streams." The object/stream comparison is not entirely accurate though. One reason for this is that sound is by definition a temporal phenomenon. If we wish to characterize a sound as something other than a simple tone (defined previously), then it must undergo some kind of temporal variation. This contrasts with our ability to distinguish objects from a static visual scene, like a photograph. Accordingly, Bregman makes a distinction between a physical *sound* and a perceptual *stream.* The former tells us information about an event in the environment, while the latter is an internal representation of the event. Because a single event can produce many sounds over time, our internal representation must be able to combine them into a single, atemporal perceptual entity. Thus, discrete sounds like footsteps produce a perceptual stream that represents a person walking.

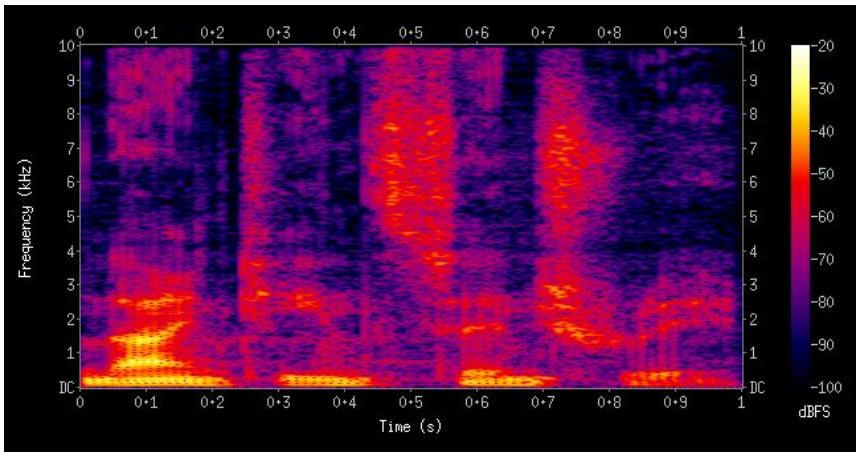


Figure 8: "Nineteenth Century" Spectrogram

Like objects, streams allow the separation of different parts of the environment. They facilitate the reaction to concurrent auditory events as distinct, instead of as discordant noise coming from a single source. They also form the perceptual anchor for descriptions or properties. "[We] say that an *object* is red, or that it is moving fast, that it is near, or that it is dangerous. [The] notion of an object...serves as a center around which our verbal descriptions are clustered" (Bregman, 1990, p. 10). When hearing two concurrent sounds, one high and near, the other low and far, streams allow us to perceive them as two different events by virtue of the disparate properties of each stream (this is not to say that they are *necessarily* two different events--a single event might have produced both). This process is appropriately termed stream segregation (Bregman, 1990).

We have discussed how sound travels from the air to the CANS, and defined an auditory stream, but how does sound get integrated into an auditory stream? Bregman proposed two

methods by which we solve the problem of ASA: sequential and spectral integration. In sequential integration, the auditory system uses changes in speed and in the frequency spectrum to inform grouping. Referred to as "horizontal" organization, sequential integration allows for the grouping of longer auditory event like music, as well as grouping sounds based on frequency proximity. Spectral integration involves the segregation of *simultaneous* sounds into different auditory streams based on spectral and spatial characteristics. This "vertical," or simultaneous, organization creates the cocktail party effect, allowing us to focus on a set of sounds based on their inclusion into a perceptual stream, which itself is being built through sequential integration.

Spectral integration is of particular interest to the present study; if listeners were able to better recall words in multichannel environments, it could only have been because they *perceived* those words being said, instead of a jumbled mass of noise. This perception in turn must occur through stream segregation, which we hypothesized was based on the localization of sound sources. However, Bregman notes that stream segregation based on spatial location is difficult to induce (1990, p. 79). Sounds, unlike everyday objects, reflect off of many surfaces before reaching the ears, obscuring the true location of their source. Instead he suggests that localization could have a "multiplying effect" on perceptual certainty when corroborated by other factors like frequency cues.

Especially relevant to the present study is the organization of speech sounds. It is clear that speech is organized differently than non-speech sounds, largely due to the existence of speech-sound schemas. Bregman defines a schema as set of learned constraints that dictate ASA, as opposed to primitive constraints that are ingrained in our auditory system (1990, p. 38). The two modalities of ASA, sequential and spectral organization, operate based on more or less the same primitive constraints regardless of the listener. However, much of how we perceive speech is schema-based, and depends on the language and its sounds. "The existence of speech-sound schemas, innate or learned, makes it harder to uncover the contribution of primitive organization" (Bregman, 1990, p. 684). Whether primitive or schema-based, we do know some characteristics of speech organization. Sequentially, streaming of speech must operate variably on short and long time-scales, using pitch continuity as a primary cue to anchor a voice in a perceptual stream. Silence is also significant, as it facilitates the recognition of words. In terms of spectral organization, pitch is the most important factor, allowing the dissection of a voice from

background noise as in the cocktail party effect. Moreover, the unique pitch contour of a voice will make it more recognizable than a monotone (Bregman, 1990, p. 690).

In order to test the effects of localization in isolation, we were forced to consider the effects of pitch continuity on ASA in speech. Likewise, all audio in our study was synthesized in the same digital "voice." That is, all words were spoken with the same timbre and pitch, an effective monotone. Each word was easily recognizable on its own, but the value of pitch differences between speakers quickly became apparent when listening to more than one at the same time.

2.4. Memory

Memory is a fundamental cognitive function, and thus one of the principle fields of psychology. Systematic writings on the topic date back 250 years (Winslow, 1861; Squire, 2011). William James (1890) was the first to identify qualitative differences between what we now call long-term and short-term memory. Since then, various models of memory function have been accepted, revised, and rejected. This overview will look at the current understanding of memory function, relevant models, and the experimental difficulties of testing memory. Especially salient is short-term memory and its function in word recall.

One of the first widely accepted theories on memory function was the Atkinson-Shiffrin memory model. Also known as the multi-store model, it asserted a partitioning of memory into three distinct parts: sensory register/buffer, short-term store/memory (STS or STM), and long-term store/memory (LTS or LTM). In this model, the sensory register first receives sensory information. The STS then incorporates information from both sensory register and the LTS. Finally, the LTS houses information for manipulation and alteration in working memory (Atkinson & Shiffrin, 1968).

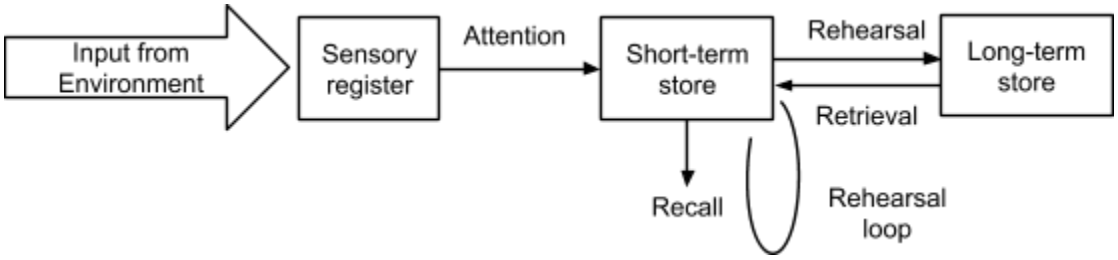


Figure 9: Atkinson-Shiffrin Memory Model

Figure 9 depicts the basic processes that move information between each area of memory. If we follow an environmental stimulus through this model, the first store we encounter is the

sensory register. The sensory register prevents overload of higher cognitive functions by limiting the amount of sensory input. Attention then selectively moves sensory input into the STS. Without selective attention, the STS would be constantly bombarded with new information, defeating the purpose of memory. A multitude of processes occur within the STS that manipulate, reinforce, or otherwise integrate information into the LTS. From LTS, complex retrieval processes move information back into the STS for manipulation.

Over time the Atkinson-Shiffrin model received scrutiny, prompting further investigation. The most influential memory model to date, developed by Baddeley and colleagues, was a response to the Atkinson-Shiffrin model (Baddeley, 1974; Jonides, 2008). Baddeley's working memory model was proposed as an alternative to the short-term store of the Atkinson-Shiffrin model. Likewise, it incorporates the multi-store nature of its predecessor. According to Baddeley's model, the sensory register and short-term store are combined into what he calls working memory. However, Baddeley's model contains different sensory buffers for different types of input. Figure 10 shows the working memory model inserted into the conceptual periphery of the Atkinson model.

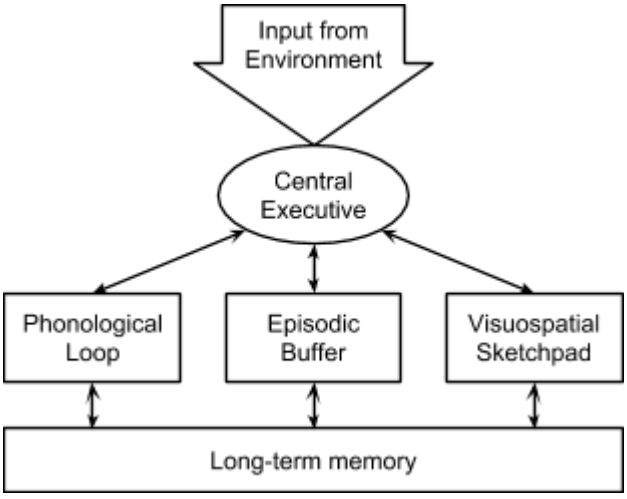


Figure 10: Working Memory Model

The working memory model makes a clear distinction between storage and processing. To use the analogy of a computer, the central executive carries out functions on the information in the sensory buffers much like a CPU functions on the information in RAM. The buffers themselves are divided between the verbal and visual domains: The visuospatial sketchpad is responsible for maintaining visual information, and the phonological loop for information that can be rehearsed verbally (words, numbers). The episodic buffer was later added to account for retention of multimodal information (Jonides, 2008; Baddeley, 2000).

The capacity of STM is notoriously small, and subject to debate. One of the most famous publications on the topic is "The Magical Number Seven Plus or Minus Two" (Miller, 1956). In it, author George Miller estimated the capacity of STM to average seven items, plus or minus

two. There has been debate about what constitutes an item, and around Miller's proposal in general (Murdock, 1962; Tarnow, 2010; Cowan, 2001). Miller himself used so-called "chunks" as the fundamental unit of memory. Chunks vary in size depending on the type of information they contain, be it letters, digits, or words. Likewise, STM capacity will change depending on the type of chunks being remembered.

The most intuitive function of working memory is recall, in which memories arise in consciousness. But storage in working memory is not a free ride, and information does not simply remain in working memory until it's recalled. Unless it is retained in LTM, information in working memory is subject to two major types of corruption: interference and decay. Interference refers to the displacement of items already in memory with new ones. As focus shifts, new sensory items can partially or completely overwrite previous ones. Interference is similarity-based, i.e., the degree of interference is a function of the similarity between the old and new items. Decay, on the other hand, refers to the default state of information loss over time. While intuitively appealing, the concept of decay is controversial. Experiments have found it difficult to eliminate alternative explanations for it, and without a functional framework, it can appear to be merely "a restatement of the problem" (Jonides, 2008).

Another integral function of working memory, then, must be to combat information loss. This process is dubbed rehearsal. When information enters working memory, it can only remain there if it is being regularly reinforced. Otherwise it will be replaced through interference, decay over time, or both. Rehearsal is most commonly performed through audible or sub-audible repetition. This is especially important in the phonological loop. In the Atkinson-Shiffrin model (see Fig. 9) we see that a "rehearsal loop" operates on the STS. This loop is what gave the phonological loop its name; as the buffer for audible information, information is maintained in the phonological loop through audible and sub-audible repetition.

The capacity of the phonological loop to maintain information through repetition is of special significance to this study. In testing how many words our listeners could recall, it was imperative that minimal time was given for rehearsal. This is because, as the study progressed, words heard later on would replace previous ones in memory through interference. If rehearsal was allowed before the answering portion of the experiment, it was theorized, then words heard later on would be remembered better than those heard earlier. It should be said that this study would not be the first to be complicated by the phenomenon of rehearsal, which has been a

difficulty for those attempting to study both memory capacity and input bandwidth for years. Rehearsal is a fairly automatic tendency, which makes it difficult to test information retention where rehearsal is not allowed. Nonetheless, techniques have been developed to inhibit rehearsal, including verbal repetition of a monosyllabic word like “the,” and introducing attention-demanding tasks during rehearsal intervals (times when it is likely that the listener will be rehearsing information) (Jonides, 2008).

3. Methods

3.1. Foundation

A study of clustering in word recall by Bousfield and Cohen included their methodology for preparing word lists. This served as the basis for our methodology. Their goal of testing the effect of categorization on clustering in word recall necessitated controlling for the reinforcement of words presented as stimulus, both during the study (1953) and prior to it (1955). Reinforcement in this context refers to repeated presentation of a word, either during experimentation or prior to it. Bousfield and Cohen reference two of their own reports; the earlier found that reinforcement affected "clustering appreciably beyond chance expectation" (1953), while the later used the Thorndike-Lorge tables to show a similar effect with prior reinforcement outside of a lab setting. In addition, they only used (1956) "two-syllable nouns with Thorndike-Lorge frequencies falling within the range of two to 17 per million" (p. 2). Three of their studies found that both reinforcement (1955), and categorization (1956) of stimulus words have a predictable, positive relationship on clustering in recall.

The aforementioned studies were vision based, requiring Subjects to recall words presented on a projector screen. While visual and auditory perception differ greatly, is seems plausible that these factors (reinforcement and categorization) have at least some, if not a similar, effect on recall when words are presented audibly instead of visually. Likewise, in compiling words for this study, we sought to control for these factors, in order to isolate the effect of sound localization (multichannel audio) on recall capacity.

