Comparisons of physiologic and psychophysical measures of listening effort in normal-hearing adults

Nicholas Patrick Giuliani

University of Iowa

Copyright © 2017 Nicholas Patrick Giuliani

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/5937

Recommended Citation


Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Speech and Hearing Science Commons
COMPARISONS OF PHYSIOLOGIC AND PSYCHOPHYSICAL MEASURES OF LISTENING EFFORT IN NORMAL-HEARING ADULTS

by

Nicholas Patrick Giuliani

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Speech and Hearing Science in the Graduate College of The University of Iowa

December 2017

Thesis Supervisors: Professor Carolyn J. Brown Assistant Professor Yu-Hsiang Wu
CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Nicholas Patrick Giuliani

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Speech and Hearing Science at the December 2017 graduation.

Thesis Committee:  
Carolyn J. Brown, Thesis Supervisor

Yu-Hsiang Wu, Thesis Supervisor

Paul Abbas

Lenore Holte

Camille Dunn

Elizabeth Walker
To Kristin, Nora, Madeline, Lidia, and Leo.
ACKNOWLEDGEMENTS

I would like to thank the Department of Communication Sciences and Disorders for their financial support during this study. I am also grateful to Dr. Daniel Tranel for his expertise and assistance with the skin conductance measurements.

I would also like to thank my children for all the time they spent patiently sitting in “dad’s lab” watching shows and having chair races up and down the hall while I was working on my study.

More than anyone, I would like to thank my wife, Kristin. I do not know how I would have been able to do any of this without her.
ABSTRACT

The purpose of this study was to compare and contrast within and between participant performance on three different measures of listening effort: a dual-task paradigm, pupillometry, and skin conductance; participants also subjectively rated the difficulty of their experience. A repeated measures design was used to address the reliability and validity of each measure. 20 participants were recruited and attended two sessions; the second occurred a minimum of one week after the first. Participants listened to sentences presented in stationary noise at four different signal-to-noise ratios: quiet, 0, -3, and -5 dB SNR. The variables of interest were: change in peak-to-peak pupil diameter, change in reaction time from baseline, skin conductance response amplitude, and skin conductance response quantity.

The results indicated that as SNR decreased, speech perception performance decreased and subjective listening effort increased. Participants accurately and consistently rated the more difficult conditions as requiring more listening effort. The change in reaction time from baseline, peak-to-peak pupil diameter, and skin conductance response quantity increased as SNR decreased; skin conductance response amplitude did not vary as task difficulty increased, but skin conductance response amplitude was larger for incorrect responses than it was for correct responses. There was a significant practice effect observed for the reaction time data. The dual-task paradigm and pupillometry measures had the greatest reliability and validity. This study demonstrated that listening effort can successfully be quantified both subjectively and objectively by using a variety of tasks. Future studies may be able to use these measures to further assess listening effort in the clinic and in the real-world.
PUBLIC ABSTRACT