Two lists of 80 words termed "answer sheets" were created for this study. They were compiled to meet the following requirements:

1. All words are two-syllable nouns
2. All words appear in the top 5,000 most frequent in the Corpus of Contemporary American English (COCA) (Corpus, 2017)
3. No words repeat

The answer sheets were generated with the follow procedure: Unique, two-syllable nouns were generated randomly using an online resource (Random Noun Generator, n.d.). If they

appeared in the top 5,000 on the COCA frequency list, they were added to the answer sheet. Every word on the answer sheets was unique (occurred only once between the two). The answer sheets also corresponded to shorter lists used for audio. Four lists of 10 words each termed "word-lists" were generated from each answer sheet with the following procedure: Ten words were selected at random, then checked for any obvious categorical similarities. If such similarities were found, one of the offending words was replaced with a new random word. The process repeated until the word-list contained 10 words. Every successive word-list was generated from the remaining words on the answer sheet, ensuring any word in the word-lists was never repeated. Checking for categorical similarities is somewhat subjective; the level to which anyone associates a set of words is not easily controlled for, and depends on factors outside of the lab (Bousfield & Cohen, 1956, p. 95). An effort was nonetheless made to avoid superficial similarities between words. The frequency-list used in our study was compiled from the Corpus of Contemporary American English, a more modern and much larger corpus than used for the Thorndike-Lorge tables (Thorndike & Lorge, 1944). As of December 2017, the COCA contains more than 560 million words, compared to about 18 million used to form the Thorndike-Lorge lists.

Table 1: Word-lists

Stereo				**Multichannel**			
1	2	3	4	5	6	7	8
network	profit	complex	pilot	symptom	union	panic	scholar
concept	reserve	visit	organ	merit	driver	album	monster
reason	bedroom	river	orange	rider	hallway	student	peasant
stomach	vision	detail	freshman	number	picture	system	final
package	mayor	support	version	asset	silver	hunter	lover
tourist	surgeon	welfare	bishop	city	artist	runner	ceiling
headline	section	woman	soldier	ladder	protest	cattle	tissue
action	physics	bible	father	license	discount	island	budget
mainstream	sister	highlight	salad	railroad	ally	verdict	planet
measure	country	diamond	regard	ideal	paper	elite	prayer

The audio for each word was generated using Google's WaveNet Text-to-Speech software. This technology allowed for the creation of high quality, natural sounding audio for every word, all in a single voice with very consistent timbre (Cloud, n.d.). Each word-list comprised an audio clip containing 10 words, each separated by 1 second of silence. The audio for individual words differed in length, so each word was given a one second slot in which to play. The stereo test played two words at a time, while the multichannel test played 4 at a time. Thus, the stereo test lasted twice as long as the multichannel test (see App. A). Additionally, a survey filled out prior to the experiment recorded age, sex, gender, hearing ability, and musical training.

3.2. Physical Layout of the Experiment

The "directional" speakers used in this study were model SAT 4.0 MKIII satellite speakers by Blue Sky, part of the MediaDesk 5.1 MKII system. For the purposes of this study, the included subwoofer was used only to route audio, and did not itself play audio. Speakers projected sound as indicated by their respective arrows. Numbers indicate the designation of the speakers throughout this study. Each speaker was 9.5" tall and rested 3' 2" above the floor. Subjects were seated 1' 5.5" off the floor, and faced in the direction of the arrow. The experiment was designed to be carried out in a 10' 8" by 15' 10" audio lab. In this space, 4 speakers were set up and placed at the 4 corners of a 8' 4" square; the distance from the head of the Subject to any speaker was approximately 5' 10". Note that the room itself was not square, but that the configuration below was roughly centered.

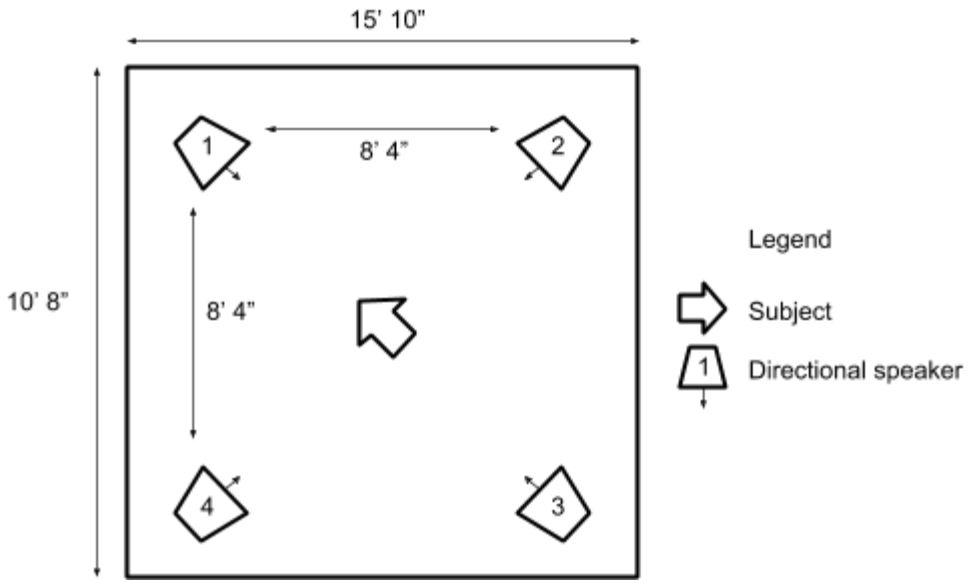


Figure 11: Lab Setup

3.3. Process

Each Subject was given a survey prior to participation (see App. B, Figs. 26-28). This survey recorded age, gender, sex, and hearing ability, as well as any relevant medical conditions. Each Subject was placed into one of two groups (A or B) prior to experimentation. The experiment was comprised of two trials, each with a listening portion and an answering portion. Upon arrival, Subjects were seated in a room with a speaker setup as depicted in figure 11. They were introduced to the experiment with the following prompt:

> In this experiment, we'll be playing a series of words from different locations in the room. Afterwards, we'll give you list of words, and you'll have five minutes to pick out any words that you heard. We'll then repeat the experiment one more time. Try to stay still and face forward while you're listening. You don't have to go completely rigid, but try not to swing your head or body. Any questions?

Subjects were prompted "Ready?" before the start of each listening portion. Four word-lists were used in each of the two trials. In one trial, speakers 1 and 2 each played two word-lists simultaneously. Words from lists 1 and 2 played simultaneously, then words from 3 and 4, repeating (see App. A) This constituted the "stereo" configuration. In the other trial, every speaker played a different word list, again simultaneously (four words, one second pause, repeat)

(see App. A). This constituted the "multichannel" configuration. With reference to figures 24 and 25, word-lists 1-4 were used in the stereo test, and 5-8 in the multichannel test. In the stereo test, speaker 2 played lists 1 and 3, and speaker 4 played lists 2 and 4. In the multichannel test, speakers 1, 2, 3, and 4 played lists 5, 6, 7, and 8 respectively.

Subjects in group A performed the stereo trial first, then the multichannel trial. Those in group B performed the reverse, doing multichannel first, then stereo. No words were repeated throughout the test. The answering portion began immediately after the listening portion. The Subject was presented with an electronic survey containing the 80-word answer sheet corresponding to the trial. Words were displayed in alphabetical order with checkboxes with the prompt "Check the words you heard" on a computer monitor (see App. B, Figs. 29-36). Subjects were given 5 minutes to complete the answer sheet but were allowed to move on before the time was up.

3.4. Analysis

Statistical calculation included averages for words remembered in both trials, as well as the percentage of correct words from total words recalled. The distribution of which words were remembered most often was also tracked, as well as their distribution over time. Participants' level of musical training was correlated with words recalled, correct words recalled, and recall percentage. Finally, incorrect word recall averages, percentages, and frequency distribution was also calculated. The data was then analyzed for any patterns, including differences in performance between group A and B, in terms of both the number of words recalled and the number of correct words remembered, the frequency in which each word was recalled, the number of words and correct words recalled based on musical training; and the percentage of correct words recalled derived from the previous points. The data was then reviewed for any notable patterns or points of interest.

4. Findings

Our participant sample consisted of 11 Subjects separated into two groups; 6 in group A, and 5 in group B. Also important is that no test Subjects reported that they had any kind of hearing impairment. This would indicate that there may be no involvement of hearing impairment for the answers in the trials.

The data in the experiment shows, upon initial inspection, that almost all test Subjects recalled fewer words, and fewer correct words, in the multichannel test than in the stereo test. On average, people taking the stereo memory test recalled 25.78 words, 19.07 of which were correct. On the other hand, people taking the multichannel audio test recalled an average of 9.28 words, 6.4 of which were correct. Since both the multi-channel test and the stereo test used a total of 80 words, and all groupings of words were given 1 second intervals between sets, the stereo test lasted twice as long as the multichannel test; hence this can imply that the number of words remembered from the multichannel test may be higher over a longer time than the tested result.

4.1. Group Data

There was a notable difference between the performance of group A, and group B (see Fig. 13). On average, the participants in group A recalled 31.17 words in the stereo test, while recalling 10.17 words in the multichannel test, where participants in group B recalled an average of 20.40 words in the stereo test, and 8.40 words in the multichannel test. As for the number of correct words remembered, the participants in group A correctly recalled an average of 21.33 words correctly in the stereo test, and 7.00 words correctly in the multichannel test, where participants in group B correctly recalled an average of 16.8 words in the stereo test, and 5.80 words in the multichannel test (see Fig.

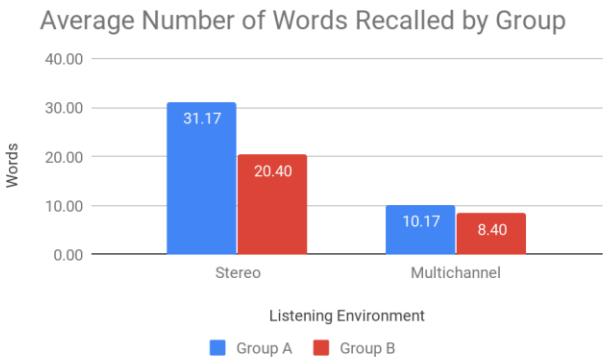


Figure 12: Average Number of Words Recalled by Group

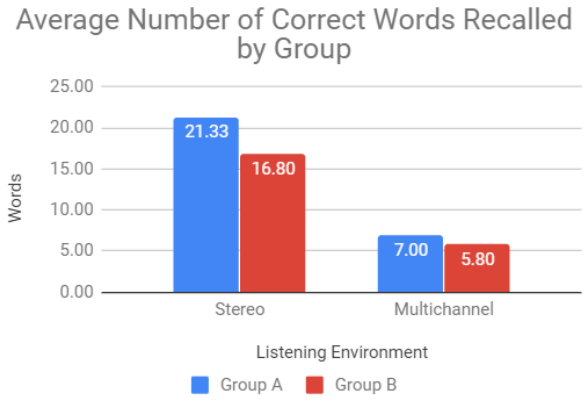


Figure 13: Average Number of Correct Words Recalled by Group

12). This results in group A recalling 68.45% correctly in the stereo test, and 68.85% in the multichannel test. Comparatively, group B recalled 82.35% of words correctly in the stereo test and 69.05% in the multichannel test. This represents the correct words guessed out of all of the words guessed, not the correct words guessed out of all options given. This means that there was a difference in the tests from group A to group B, but due to the low number of Subjects in each group, it must still be considered.

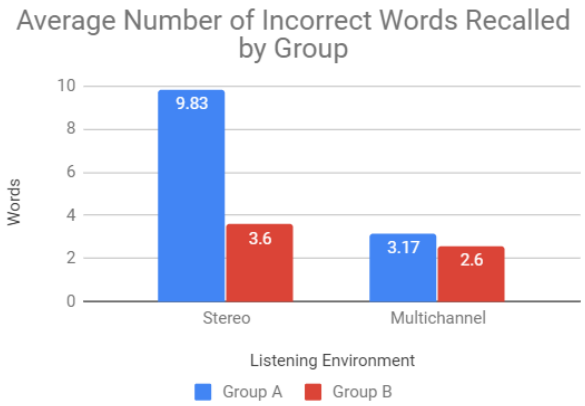


Figure 14: Average Number of Incorrect Words Recalled

Group A recalled an average of 9.83 words incorrectly in the stereo test, and 3.17 in the multichannel test. Group B recalled an average of 3.60 words incorrectly in the stereo test, and 2.60 in the multichannel test (see Fig. 15). This leads to group A recalling a total of 31.55% of words incorrectly in the stereo test, while recalling 31.15% of words incorrectly in the multichannel test. Group B recalled 17.65% of the words incorrectly in the stereo test, while recalling 30.95% of words incorrectly in the multichannel test.