Listening in difficult, noisy situations is stressful and can impact a person in many ways: they may feel anxious or stressed, they may have a harder time understanding what is being said, and they may have a difficult time doing other things at the same time—like driving. “Listening effort” is the term hearing researchers use to describe the stress someone feels when it is challenging to understand what others are saying. However, we are not sure how to best measure listening effort. Most adults can easily tell us how they feel or how hard it was to understand in a noisy place, but children and adults with disabilities may not be able to describe how they are feeling when they are in a similar situation. We decided to look at how well people can multitask, as well as how their nervous system responds, when they listen to speech in situations with no background noise and situations where there is a lot of background noise. We thought that the participants would work harder to understand the sentences as we made the noise louder. We asked them how hard they worked and measured how well they multitasked, how large their pupils grew, and how much they sweat while they were listening to the sentences. Everyone in the study was able to tell us when it was hard to listen. We also learned that it was harder to multitask, their pupils grew larger, and they sweat more when they got more sentences wrong. Going forward, we hope to use these measures to better understand how children and adults with disabilities listen, and how much stress people feel when they are listening in real-world situations, like restaurants.
**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER I-INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER II-REVIEW OF THE LITERATURE</td>
<td>3</td>
</tr>
<tr>
<td>I. Overview of Listening Effort</td>
<td>3</td>
</tr>
<tr>
<td>II Rating Scales</td>
<td>4</td>
</tr>
<tr>
<td>III Physiologic measures of listening effort</td>
<td>4</td>
</tr>
<tr>
<td>IV. Dual-Task Paradigm</td>
<td>10</td>
</tr>
<tr>
<td>V. Conclusions</td>
<td>12</td>
</tr>
<tr>
<td>VI. Goals of the proposed study</td>
<td>13</td>
</tr>
<tr>
<td>CHAPTER III-METHODOLOGY</td>
<td>15</td>
</tr>
<tr>
<td>I. Participants</td>
<td>15</td>
</tr>
<tr>
<td>Sample size calculation</td>
<td>15</td>
</tr>
<tr>
<td>II. Stimuli</td>
<td>16</td>
</tr>
<tr>
<td>Auditory</td>
<td>16</td>
</tr>
<tr>
<td>Visual</td>
<td>16</td>
</tr>
<tr>
<td>Presentation and Recording</td>
<td>17</td>
</tr>
<tr>
<td>III Procedures</td>
<td>17</td>
</tr>
<tr>
<td>Pupillometry</td>
<td>17</td>
</tr>
<tr>
<td>Pupillometry Analysis</td>
<td>18</td>
</tr>
<tr>
<td>Dual-Task</td>
<td>19</td>
</tr>
<tr>
<td>Dual-Task Analysis</td>
<td>19</td>
</tr>
<tr>
<td>Skin conductance</td>
<td>20</td>
</tr>
<tr>
<td>Skin Conductance Analysis</td>
<td>21</td>
</tr>
<tr>
<td>CHAPTER IV-RESULTS</td>
<td>27</td>
</tr>
<tr>
<td>I. General procedures</td>
<td>27</td>
</tr>
<tr>
<td>II. Percent correct and Perceived Effort</td>
<td>28</td>
</tr>
<tr>
<td>III. Pupillometry</td>
<td>30</td>
</tr>
<tr>
<td>IV. Dual-task</td>
<td>33</td>
</tr>
<tr>
<td>V. Skin Conductance</td>
<td>34</td>
</tr>
<tr>
<td>VI. Individual Linear Trends</td>
<td>42</td>
</tr>
<tr>
<td>VII. Test-Retest Reliability</td>
<td>44</td>
</tr>
<tr>
<td>CHAPTER V-DISCUSSION</td>
<td>71</td>
</tr>
<tr>
<td>I. Pupillometry</td>
<td>74</td>
</tr>
<tr>
<td>II. Dual-task</td>
<td>78</td>
</tr>
<tr>
<td>III. Skin Conductance</td>
<td>80</td>
</tr>
<tr>
<td>IV. Intersession Reliability and Validity</td>
<td>85</td>
</tr>
<tr>
<td>V. Conclusions</td>
<td>88</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>89</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1 Laboratory setup........................................................................................................24
Figure 2 Pupillometry paradigm schematic................................................................................25
Figure 3 Skin conductance paradigm schematic........................................................................26
Figure 4 Speech perception ability grouped by testing paradigm..............................................46
Figure 5 Subjective listening effort grouped by testing paradigm.............................................47
Figure 6 Grand mean change in peak-to-peak pupil diameter..................................................48
Figure 7 Grand mean change in peak-to-peak pupil diameter with standard error .................49
Figure 8 Change in peak-to-peak pupil diameter from baseline.............................................50
Figure 9 Grand mean change in reaction time..........................................................................51
Figure 10 Change in reaction time for both sessions.................................................................52
Figure 11 Ambient changes in room temperature......................................................................53
Figure 12 Skin conductance response amplitudes for Participant 15.........................................54
Figure 13 Skin conductance response amplitudes for Participant 3...........................................55
Figure 14 Skin conductance response amplitudes for Participant 12.........................................56
Figure 15 Grand mean skin conductance response amplitude, Session I.................................57
Figure 16 Grand mean skin conductance response amplitude, Session II...............................58
Figure 17 Grand mean skin conductance response quantity.....................................................59
Figure 18 Grand mean skin conductance response amplitude based on speech perception
  performance..........................................................................................................................60
Figure 19 Grand mean skin conductance response quantity based on speech perception
  performance..........................................................................................................................61
Figure 20 Correlation of SCR quantity and amplitude.............................................................62
Figure 21 Normalized linear relationship of perceived listening effort during pupillometry ................................................................. 63

Figure 22 Normalized linear relationship of perceived listening effort during dual-task testing. ........................................................................... 64

Figure 23 Normalized linear relationship of perceived listening effort during skin conductance testing. ........................................................................... 65

Figure 24 Normalized linear relationship of the change in peak-to-peak pupil diameter ........................................................................... 66

Figure 25 Normalized linear relationship of the change in reaction time from baseline. ........................................................................... 67

Figure 26 Normalized linear relationship of the change in SCR amplitude ........................................................................... 68

Figure 27 Normalized linear relationship of the change in SCR quantity ........................................................................... 69

Figure 28 Test-retest reliability ........................................................................... 70
LIST OF ABBREVIATIONS

Electrodermal Activity (EDA)
Listening Effort (LE)
Normal Hearing (NH)
Non-specific Skin Conductance Response (NS.SCR)
Skin Conductance Response (SCR)
Skin Conductance Level (SCL)
Signal-to-Noise Ratio (SNR)
CHAPTER I
INTRODUCTION

Speech perception tests quantify the benefit attributed to hearing aids and cochlear implants. Some are simple tests with one talker (IEEE, 1952) whereas others incorporate a variety of talkers (Spahr et al., 2012). Others are made more challenging by introducing background noise (Nilsson et al., 1994; Spahr et al., 2012); however, modern hearing aid and cochlear implant users describe additional benefits that are not reflected in a variety of speech perception tests. For example, digital noise reduction does not improve speech perception scores (Magnusson et al., 2013; Bentler et al., 2008; Chung et al., 2006; Bentler, 2005). Nevertheless, many participants report that the addition of digital noise reduction provides a “better sound quality” (Chung et al., 2006) and makes them “less averse to sounds” (Bentler et al., 2008). Speech perception testing alone does not assess the range of benefits experienced by modern hearing aid and cochlear implant users. Clearly, having an objective metric that complements speech perception testing would be welcomed.

In recent years, listening effort has gained attention as a metric that may supplement or, in some cases, replace speech perception testing in the audiology clinic. To date, there is no widely accepted measure of listening effort. Rating scales are simple to administer and are inexpensive; they are also variable, subjective, and difficult to compare across participants. Dual-task paradigms add an additional layer of complexity that may not allow for an uncontaminated assessment of listening effort. Pupillometry provides useful insight into the cognitive demands required for a challenging listening task, but the technology is expensive, not extremely robust to movement, and requires careful control of the testing environment. Skin conductance, yet another proposed
measure thought to provide insight into listening effort, is less expensive than pupillometry but current research on how it relates to listening effort is limited. The sensitivity of these four metrics varies and to date, no study has reported within and between participant comparisons of these measures. The following sections will briefly describe the history of each measure, how each may be used to assess listening effort, and their strengths and weaknesses.
CHAPTER II
REVIEW OF THE LITERATURE

I Overview Of Listening Effort

The concept of listening effort in the literature is encountered only sparingly prior to the early 2000s. Most of the early studies seeking to assess listening effort did so by either a dual-task paradigm (Downs and Crum, 1978; Downs, 1982) or a rating scale (Preminger and Van Tassel, 1995; Bentler et al., 2005). Many of these studies suggested that listening effort was greater in challenging listening environments (e.g., background noise) and was greater for individuals with hearing loss (Downs, 1982). With the exception of this small collection of early studies, listening effort was explored only as an anecdote until recently. In fact, the first formal definition for listening effort was not established until 2014 when a British Working Group defined it as: “The mental exertion required to attend to, and understand, an auditory message” (McGarrigle et al., 2014). This definition held until 2016, when a consortium of listening effort researchers suggested a broader definition. They first defined mental effort, the overarching mental process that encompasses listening effort, as: “The deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task” (Pichora-Fuller, et al., 2016); listening effort was then specified as the direction of mental effort towards a specific listening task. Whatever the exact definition may be, the main assumption behind listening effort is that as a listening task becomes more challenging, the listener must work harder (use more cognitive resources), and thus exert more effort to accurately attend to the target. This allocation of cognitive resources may be reflected in self-reported measures, psychophysical tasks, and as physiologic responses; however,
the most appropriate method for listening effort has yet to be established. The following section will explore the historical context of the many techniques used to assess listening effort.

II Rating Scales

Rating scales are a straightforward and cost-effective ways to assess listening effort, but lack face validity (Ohlenforst et al., 2017). One potential explanation for this is that no currently available rating scale was developed specifically to assess listening effort. Researchers have asked listeners to rank their experience using a visual-analog or Likert scale (Zekveld et al., 2011, Mackersie et al., 2009); other researchers (Bentler et al., 2005, Wu et al., 2016) have adapted existing scales, such as the NASA Task Load Index (Hart and Staveland, 1988) or APHAB (Cox and Alexander, 1995), to assess subjective listening effort. We know that individual participants exhibit poor reliability and trial-by-trial coherence, meaning that the same participant may rate their experience differently from trial to trial, even if the task is identical (Marks and Florentine, 2011). As Borg (1982) notes: “good general functions for a group of participants can be obtained, but it is difficult to compare the participants with each other because they are asked to make only relative comparisons.” Ohlenforst et al. (2017) conducted a systematic review of recent listening effort studies and found exactly that: current subjective rating scales are not able to illuminate differences in listening effort across participants.