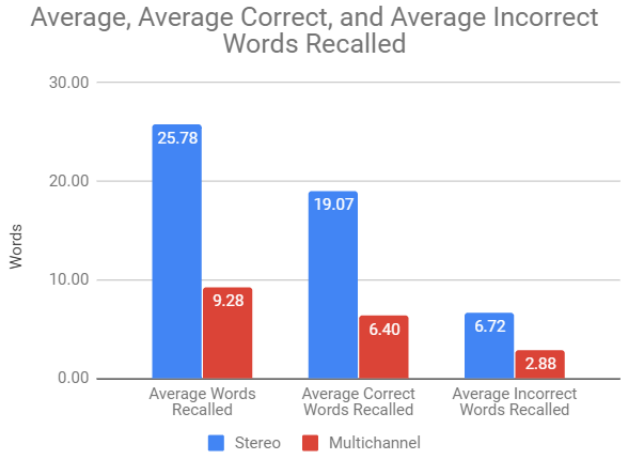


Figure 15: Average, Average Correct, and Average Incorrect Words Recalled

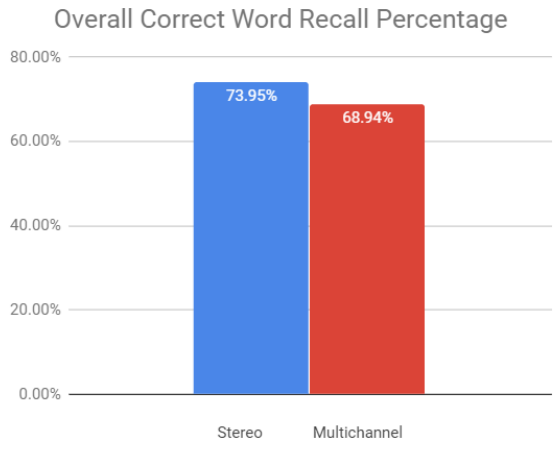


Figure 16: Overall Correct Word Recall Percentage

Considering all participants together, the group recalled an average of 25.78 words in the stereo test, and 9.28 in the multichannel test. The overall group *correctly* recalled an average of 19.07 words in the stereo test and 6.4 in the multichannel test. The percentage of correct words recalled out of the total words recalled was 73.95% in the stereo test and 68.94% in the

multichannel test. The number of words recalled incorrectly for the total group averaged 6.72 in the stereo test, and 2.88 in the multichannel test. The percentage of words recalled incorrectly compared to the total number of words recalled for the group was 26.05% in the stereo test and 31.06% in the multichannel test.

4.2. Musical Training Data

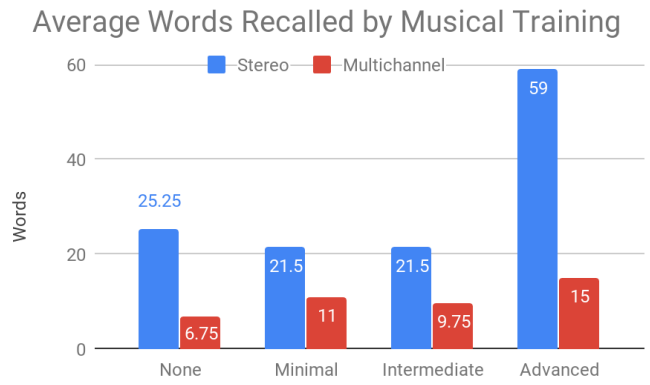


Figure 17: Average Words Recalled by Musical Training

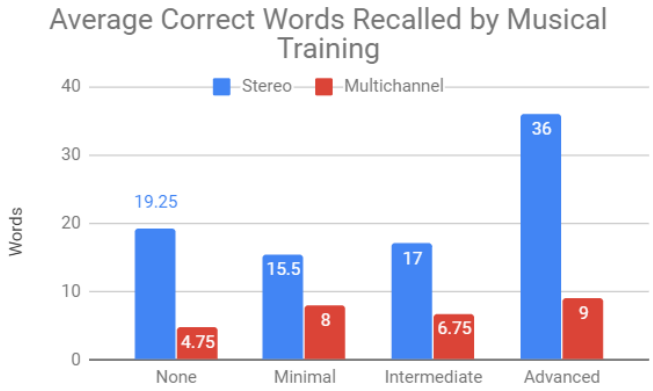


Figure 18: Average Correct Words Recalled by Musical Training

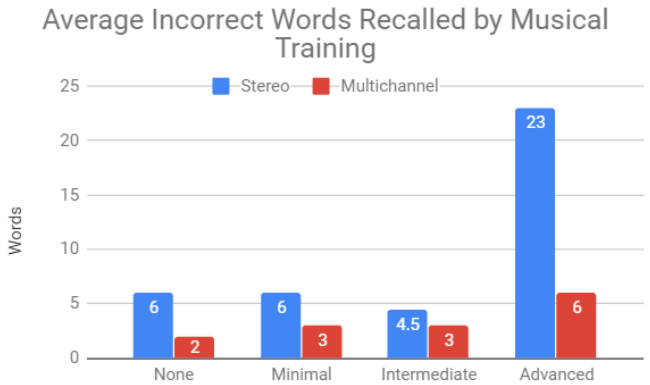


Figure 19: Average Incorrect Words Recalled by Musical Training

Before participating in the experiment, Subjects reported a self-identified level of musical skill. A total of four (4) Subjects described their level of musical training as "None", two (2) as "Minimal", four (4) as "Intermediate", and one (1) described their training as "Advanced".

Test Subjects who reported having "None" as musical training recalled an average of 25.25 words in the stereo test, and 6.75 in the multichannel test (see Fig. 17). They recalled an average of 19.25 words correctly in the stereo test, and 4.75 in the multichannel test (see Fig. 18). As a percentage of total words recalled, they recalled 76.24% of words correctly in the stereo test, and 70.37% in the multichannel test (see Fig. 19). Inversely, 23.76% of words were recalled incorrectly in the stereo test, and 29.63% in the multichannel test.

Test Subjects who reported their training as "Minimal" recalled an average of 21.5 words in the stereo test, and 11.00 words in the multichannel test (see Fig.

17). The Subjects recalled an average of 15.5 words correctly in the stereo test, and 8.0 in the multichannel test (see Fig. 18). As a percentage of total words recalled, they recalled 72.09% of words correctly in the stereo test and 72.73% in the multichannel test (see Fig. 19). Inversely, 27.91% of words were recalled incorrectly in the stereo test, and 27.27% in the multichannel test.

Test Subjects who reported their training as "Intermediate" recalled an average of 21.5 words in the stereo test, and 9.75 words in the multichannel test (see Fig. 17). The Subjects recalled an average of 17.0 words correctly in the stereo test and 6.75 in the multichannel test (see Fig. 18). As a percentage of total words recalled, they recalled 79.07% of words correctly in the stereo test and 69.23% in the multichannel test (see Fig. 19). Inversely, 20.93% of words were recalled incorrectly in the stereo test, and 30.77% in the multichannel test.

The test Subject who reported their training as "Advanced" recalled 59 words in the stereo test, and 15 words in the multichannel test (see Fig. 17). The Subject recalled 36 words correctly in the stereo experiment and 9 in the multichannel test. They recalled a total of 61.02% of words correctly in the stereo test and 60.00% in the multichannel test (see Fig. 19). Inversely, they recalled 38.98% of words incorrectly in the stereo test, and 40.00% in the multichannel test. This test Subject was in group A.

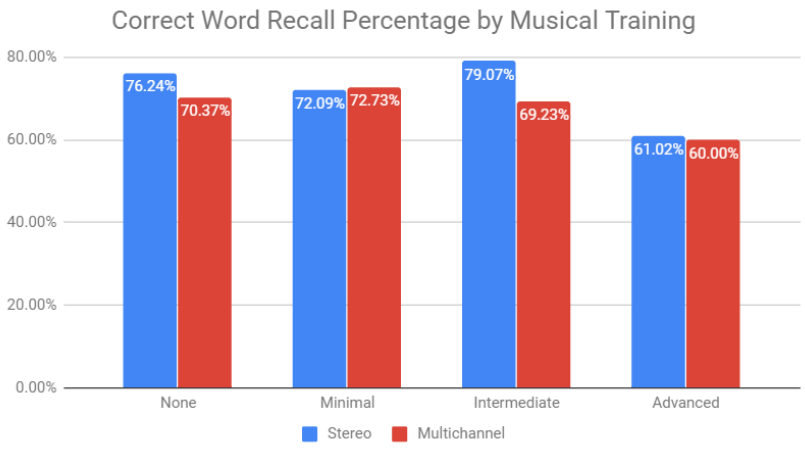


Figure 20: Correct Word Recall Percentage by Musical Training

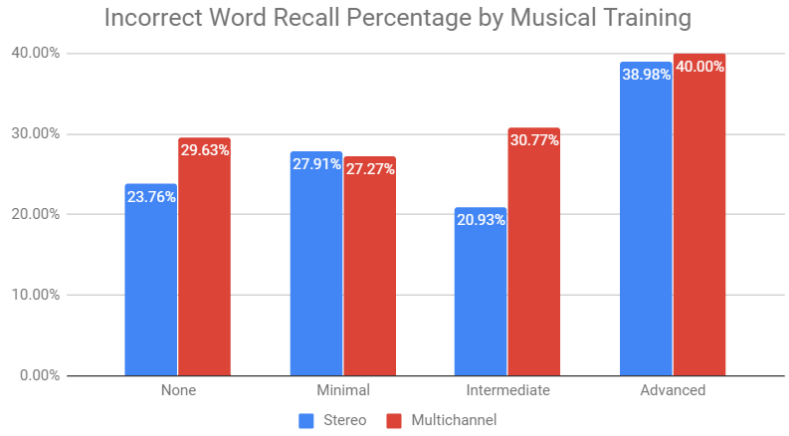


Figure 21: Incorrect Word Recall Percentage by Musical Training

4.3. Discussion

One of the notable patterns in the experiment was that group A, the group that performed the stereo experiment first, recalled more words overall compared to the test Subjects in group B, who performed the multichannel experiment first. However, the percentage of words they recalled correctly was somewhat smaller in the stereo experiment. Test group A correctly recalled an average of 21.33 words in the stereo experiment while group B correctly recalled an average of 16.8 words in the stereo experiment. Group A correctly recalled an average of 7.0 words in the multichannel experiment while group B recalled an average of 5.8 words correctly in the multichannel experiment. Comparing the number of words correct to the number of overall words recalled by the groups, however, shows that the test Subjects in group B were more accurate than those in group A for the stereo experiment. Group A recalled 68.45% of their words correctly while group B recalled 82.35% correctly in the stereo experiment. This change may be attributed to test group B having more practice in memorizing words and answering questions than group A by the time they reach the first experiment. The inverse, however, does not prove true, as group A had performed the multichannel test with an accuracy within .30% of group B. It is also quite possible that group A performed with similar percentage of words correct to group B on the multichannel experiment because of the increased auditory stress on the test Subjects. Test Subjects in group B may have been able to improve upon their performance because they had to focus on a smaller number of words at once in their next test, while group A had an increased auditory load in their next test. The increase of workload may have been unexpected or disorienting for group A, while the decrease of workload for group B could have been unexpected but encouraging.

The change in performance of the groups in terms of behavior when answering questions is more evident in the data showing the overall number of words recalled by the groups and their accuracy, especially in the stereo test. Test Subjects in group A recalled more words overall compared to those in test group B, with an average of 31.17 and 20.40 in the stereo test respectively, and an average of 10.17 and 8.40 in the multichannel test respectively. This difference in the average of number of words remembered may be attributed to the confidence and expectations of the test Subjects. One possible explanation for this is that test Subjects who were assigned to group A started with the stereo test, arguably the easier of the two tests—due to

having the least number of words playing at once—and, without any insight as to how difficult the next part of the test would be, were still expecting themselves to guess a number of words closer to their result in the first experiment. On the other hand, group B may have become confused at the difficulty of the multichannel experiment and consequently become less confident during the second trial. However, tying into their increased percentage of words correct compared to group A, it is also possible that the test Subjects had become more cautious when answering questions due to their wariness about the difficulty of the experiment.

Another notable pattern in the experiment for both group A and group B in the multichannel experiment and group A in the stereo experiment, test Subjects recalled words with an accuracy remarkably close to 69% (68.85%, 69.05%, and 68.45%, respectively) (see Figs. 16, 20). The only major difference in the percentage of words recalled correctly was in the stereo experiment for group B, who had an accuracy of 82.35%. This difference may be attributed to the aforementioned change in confidence and/or cautiousness of group B. Otherwise, this remarkable similarity in the percentage of words remembered may be something natural and inherent to the human brain and its psychology. It is very possible the test Subjects have the natural urge to answer their tests with as many words as possible until they are only sure that about 69% of their words are correct, unless otherwise influenced.

It is also possible that the difference in the words remembered between the experiments is due to the "cocktail party" effect. People have the ability to selectively tune in to certain sounds when presented with more than one audio source at once. In turn, our Subjects may have only heard one word being played at a time. If this was the only deciding factor in the number of words the test Subjects remembered, then the number of words correctly recalled in the stereo test would be approximately double the number remembered in the multichannel test. With the average number of words correctly remembered in the stereo test among all test Subjects standing at 19.07 and the average number of words correctly in the multichannel test standing at 6.40, the test Subjects remembered 2.98 times as many words in the stereo test as in the multichannel test (see Fig. 15). Because 2.98 is not approximately 2, it is likely that the cocktail party effect is not the sole factor in determining how many words the test Subjects remembered. However, this does not exclude the possibility that the cocktail party effect was a factor in the difference in the number of words recalled by the test Subjects. It is quite possible that the

cocktail party effect contributed to the difference in the number of words recalled, but its influence cannot be determined from the data gathered in the experiment.

It may seem likely that the cocktail party effect is the sole determinant for how many words the test Subjects remembered in the first test; since two words were playing at once and the Subjects would focus in on only one of them, 1 out of every set of 2 words would be recalled. With a total of 40 words, this would result in the test Subjects remembering 20 words, which is close to the results provided in the test (19.07 words remembered correctly). This matches the estimate with a 4.65% difference. For the multichannel test, Subjects would theoretically remember 1 of every set of 4 words recalled, which would mean a total of 10 out of the 40 words played back at them; however, they only remembered a total of 6.40 words. This is a 36% difference between the theoretical and test result, therefore the number of words remembered is not solely influenced by the cocktail party effect. The massive reduction in the performance of the test Subjects may have been due to the overwhelming or disorienting nature of the multichannel test, as the sounds possibly interfered with one another. It is also possible that the stress of having to distinguish any of the words out of the grouping impeded their ability to remember and distinguish the words.

The later 3 of the 4 spikes in words heard for the multichannel audio experiment (see App. C, Fig. 38) all come from the fourth channel. With the one outlier spike as coming from the second channel. It can indicate that our audio balance may have been off, or the audio file itself may have been recorded with a naturally higher gain level than the other audio files, resulting in the speakers having uneven output levels.

There are several low peaks in the stereo sample as well (see App. C, Fig. 37). They are mostly from the first channel and as a result it may be that we had lower-volume audio files on that channel as there are also major spikes distributed between both channels, showing that it is likely not the speakers were off balance.

5. Conclusions and Recommendations

5.1. Conclusions

The goal of our experiment was to determine the word recall efficacy of stereo audio compared to multichannel audio. This was to define and demonstrate the viability of using multichannel audio to increase learning capabilities, and so by studying and testing the efficacy of stereo audio compared to 4-directional multichannel audio in the short-term retention of information, we intended to discern the feasibility of bringing this technology into common use.

Within the experiment we conducted, we can conclude that in the simultaneous perception and retention of words, multichannel audio is not as effective as stereo. From our findings we can see, with the limited sample size we were able to obtain, that there is a stark difference between the two localization schemes. There is a considerable and significant decrease in the average correct words recalled, and an increase in average words incorrectly recalled by our Subjects between the two audio styles. Our Subjects recalled significantly more words, and more correct words, in our stereo experiment than in our multichannel experiment, where in our multichannel experiment, our Subjects were more likely to recall incorrect words than correct ones. However, our tests were intended to determine how we could use multichannel audio to improve methods of teaching and integration of auditory media as a means of utilizing the working memory. The flaws in our experimental procedure mean that our data suggests more about the cocktail party effect and means of disorienting a person than the psychological benefits or drawbacks of multichannel audio in a learning environment.

Our methods of controlling variables were effective in a number of ways but fell short in addressing the balance of our speakers, the timing of the spoken words such that they were comprehensible, and preventing the cocktail party effect from heavily impacting what the Subjects were able to focus on and remember. We were, to a great degree, able to manage the effects of reinforcement and categorization by randomizing our word pool from a list of the 5000 most frequent 2-syllable words, and vocal, tonal, and timbre variability was accounted for using Google's WaveNet

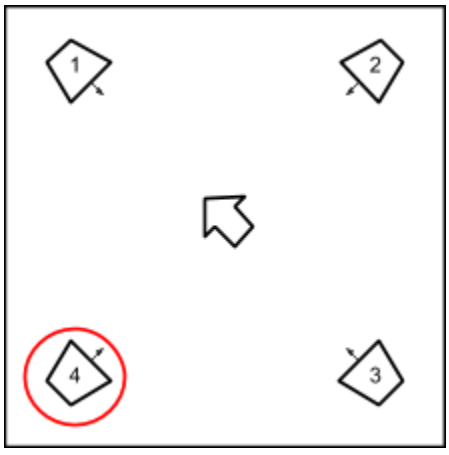


Figure 22: Unbalanced Speaker

Text-to-Speech software. What fell short in vocal control was the setup of the speakers and the individual balance of the words, where variance in the tonality of the text-to-speech program and the volume of the speakers made it easier to understand words from speaker 4 than any of the others.

The use of an answer sheet to select words instead of writing them out is another possible source of error. It could have caused a placebo-like effect, where the Subjects may have wrongly chosen words that could have been an incorrect assumption as to what they heard. The writing of the words would have removed this potential source of bias, but despite its benefits, as it made possible a greater degree of variability, a multiple-choice list was provided to simplify the gathered data.

Regarding musical ability, we were trying to see if there is any correlation between hearing and memory for the multiple channels, and musical training seemed to be a reasonable means of approximating a Subject's capability. With the sample size we were able to gather there was no discernible correlation between musical ability and the recall of correct words, but people claimed to hear more words when they identify themselves as having a more advanced musical training. However, as there was only one Subject in this category, this is certainly not conclusive.

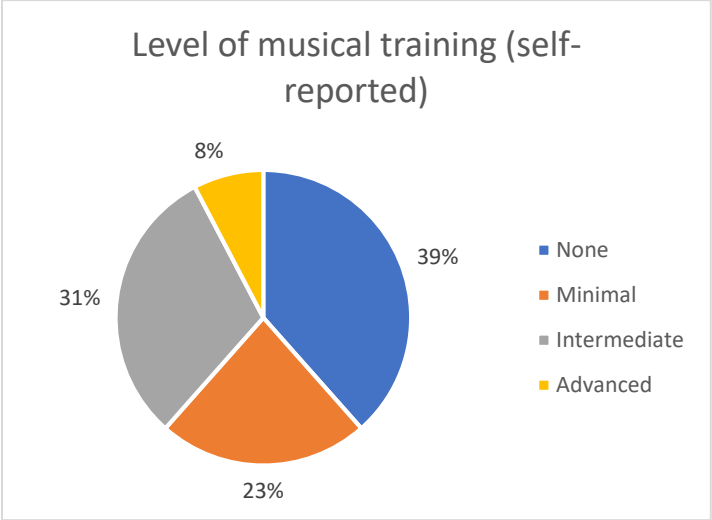


Figure 23: Level of Musical Training (self-reported)

While our tests were intended to better inform our use of multichannel audio to improve teaching methods and auditory media integration, they most likely demonstrate how the cocktail effect can interfere with auditory perception, even despite our lack of determinate data during the multichannel test. As this was not what we intended to test for, our data cannot support beyond doubt this likelihood, but while further testing would have clarified this point, we were unable to perform a second round of testing due to our lack of volunteers.

It is difficult to conclude that there is any significant change in auditory perception or short-term memory retention between stereo and 4-directional multichannel audio without clear

experimental data to base such supposition upon. Our tests clearly demonstrate a decrease in comprehensibility where the words played to the Subject are simultaneous, but where there is sufficient accommodation for the variables above and a more complete means of accumulating meaningful data, it is our belief that the results might be altered significantly.

5.2. Recommendations

Based on our understanding of the subject matter and testing throughout the experiment we have made some observations on the methods and concepts used to help develop and setup this project. Our recommendations are to further study this topic, and to make the following changes to our experimental procedure so as to ensure a more definitive result:

- Determine the limits of human hearing and memory so as to better utilize the maximum number of words reasonably perceivable at once.
- Define and operate under the optimal conditions for hearing multiple voices and perceiving them effectively, making sure the variables are adjusted for and controlled.
- Use a much larger sample size to gather more statistically accurate data and to collect more complete ranges of hearing abilities, better informing the separation of test participants into meaningful groups for data analysis.
- Look at correlations in data with other factors such as participant gender, sex, race, age, and education to group participants into more useful clusters.
- Continuing with this as an IQP or even MQP, planned and executed during the A-D terms and not the E terms, as there will be more possible Subjects during the school year.
- Test the possibility of offsetting audio tracks (in time) in a multichannel setup to benefit the listener's processing and recall ability.
- Develop a better system for the Subjects to recall and record the words they were able to recollect, as providing the list allowed for guessing to create higher scores and providing the participants with words they didn't hear potentially lowered their accuracy.
- Further explore the relationship between musical ability, and increased recall.

With the consideration of these points we can develop a better understanding of the limits of human hearing and memory to more completely account for what we failed to include or

provision for within this experiment. From this, our methods could possibly be adapted and used for better optimized work and learning environments, hopefully aiding any to employ them.

We recommend the further exploration of this topic to those groups most able to apply its benefits; developmental educators at every level of the public and private sectors, researchers inventing techniques to better use and understand the brain, civil servants working to stay up to date with the latest concepts and techniques, military training and communications, as well as any others able to make the concepts we have explored viable. Any technique that can fundamentally increase human ability to learn and grow should be developed to its furthest potential for the good of all.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of Learning and Motivation* (Vol. 2, pp. 89-195): Academic Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417-423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47-89): Academic Press.

Blauert, J. (1969). *Sound Localization in the Median Plane* (Vol. 22).

Bousfield, W. A., & Cohen, B. H. (1953). The effects of reinforcement on the occurrence of clustering in the recall of randomly arranged associates. *The Journal of Psychology*, *36*(1), 67-81. doi:10.1080/00223980.1953.9712878

Bousfield, W. A., & Cohen, B. H. (1955). The occurrence of clustering in the recall of randomly arranged words of different frequencies-of-usage. *The Journal of General Psychology, 52*(1), 83-95. doi:10.1080/00221309.1955.9918346

Bousfield, W. A., & Cohen, B. H. (1956). Clustering in recall as a function of the number of word-categories in stimulus-word lists. *The Journal of General Psychology, 54*(1), 95-106. doi:10.1080/00221309.1956.9920263

Brainard, M. S., Knudsen, E. I., & Esterly, S. D. (1992). Neural derivation of sound source location: Resolution of spatial ambiguities in binaural cues. *The Journal of the Acoustical Society of America, 91*(2), 1015-1027. doi:10.1121/1.402627

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

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), 975-979. doi:10.1121/1.1907229

Cloud text-to-speech. Retrieved from https://cloud.google.com/text-to-speech/

Corpus of contemporary american english. (December 2017). Retrieved from https://corpus.byu.edu/COCA/

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87-114. doi:10.1017/S0140525X01003922

Drullman, R., & W. Bronkhorst, A. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America, 107*(4), 2224-2235. doi:10.1121/1.428503

Goupell, M. J., & Stakhovskaya, O. A. (2018). Across-channel interaural-level-difference processing demonstrates frequency dependence. *The Journal of the Acoustical Society of America, 143*(2), 645-658. doi:10.1121/1.5021552

Impairments, N. R. C. U. C. o. D. D. f. I. w. H. (2004). *Hearing loss: determining eligibility for social security benefits* (R. A. Dobie & S. Van Hemel Eds. Vol. 2). Washington (DC): National Academies Press (US).

James, W. (1890). *The principles of psychology*. New York,: H. Holt and Company.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology, 41*(1), 35-39. doi:10.1037/h0061495

Jones, H. G., Brown, A. D., Koka, K., Thornton, J. L., & Tollin, D. J. (2015). Sound frequency-invariant neural coding of a frequency-dependent cue to sound source location. *Journal of Neurophysiology, 114*(1), 531-539. doi:10.1152/jn.00062.2015

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual review of psychology, 59*, 193-224. doi:10.1146/annurev.psych.59.103006.093615

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

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482-488. doi:10.1037/h0045106

Random noun generator. Retrieved from https://randomwordgenerator.com/noun.php

Squire, L. R., & Wixted, J. T. (2011). The Cognitive Neuroscience of Human Memory Since H.M. *Annual review of neuroscience, 34*, 259-288. doi:10.1146/annurev-neuro-061010-113720

Tarnow, E. (2010). There is no capacity limited buffer in the Murdock (1962) free recall data. *Cognitive Neurodynamics, 4*(4), 395-397. doi:10.1007/s11571-010-9108-y

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Winslow, F. (1860). *On obscure diseases of the brain and disorders of the mind*. Philadelphia: Blanchard & Lea.

Word frequency data. (December 2017). Retrieved from https://www.wordfrequency.info/free.asp

Yost, W. A. (2015). Psychoacoustics: a brief historical overview. *Acoustics Today, 11*(3), 46-53.

Appendix A: Word-list Audio

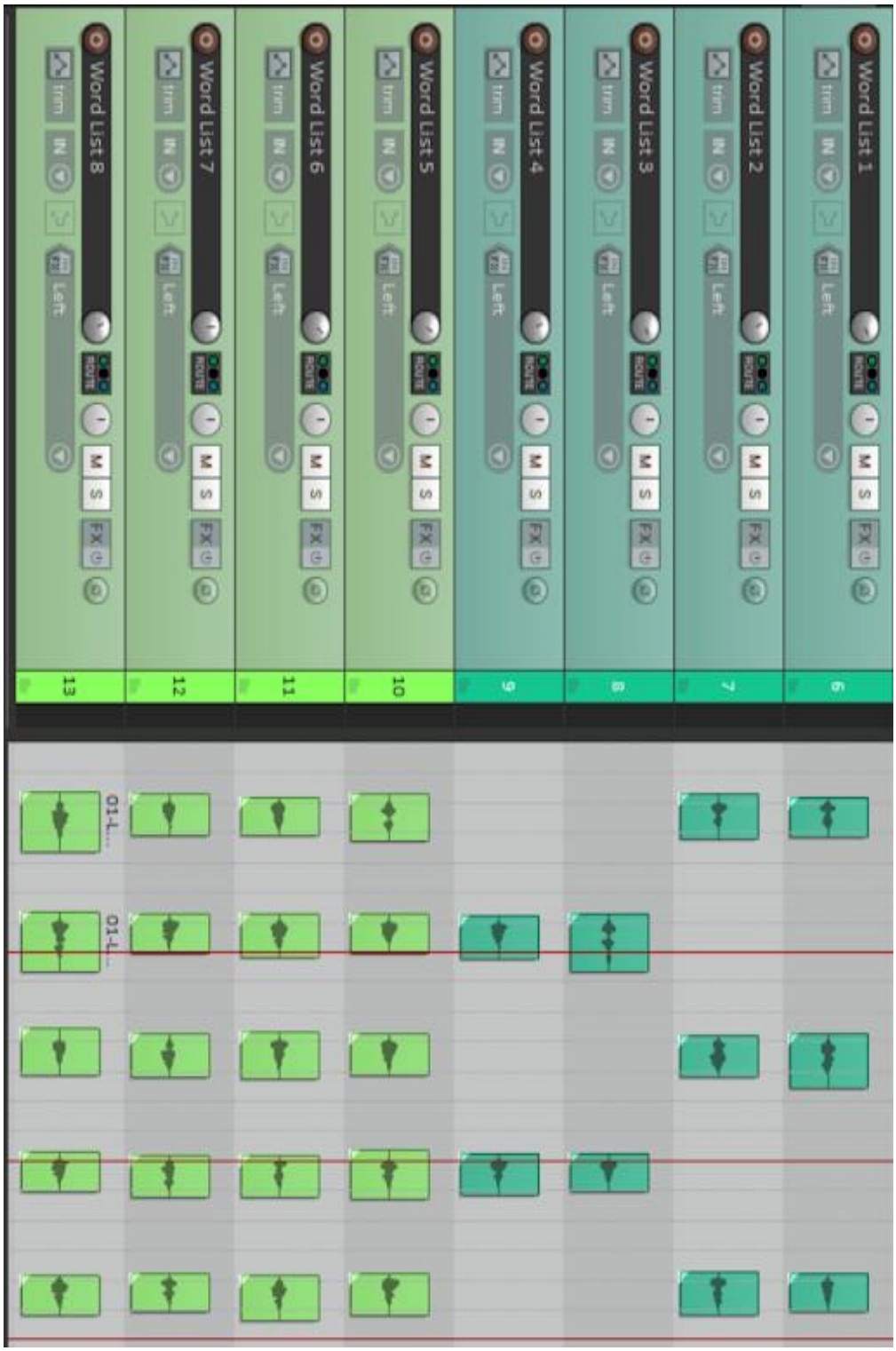


Figure 24: Screenshot of Word-List Audio (a)