III Physiologic Measures Of Listening Effort

The autonomic nervous system

The autonomic nervous system controls bodily functions that are not under direct, conscious control of an individual. It can be further divided into two parts: the sympathetic and parasympathetic. Many people learned in a high school or college
human physiology course that the sympathetic branch governs the "fight-or-flight" response, while the parasympathetic branch governs the "rest and digest" response. There are many exogenous factors that can evoke a change in sympathetic nervous system activity. This change in state is commonly referred to as a change in arousal level.

Merriam-Webster defines *arouse* as "to stimulate to action or to bodily readiness for activity." Stimuli that have strong affective value, in both the negative and positive sense, can elicit a change in arousal state. The effects of strong stimuli in the visual (Lang et al., 1998; Tranel, 2004), olfactory (Bensafi et al., 2002), and auditory (Bradley and Lang, 1999) domains are well-studied and can be indexed by non-conscious changes throughout the body. Pupil dilation and skin conductance are two metrics that are known to reflect changes in the sympathetic nervous system and may be useful tools for assessing physiologic changes during effortful listening.

**Pupillometry**

The pupil is an opening in the iris that changes size in response to a wide variety of stimuli; pupillometry is simply the process for measuring such changes. The light reflex, which causes the pupil to dilate or constrict several millimeters, is commonly used by physicians to assess a patient’s neurological state. A physician will shine a light in a patient’s eye, making sure that the pupils dilate symmetrically and simultaneously. The pupil will also dilate, although only tenths of a millimeter, while a person is engaged in a cognitively demanding task (Kahneman and Beatty, 1966; Beatty, 1982; Laeng et al., 2012). Several investigators have exploited these findings to demonstrate how pupillometry reflects listening effort (Zekveld et al., 2010, 2011, 2014; Zekveld and Kramer, 2014; Steel et al., 2015; Winn et al., 2015).
Early pupillometry research was time-consuming and labor intensive. Researchers would take hundreds of pictures and painstakingly measure changes in pupil diameter by hand. One early such study was conducted by Kahneman and Beatty. In 1966, they showed that the pupil dilates in response to increasing cognitive demands and constricts as the participant reports the results. This was further expanded upon in 1982 by Beatty who reviewed several studies and classified changes in pupil diameter by various task types. Winn et al. (2015) further expanded on the results from Beatty by adding the unique challenges of listening to vocoded speech, which simulates how speech might sound to a CI user. The researchers distorted speech passages by processing them using a vocoder by: 1) varying the number of channels (fewer channels represent less-clear speech simulations) and 2) changing the steepness of the filter cutoffs (broader filters equate to more distorted speech). They reported that pupil diameter increased as the passages became more distorted, and thus requiring more listening effort to comprehend. Perhaps the most important finding was that pupil dilation continued to increase, even in conditions where there was no significant decrement in speech perception abilities; the only exception was the most highly degraded condition where participants did perform poorly on speech discrimination. If these findings hold across studies, they suggest pupillometry might be a useful supplement to clinical assessments of speech perception where a simple percentage correct may not encapsulate everything that a listener is experiencing. A change in listening effort may explain why these physiologic responses vary in the absence of changing speech perception abilities.

While pupillometry has considerable clinical potential, there are also several drawbacks. For one, the equipment required to measure pupil dilation
is expensive, making it inaccessible to many clinics. Proper pupil diameter measurements also require strict control of ambient light levels, which means that testing environments will likely require modification. Furthermore, obtaining accurate measures of pupil diameter requires that the listener wear goggles or hold their head relatively still while keeping their eyes fixated on a computer screen. If a participant is unable to do so, the eye tracker will lose its “lock” on the pupils, resulting in a loss of data. Additionally, there are wide variations in iris color from person-to-person. Common iris colors range from light blue to dark brown or black and include rare violet and red variants (White and Rabago-Smith, 2011)–some color variants may be more difficult to assess than others. While eye-tracking units state that they are robust to such factors, it is not yet clear how much and to what degree elements such as contact lenses, eyeglasses, eye makeup, and/or head movements have on the equipment. Finally, hardware and software manufacturers will need to determine how to separate dilation due to the light reflex from smaller changes due to increasing cognitive demands. Many studies are looking at the real-world implications of listening effort (Wu et al., 2014) and if these potential weaknesses of pupillometry can be controlled, the additional information it may provide is appealing.