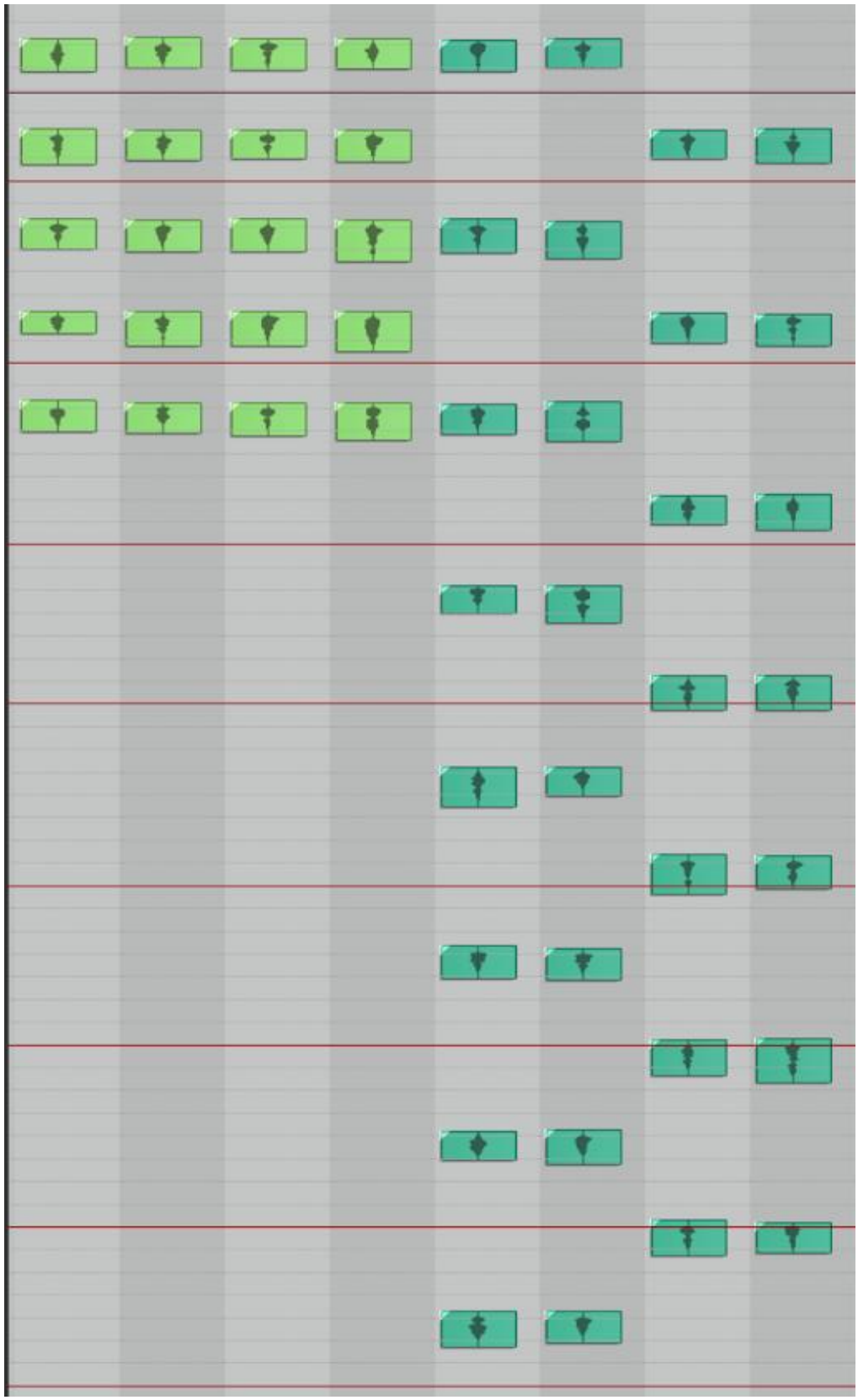
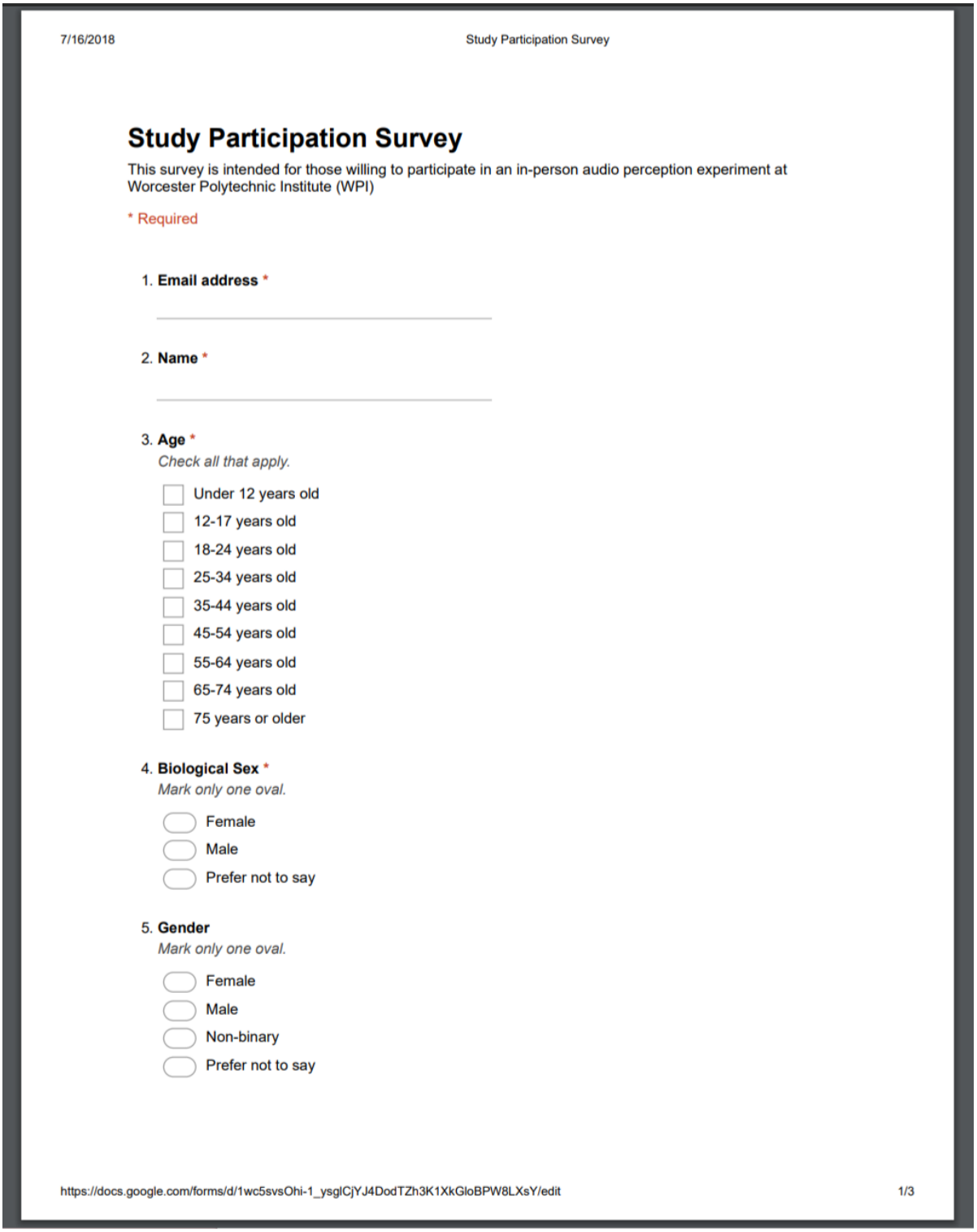


Figure 25: Screenshot of Word-List Audio (b)

Appendix B: Subject Answering Material



7/16/2018 Study Participation Survey

Study Participation Survey

This survey is intended for those willing to participate in an in-person audio perception experiment at Worcester Polytechnic Institute (WPI)

* Required

1. **Email address** *

2. **Name** *

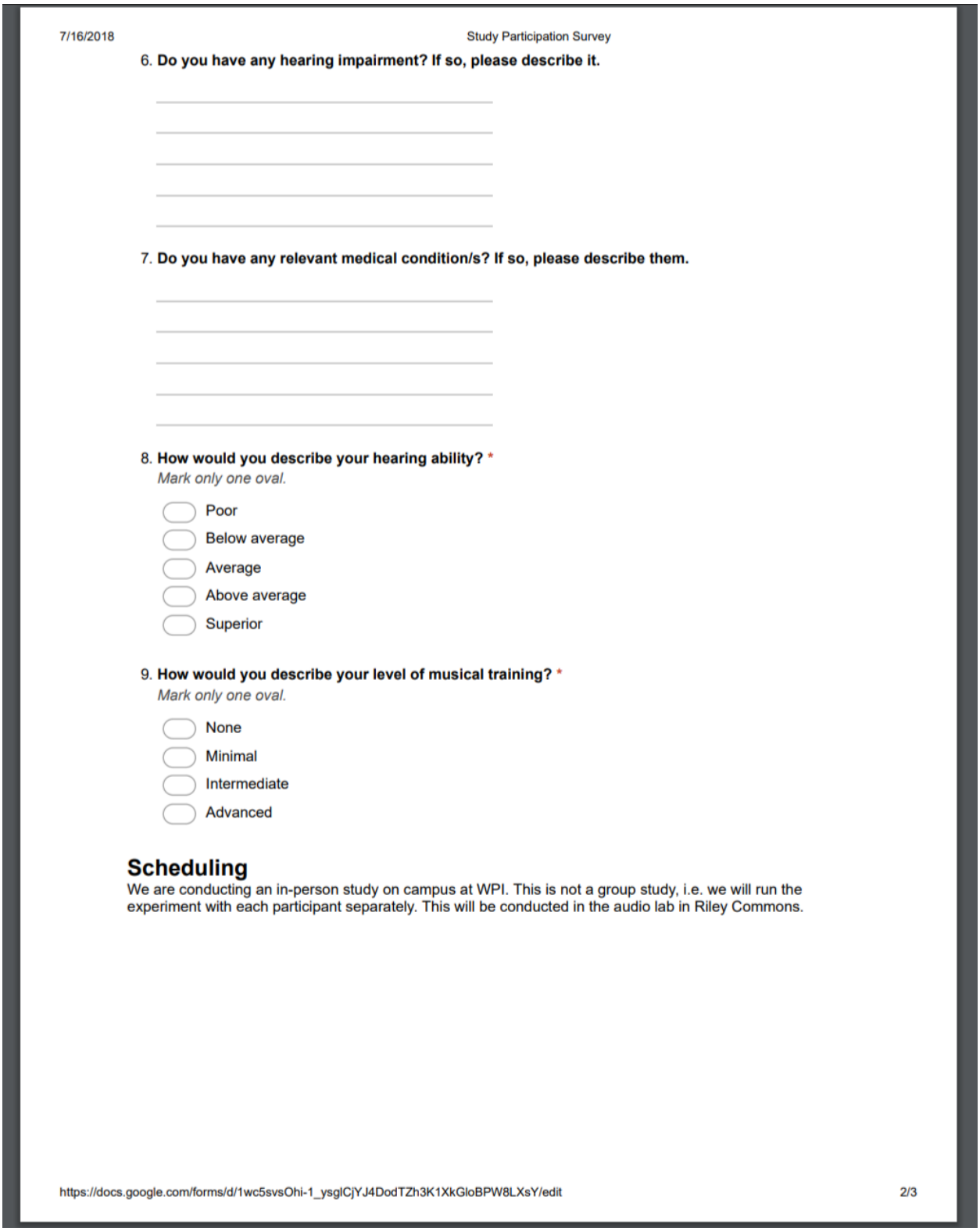
3. **Age** *
Check all that apply.
- [ ] Under 12 years old
- [ ] 12-17 years old
- [ ] 18-24 years old
- [ ] 25-34 years old
- [ ] 35-44 years old
- [ ] 45-54 years old
- [ ] 55-64 years old
- [ ] 65-74 years old
- [ ] 75 years or older

4. **Biological Sex** *
Mark only one oval.
- ◯ Female
- ◯ Male
- ◯ Prefer not to say

5. **Gender**
Mark only one oval.
- ◯ Female
- ◯ Male
- ◯ Non-binary
- ◯ Prefer not to say

https://docs.google.com/forms/d/1wc5svsOhi-1_ysglCjYJ4DodTZh3K1XkGloBPW8LXsY/edit 1/3

Figure 26: Pre-Experiment Survey (a)



7/16/2018 Study Participation Survey

6. **Do you have any hearing impairment? If so, please describe it.**

7. **Do you have any relevant medical condition/s? If so, please describe them.**

8. **How would you describe your hearing ability?** *
Mark only one oval.

- Poor
- Below average
- Average
- Above average
- Superior

9. **How would you describe your level of musical training?** *
Mark only one oval.

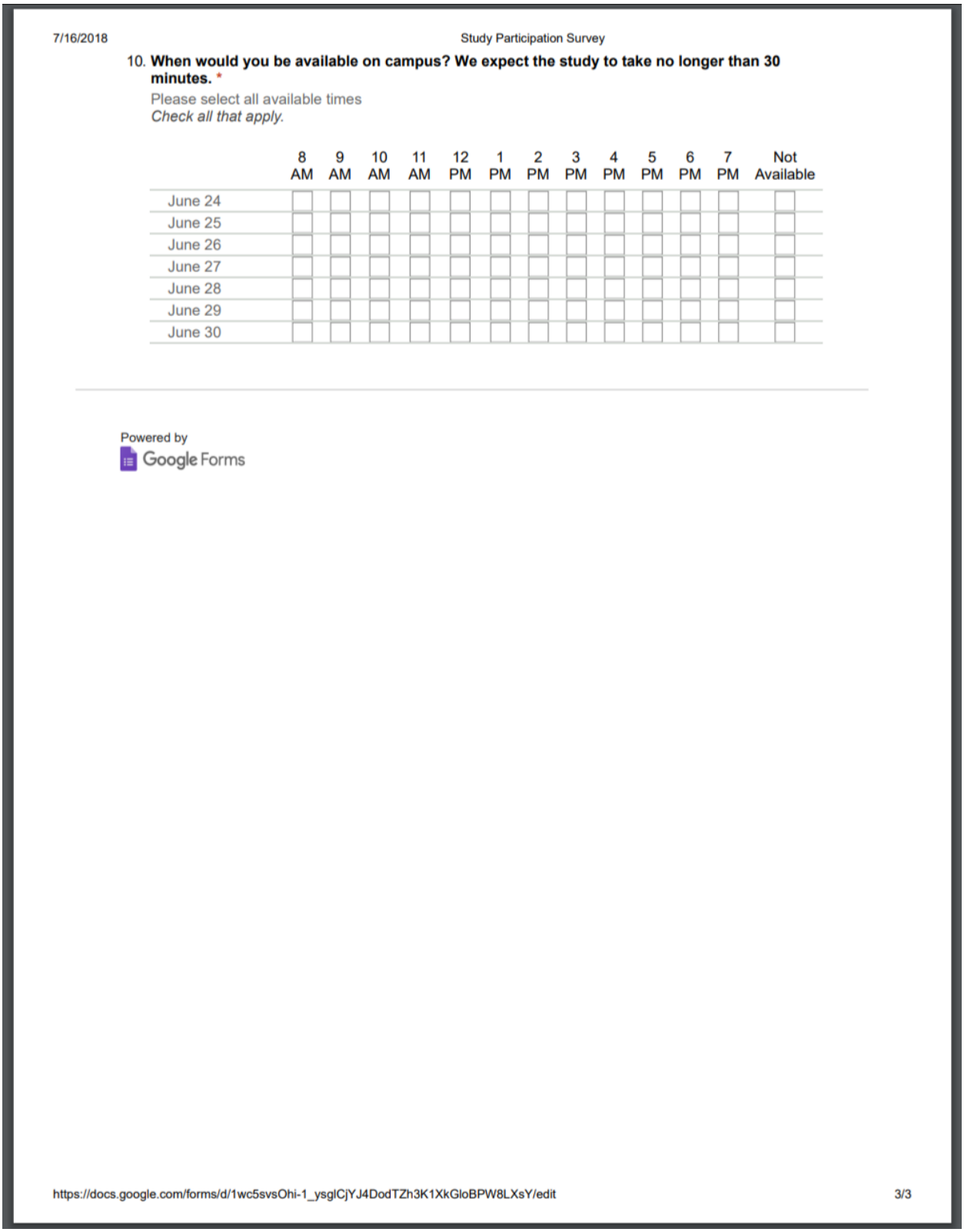
- None
- Minimal
- Intermediate
- Advanced

Scheduling

We are conducting an in-person study on campus at WPI. This is not a group study, i.e. we will run the experiment with each participant separately. This will be conducted in the audio lab in Riley Commons.

https://docs.google.com/forms/d/1wc5svsOhi-1_ysglCjYJ4DodTZh3K1XkGloBPW8LXsY/edit 2/3

Figure 27: Pre-Experiment Survey (b)



7/16/2018 Study Participation Survey

10. **When would you be available on campus? We expect the study to take no longer than 30 minutes.** *

Please select all available times
Check all that apply.

	8 AM	9 AM	10 AM	11 AM	12 PM	1 PM	2 PM	3 PM	4 PM	5 PM	6 PM	7 PM	Not Available
June 24	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
June 25	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
June 26	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
June 27	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
June 28	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
June 29	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
June 30	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Powered by Google Forms

https://docs.google.com/forms/d/1wc5svsOhi-1_ysglCjYJ4DodTZh3K1XkGloBPW8LXsY/edit 3/3

Figure 28: Pre-Experiment Survey (c)

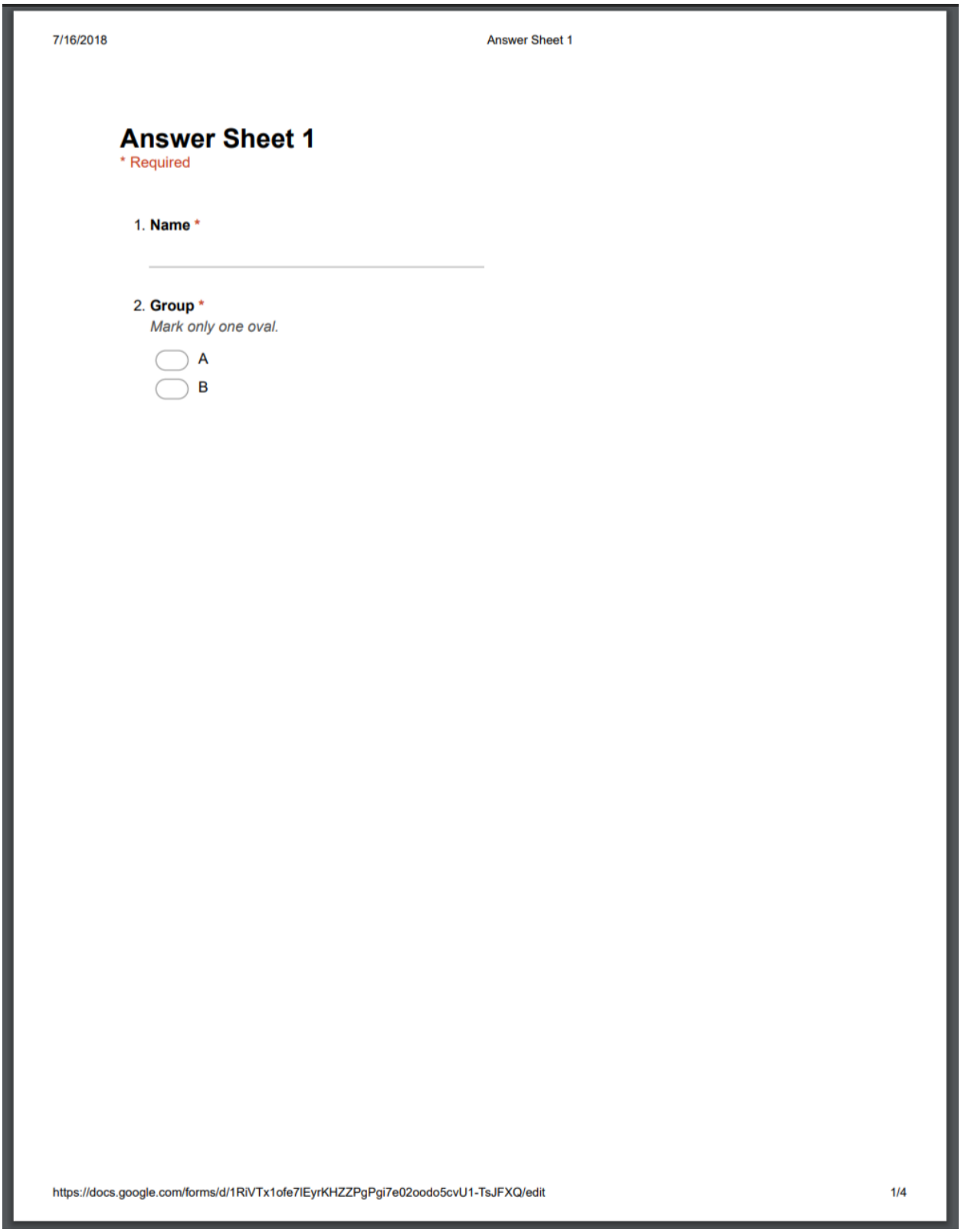
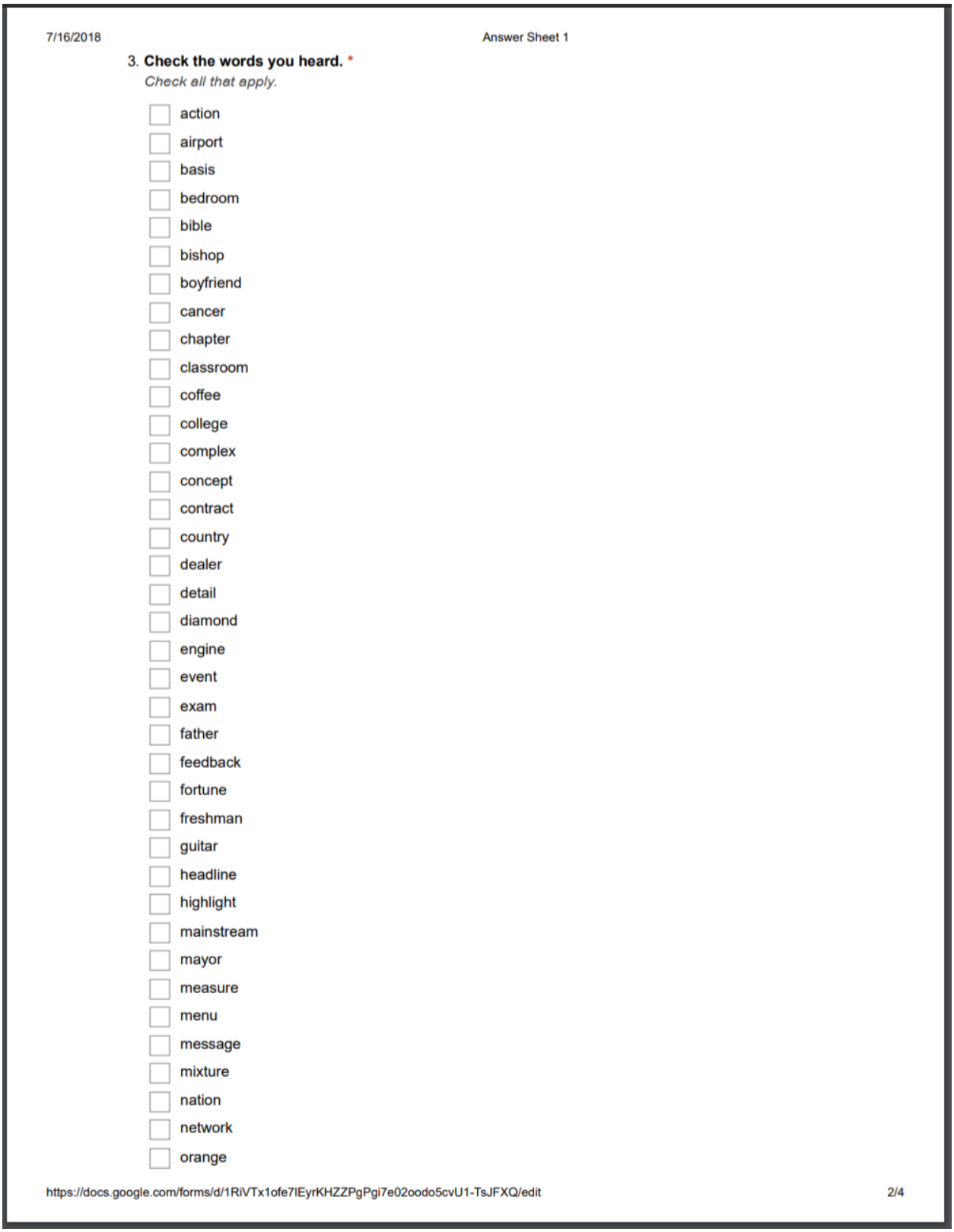


Figure 29: Answer Sheet 1 (a)



7/16/2018 Answer Sheet 1

3. **Check the words you heard.** *

Check all that apply.