**Skin conductance**

Skin conductance quantifies the electrical admittance of the skin—specifically the sweat glands—and may serve as another objective measure of listening effort. When the sweat glands become active, they excrete a conductive solution, which in turn alters the resistance of the skin. A small, undetectable current can then be applied to the skin and, while keeping the voltage constant, the change in resistance between two electrodes can
be measured. Skin conductance, which is measured in micro Siemens (µS), is simply the reciprocal of skin resistance.

Skin conductance measures have been used for decades in neuropsychological research to measure changes in a participant’s state of arousal. Two subcomponents are commonly identified: Skin Conductance Level (SCL), or tonic component, and Skin Conductance Response (SCR), or phasic component (Fowles et al., 1981; Boucsein et al., 2012). SCLs can range from 2-100 µS and the smaller SCRs, which are superimposed on the SCL, typically range from 0.01-5.0 µS (Fowles et al., 1981). While SCL varies continuously, SCRs occur within a discrete 1-5 second time window after the onset or offset of a stimulus (Fowles et al., 1981; Roth et al., 2012).

Many different types of stimuli have been used to elicit changes in skin conductance. Such stimuli include emotionally charged photographs of nudes or body mutilation (Tranel, 2004). Both smells (Bensafi et al., 2002), and auditory (Bradley and Lang, 1999) stimuli can cause a change in skin conductance. For example, studies have shown that SCR amplitude increases as the intensity of an auditory signal is increases (O’Gorman et al., 1970; Zuckerman et al., 1987); larger contrasts in intensity result in larger SCR amplitudes (O’Gorman et al., 1970). Khalfa and colleagues (2002) were able to demonstrate that music with strong emotional salience results in an increase in SCL.

To date, few studies have explored skin conductance as a method of assessing listening effort; the few that have concentrated on SCL, while no studies have assessed SCR amplitude changes during effortful listening. Mackersie and Cones (2011), reported that SCL increases in situations where greater intense cognitive listening demands are
placed on the participants. The researchers required participants to repeat either a single digit (1-10) delivered monaurally (least difficulty), dichotic pairs delivered binaurally (moderate difficulty) or two sets of dichotic pairs delivered binaurally (high difficulty). As the task became more challenging, the average SCL increased. The authors suggested that a challenging auditory task requires greater listening effort, which was observed physiologically as a change in SCL. Another study by Mackersie and colleagues (2015) examined how mean SCL varies across groups of listeners on a speech perception in noise task. They presented speech at a variety of SNRs and found that NH listeners had lower mean z-transformed SCLs than a group of listeners with hearing impairment; however, there were no significant changes in SCL within the groups across each of the different SNRs. Although promising, the results from both from these studies are difficult to evaluate given that stimulus presentation intervals and recording parameters, which are critical when assessing skin conductance, were not fully described. Additional studies are needed to determine whether skin conductance can serve as an objective measure of listening effort. If so, skin conductance may have several advantages over pupillometry. For example, the equipment used to record skin conductance is relatively inexpensive (Fowles et al., 1981; Boucsein et al., 2012). Secondly, skin conductance measures do not require that a participant fixate on a particular object or wear headgear, which suggests that skin conductance may be of broader utility for young or disabled participants. The recording system is also less obtrusive and more comfortable for participants to wear; in fact, these responses can be recorded from a variety of sites of the body (e.g., palms, fingertips, soles) (Fowles et al., 1981; Boucsein et al., 2012). All of these factors are important in clinical settings. However, skin conductance also has
limitations. For one, such measures are strongly affected by environmental changes and must be conducted in a climate controlled environment. Second, there is no clear method for assessing skin conductance—neither SCL or SCR—in the hearing science literature. Additionally, the known literature has used highly contrasting—and sometimes unpleasant—stimuli to evoke changes in skin conductance; it is not clear if skin conductance will be sensitive to small variations in listening effort.

IV Dual-Task Paradigm

By far, the most common way to quantify listening effort has been through the use of a “dual-task” paradigm. This experimental design requires that the listener be engaged in a primary task, normally speech perception, while performance on a secondary task, presented simultaneous, is monitored. The central tenant of the dual-task paradigm is that more effort is required as the complexity of the primary increases (Kahneman 1973; Ninio and Kahneman, 1974); this increase in effort is reflected as a decrement in secondary task performance. In 1974, Ninio and Kahneman published an article that should be seen as the inspiration for using a dual-task paradigm to assess listening effort. At the time, there were two main competing theories regarding divided attention: the filter theory and, as Ninio and Kahneman postulated, the effort theory. The filter theory (Triesman, 1964) assumes that "performance in divided attention should not differ materially from performance in focused attention;" that is, the addition of a secondary task should not greatly impact the primary task as the individual can “filter out” the secondary task. For example, a person driving a vehicle does not immediately crash once they turn on the radio.
Ninio and Kahneman (1974) found that while divided attention is possible, it comes at a premium: namely, a decline in secondary task performance. The authors demonstrated this as an increase in reaction time as the participants transitioned from a focused attention (single-task) condition, to the divided attention (dual-task) condition. For example, a person listening to the radio in a car may take longer to suddenly apply pressure to the breaks than someone who is not listening to the radio. The outcome largely depends on the difficulty of each task, and how closely related they are.

If a secondary task is too similar to a primary task, “divided attention” cannot be fully realized. This is perhaps why texting while driving is so inherently dangerous: the primary and secondary tasks overlap in both the motor and visual domains. Some of the many examples of secondary tasks used in listening effort studies include: word recall (Hornsby et al., 2013), performance in a driving simulator (Wu et al., 2014), and a simple, cued button press (Piccou et al., 2011, 2013; Wu et al., 2016), which assesses reaction time (it should be noted here that of these examples, word recall as a secondary task violates the primary principle of separate primary and secondary modalities laid out in the “divided attention” theory). By using a variety of secondary tasks, many studies have found that increasing the difficulty of the primary listening task results in a decrement in performance on the secondary task. This change in secondary tasks performance thus reflects a change in listening effort.

We know that digital noise reduction does not improve speech perception scores (Bentler, 2005; Chung et al., 2006; Bentler et al., 2008; Magnusson et al., 2013), yet it is incorporated into every modern hearing prosthetic. A likely explanation is that it does provide some benefit to the listener, but this benefit is not captured by speech perception.
testing alone. Sarampalis and colleagues (2009) used a dual-task paradigm with sentence repetition as the primary task and a cued button press as the secondary task to learn how digital noise reduction may reduce listening effort. They found that the listeners’ visual reaction time decreased with the noise reduction engaged, despite a lack of significant changes to speech perception abilities. These results suggest that listening effort may be more sensitive to changes provided by signal processing algorithms than speech perception testing alone; however, the dual-task paradigm is not without its own set of limitations. Wu et al. (2014) suggested that such paradigms may not be ecologically valid as in a laboratory environment, the secondary tasks are often novel and do not reflect typical real-world “multi-tasks” such as walking or writing (Wu et al., 2014). Recently, another study by Wu et al. (2016) showed that a dual-task paradigm is only a sensitive measure of listening effort within a small range of SNRs. At favorable SNRs, listeners are able to perform primary and secondary tasks with little difficulty. But, as the SNR becomes less and less favorable, the primary task becomes too difficult, listening effort saturates, and the participants “give up” on the primary task and instead focuses all of their effort on the secondary task (i.e., the cued button press). When a listener abandons one of the tasks—whether from fatigue, lack of attention, or some other factor—no inference about listening effort can be made. These observations suggest that for some study designs, the dual-task paradigm may not be the most appropriate method to assess listening effort.

V Conclusions

In recent years, listening effort has gained attention as a metric that may supplement or, in some cases, replace speech perception testing in the audiology clinic.
However, there is no widely accepted primary method of measuring listening effort. Rating scales are simple to administer and are inexpensive but are variable, subjective, and are difficult to use if the goal is to explore changes in listening effort across subjects. Dual-task paradigms can be challenging and provide only an indirect measure of listening effort. Studies have shown that pupillometry may provide useful insight into the cognitive demands placed on a subject in challenging listening environments, but the technology is expensive, not extremely robust to movement, and requires careful control of the testing environment. While they are less expensive to record than changes in pupil dilation and are more robust to changes in the test environment, research that supports the use of skin conductance as a measure of listening effort is limited. More exploration is needed before any of the metrics described in this review are ready to be incorporated into routine clinical practice. The sensitivity of these four metrics varies and to date, no study has reported within subject comparisons of these four different methods that have been proposed as measures of listening effort.