- [ ] action
- [ ] airport
- [ ] basis
- [ ] bedroom
- [ ] bible
- [ ] bishop
- [ ] boyfriend
- [ ] cancer
- [ ] chapter
- [ ] classroom
- [ ] coffee
- [ ] college
- [ ] complex
- [ ] concept
- [ ] contract
- [ ] country
- [ ] dealer
- [ ] detail
- [ ] diamond
- [ ] engine
- [ ] event
- [ ] exam
- [ ] father
- [ ] feedback
- [ ] fortune
- [ ] freshman
- [ ] guitar
- [ ] headline
- [ ] highlight
- [ ] mainstream
- [ ] mayor
- [ ] measure
- [ ] menu
- [ ] message
- [ ] mixture
- [ ] nation
- [ ] network
- [ ] orange

https://docs.google.com/forms/d/1RiVTx1ofe7lEyrKHZZPgPgi7e02oodo5cvU1-TsJFXQ/edit 2/4

Figure 30: Answer Sheet 1 (b)

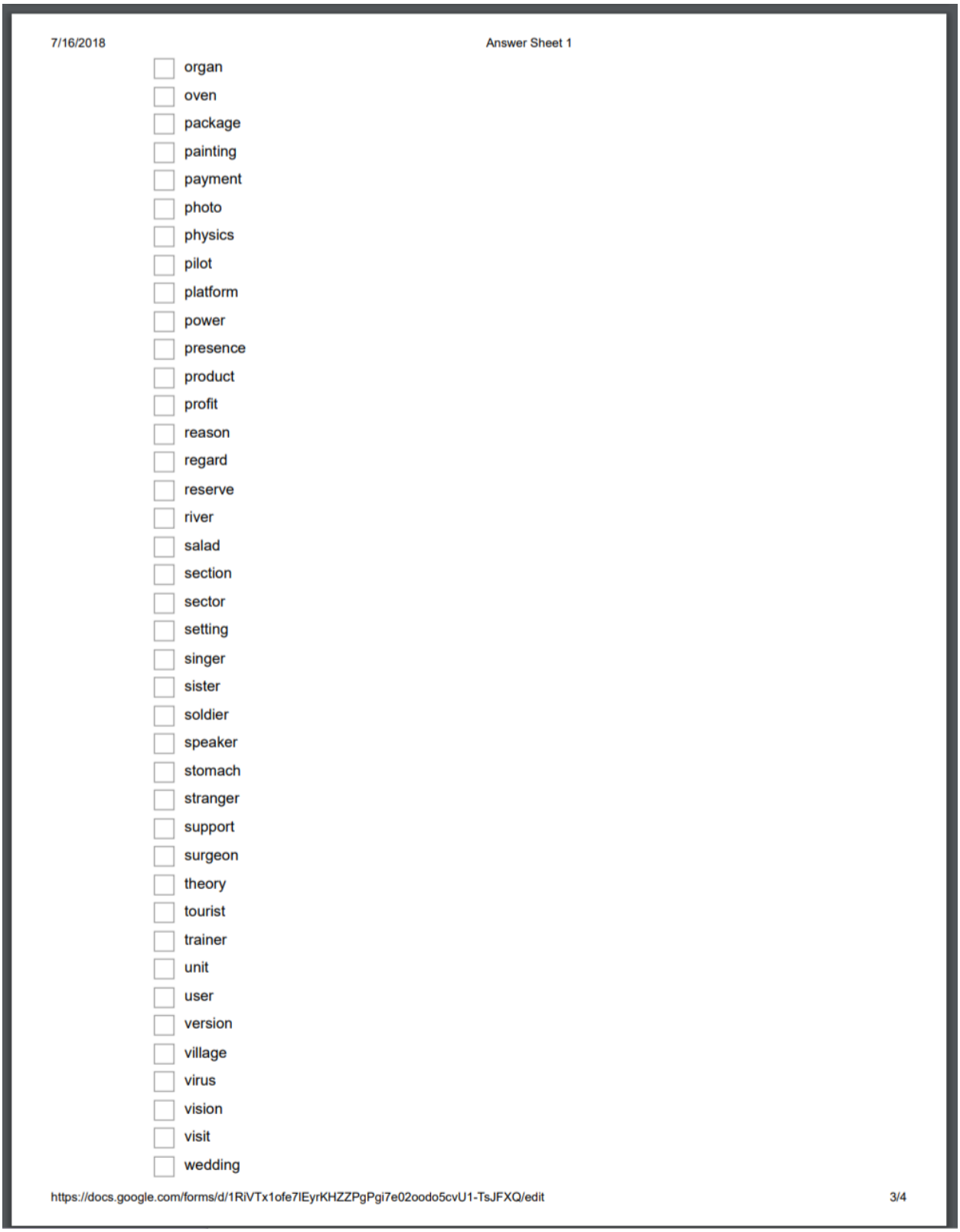


Figure 31: Answer Sheet 1 (c)

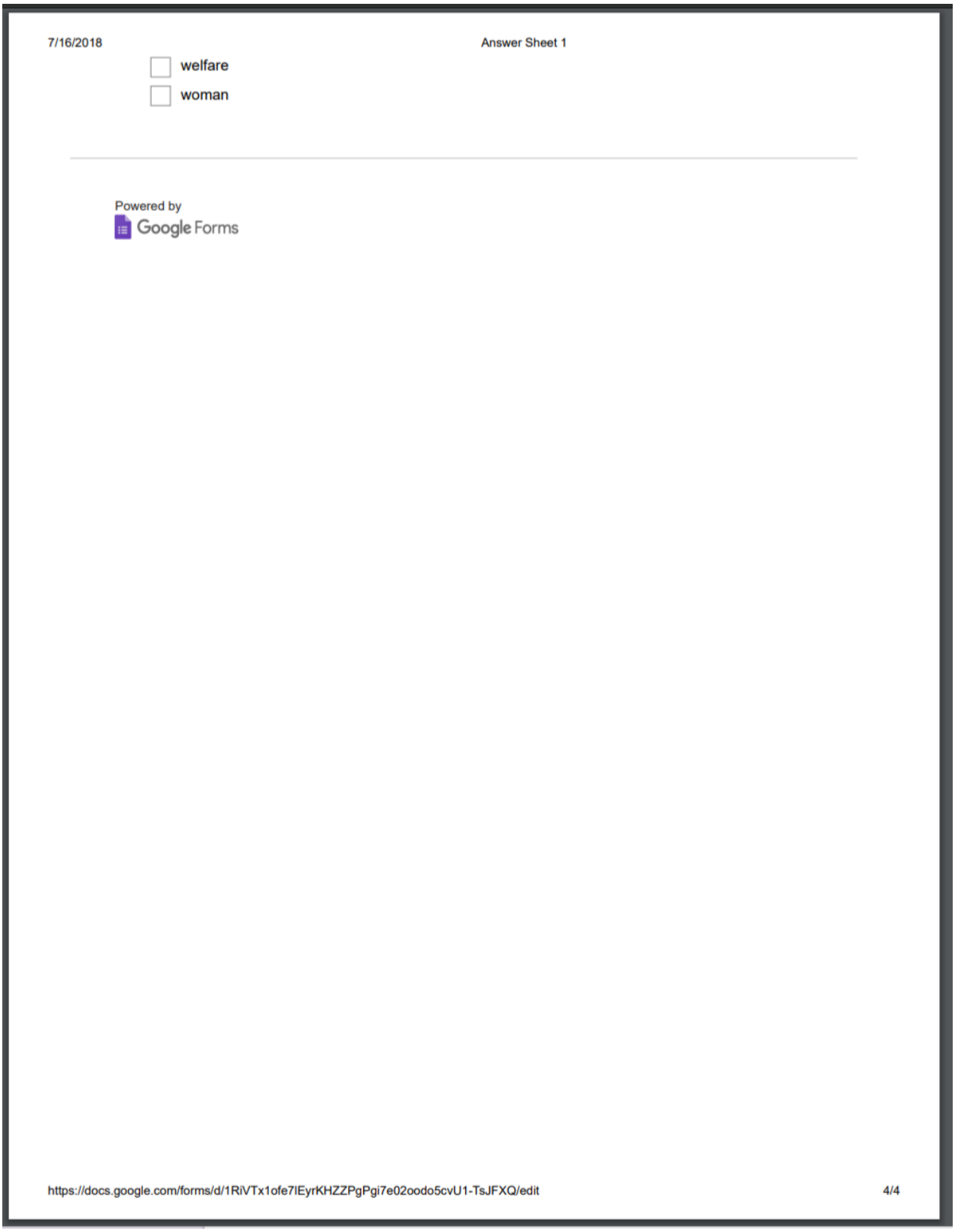


Figure 32: Answer Sheet 1 (d)

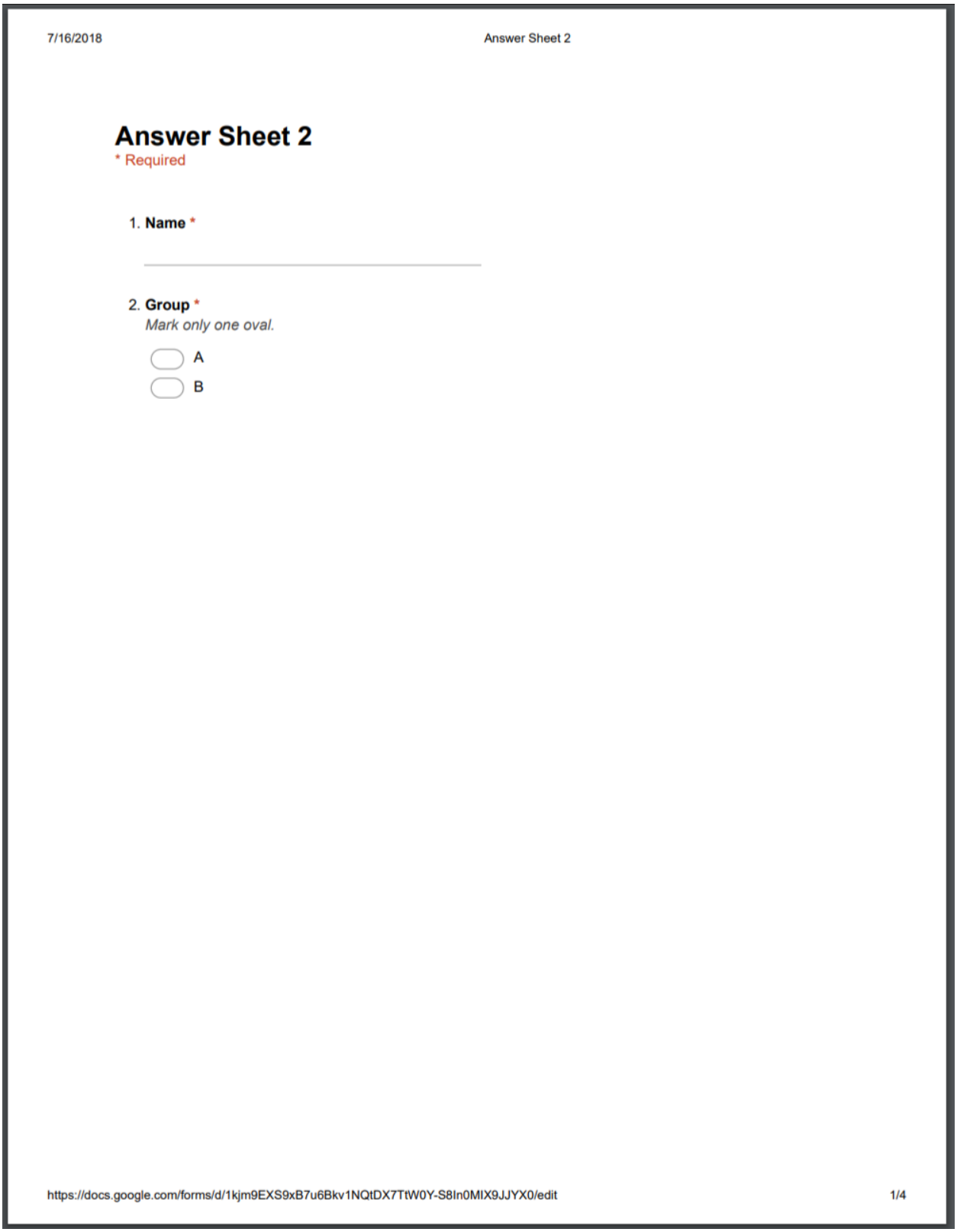


Figure 33: Answer Sheet 2 (a)