VI Study Goals

This study describes within and between participant comparisons of three different measures of listening effort: a dual-task paradigm, pupillometry, and skin conductance. A rating scale is used as a means to verify increasing task difficulty. We hypothesize that listening effort will change as the signal-to-noise ratio is manipulated. We predict that changes in listening conditions that require greater effort will result in a larger pupil diameter, greater skin conductance, increased reaction time, and higher self-reported assessments of listening effort. The clinical relevance of our findings may be
important and as such, will need to be reliable and repeatable. Thus, we will also compare the validity and reliability of each of the different metrics of listening effort.
CHAPTER III
METHODOLOGY

I Participants

Sample size calculation
First, Cohen’s d was calculated based on the smallest and largest mean values found in three studies: Zekveld et al., (2010), Mackersie and Cones (2014), and Winn et al., (2015). Those effect size values were 1.11, 9.81, and 15.6, respectively—all of these values are considered large effect sizes. As a conservative estimate, the means from the smallest effect size, in this case Mackersie and Cohens (2014), were entered into the following formula to calculate the number of participants required for this study:

\[ n = \frac{2\sigma^2 (Z_{\alpha} + Z_{\beta})^2}{\text{difference}^2} \]

\[ n = 2\left(\frac{.1259}{.8+1.96}\right)^2 \]

\[ \frac{.0317(7.6176)}{.0196} \]

\[ n = 12.32 \]

Where \( Z_\alpha \) and \( Z_\beta \) are the desired effect size and significance level, respectively. A minimum of 13 participants were required to achieve a large effect size.

20 normal hearing adults participated in two separate sessions for this study. Normal hearing was defined as pure tone air conduction thresholds of \( \leq 20 \) dB HL for octave frequencies from 250-4000 Hz in both ears. Participants were healthy adults between 18-50 years of age (14 females, 6 males; mean age=28.3, SD=8.65). All participants were native English speakers and had normal or corrected to normal vision in both eyes. 11 participants wore corrective lenses and nine did not require corrective
lenses of any kind. Two participants were left-handed and 18 participants were right
handed (see Table 1 for demographic information). Exclusion criteria included a positive
history of: hearing loss, age greater than 50 years at the time of testing, or a history of
psychiatric illness. Each participant was asked to remove his/her glasses and any eye
makeup prior to testing. All participants were consented in accordance with the
procedures submitted to the Institutional Review Board at the University of Iowa and
were compensated monetarily for their time.

II Stimuli

Auditory
IEEE sentences were used for speech perception testing. The entire corpus of
IEEE sentences consists of 72 lists of 10 simple sentences (i.e., “The birch canoe slid on
the smooth planks.”). Each sentence ranges from 2-4 seconds in length and is spoken by a
single male talker (IEEE, 1969). The ends of each sentence were aligned and the
beginnings were zero-padded in MATLAB® until each stimulus was exactly five
seconds long. Speech-shaped noise based on the long-term average speech spectrum of
20 different stimuli was used to calibrate the sentences to a level of 65 dBA in the sound
field; the same noise was also used during testing as the masker. The presentation level of
the sentences was varied to achieve three different SNRs: 0 dB SNR, -3 dB SNR, and -5
dB SNR. Unlike previous studies, the noise was not gated but presented continuously
during the speech perception task. The purpose of this was to limit any latent reactions to
the noise itself.

Visual
Visual prompts were used to cue the participants during each task. Initially, a
white cross was presented on a dark background to alert the participant that a trial was
going to begin. The cross changed from white to green at a set interval after the conclusion of each sentence; this interval varied based on paradigm and will be described in further detail in the following sections. The green cross alerted the participant to repeat back the sentence. Each visual stimulus was evaluated in Adobe Photoshop® and with a light meter to ensure equal luminance (~20 lux).

**Presentation and Recording**

Presentation of the visual and auditory stimuli was controlled using MATLAB®. The pupillometry and reaction time responses were recorded by Attention Tool® software (version 6.2, iMotions Global). Skin conductance recordings were acquired using Acqknowledge® software (v. 3.9).

**III Procedures**

The test protocol was divided into two sessions with each lasting approximately three hours. The order of testing (pupillometry, dual-task, and skin conductance), was counterbalanced between sessions to reduce any variance attributed to test order. Each SNR was also randomized within a given condition to control for fatigue and any learning effects.

**Pupillometry**

Participants were seated in a chair directly in front of a stationary table located inside a sound-treated booth. They were positioned 1 m from a loudspeaker at 0° azimuth (Figure 1). A Tobii X2-60 distal eye tracker, which sampled data at a rate of 60 Hz, was secured to the bottom of a computer monitor (Dell, 4:3 aspect ratio with a resolution of 1280x960 pixels). A chin rest was fastened to a stationary table at a fixed distance of 60 cm from the monitor. The height of the chin rest and monitor were adjusted so that the participant's eyes were located in the center of the recording field. The ambient light
levels in the sound booth were set to their maximum and minimum values and the light level used for testing was adjusted such that their pupil diameter was located in the middle of their dynamic range. Pupil diameter measurements from the right eye were used for analysis.