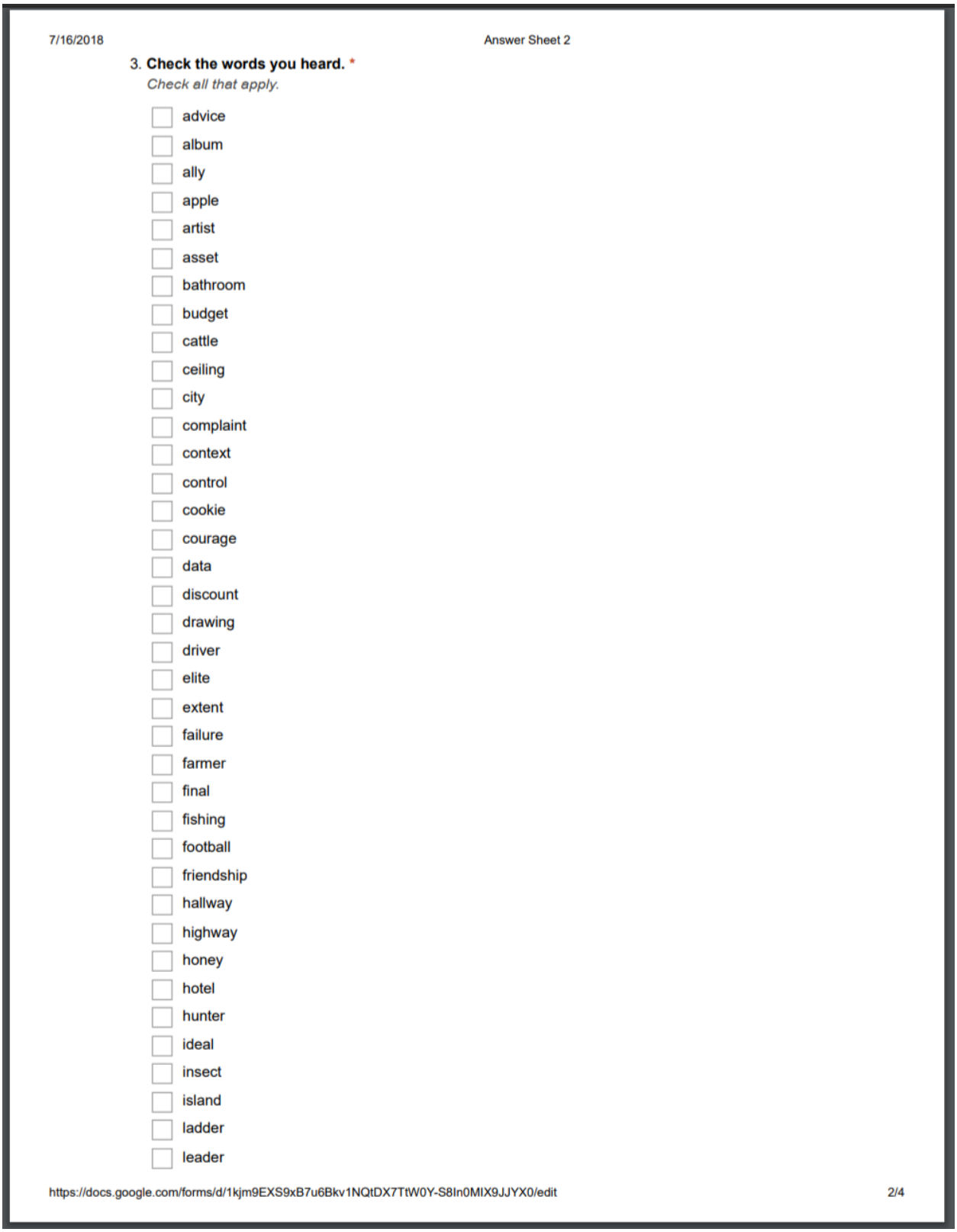


Figure 34: Answer Sheet 2 (b)

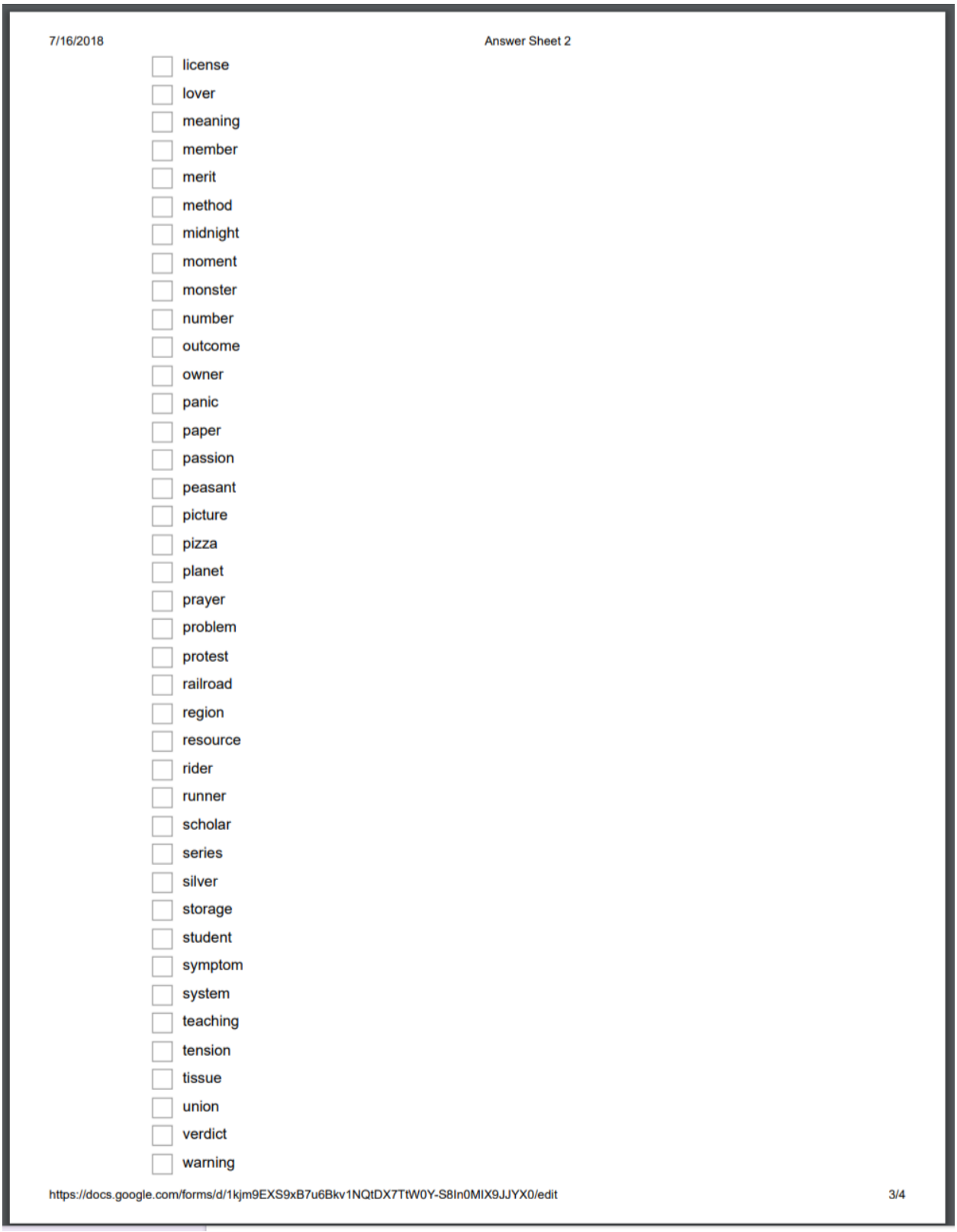


Figure 35: Answer Sheet 2 (c)

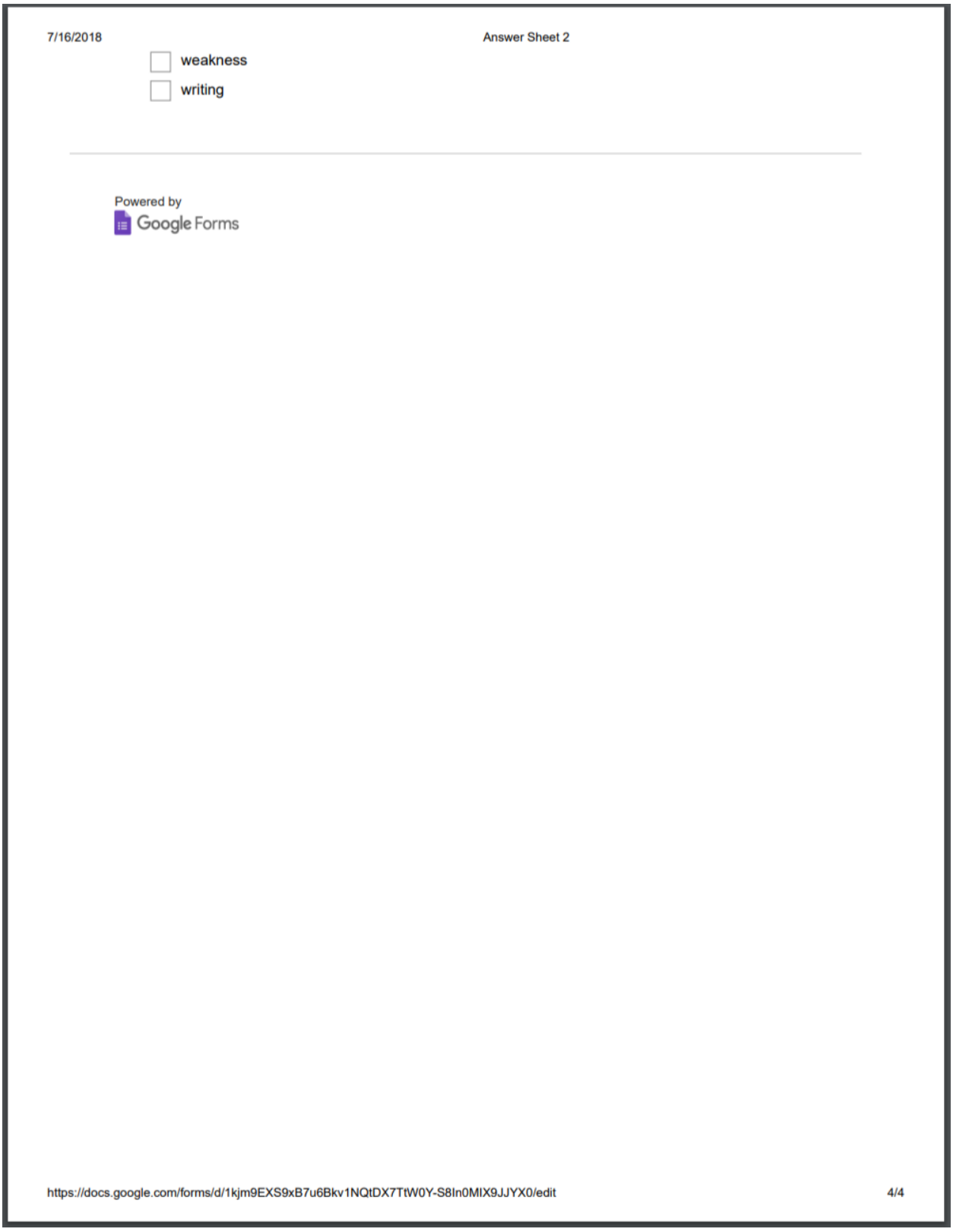


Figure 36: Answer Sheet 2 (d)

Appendix C: Recall Frequencies by Listening Order

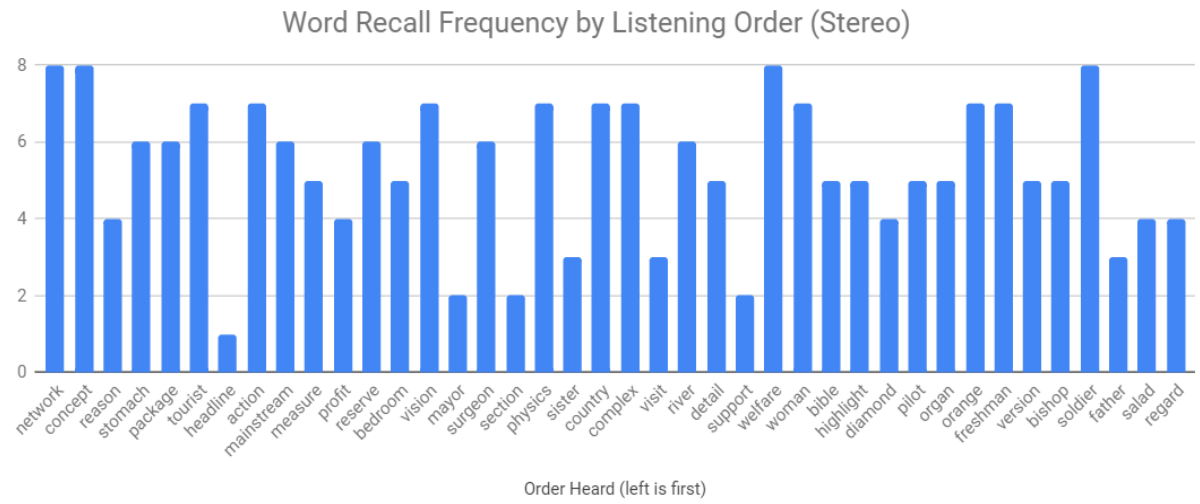


Figure 37: Word Recall Frequency by Listening Order (Stereo)

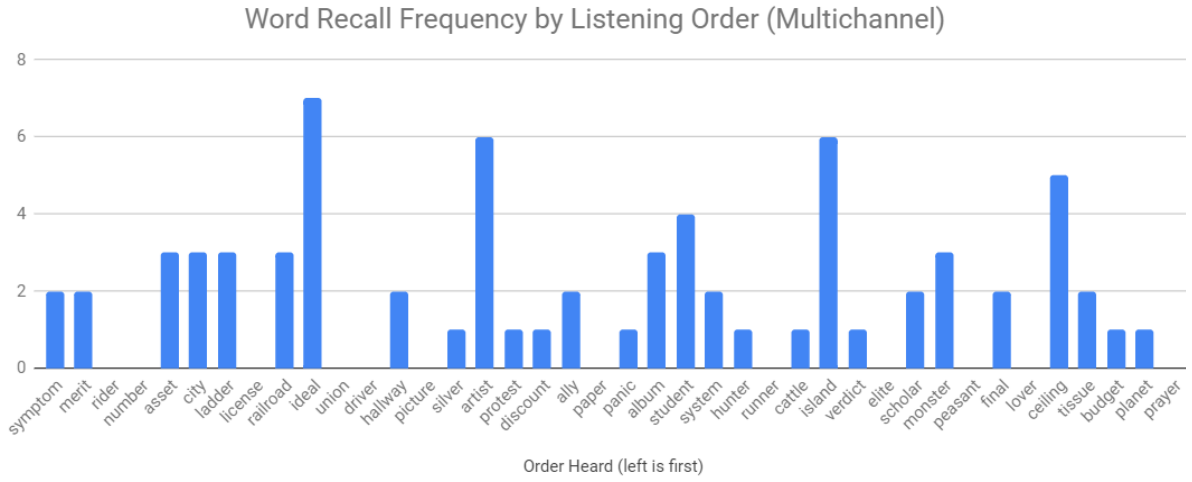


Figure 38: Word Recall Frequency by Listening Order (Multichannel)