A practice session consisting of one block of 10 sentences was used to familiarize the participant with the task. After completion of the practice session, the participant underwent 20 trials (2 blocks of 10 sentences) at each of the four SNR values. After the completion of each block, the participant was asked to rank their degree of effort on scale of 0-100 (as in Hart and Staveland, 1988); other than the floor and ceiling values, there were no explicit numerical constraints placed on the rating scale.

**Pupillometry analysis**
Prior to averaging, eye blinks were extracted and a 10-point moving average filter was used to smooth the data. Data that were excessively noisy or contained a large number of eye blink artifacts were removed by the author; ultimately, <2% of the data was excluded due to large artifacts. The filtered and de-blinked data were then averaged together for each condition. These filtered tracings were then used to determine the task-evoked pupil response, which was calculated by finding the mean pupil diameter during the 1-second interval preceding each stimulus and subtracting that value from every remaining sample in the trial. Several of the participants had pupil values that actually decreased during the initial listening phase, which is a departure from the data reported in previous studies. Many values were large, but negative with respect to baseline. To account for these differences, rather than obscure the maximum dilation, the peak-to-peak change in pupil diameter was analyzed in a time window beginning 2 seconds before the
end of the sentence and ending at the response cue, which always occurred 2 seconds after each sentence ended. Figure 2 shows a schematic of the pupillometry paradigm.

**Dual-Task**

Speech perception served as the primary task and a simple button press, cued visually, as the secondary task. The physical setup used for the dual-task procedure was the same as that used for pupillometry, except the participant was asked to place one finger of their dominant hand on the "enter" button of a standard computer keyboard. The participant was again asked to gaze at a computer screen and follow the visual cues. After the sentence began, the participant was cued to depress the response button when the words “Press enter” appeared on the screen. The time between when the sentence began and when the participant was cued to press the response button was randomized between 0.5 and 2 seconds; this cue always occurred before the response prompt. 20 replications of only the reaction time task served as the single-task baseline. Following the single-task baseline and similar to the pupillometry and skin conductance paradigms, a practice session consisting of 10 sentences was conducted to familiarize the participant with the dual-task paradigm. This was done to keep the number of practice trials for the dual-task commensurate with the practice trials for the physiologic paradigms. The participants were encouraged to prioritize the speech perception task while attempting to press the "enter" button as fast as possible after being cued to do so. After each block of 10 sentences, the participant was asked to rate how much effort was required to complete the task, using the same scale as they had for pupillometry.

**Dual-task analysis**

Median reaction time values are more robust to outliers and as such, are typically used for the analysis of reaction time data (Sarampalis et al., 2009; Wu et al., 2016).
Median reaction time values were calculated for the 20 trials of the baseline condition (button press only) and the 20 trials for each subsequent condition (quiet, 0 dB SNR, -3 dB SNR, and -5 dB SNR) and then averaged together. Each participant’s mean (average of the median) reaction time value for the baseline condition was subtracted from the mean value of each experimental condition to achieve a “change in reaction time from baseline” metric; a value of “0” indicated no change from baseline while positive and negative values indicated an increase or decrease in reaction time from the baseline condition, respectively.

**Skin Conductance**

Before testing, two Ag-AgCl electrodes with a 0.05 molar NaCl solution were placed over the thenar and hypothenar eminences of the non-dominant hand. The electrodes were connected to a Biopac GSR 100 amplifier and MP100 recording unit. 16 participants required an amplifier gain of 2 µS/V; four participants had much larger SCRs and required a lower gain setting of 5 µS/V. The participant was encouraged to rest his/her hand palm up in their lap or on the examination table and was instructed not to move their arm, hand, or fingers during testing.

After the electrodes were secured, a five-minute acclimatization period during which time the participant was encouraged to relax was initiated. Skin conductance activity was recorded during this time, even though no stimuli were presented. This acclimatization period allows for sufficient time for the conductive paste to be absorbed and the tonic activity to stabilize (Boucsein, 2012; Roth et al, 2012). After this period, data collection proceeded in a fashion similar to that described for pupillometry, but with a longer interval between the sentence offset and the response cue, which always happened five seconds after the end of each sentence, as well as a longer interval between
trials (~18-21 seconds). The participant engaged in a 10-sentence practice session before data collection began. With the exception of the first three participants, temperature and humidity levels were monitored and recorded for the start and end of each session. Figure 3 shows a schematic of the skin conductance paradigm.

**Skin conductance analysis**

A low-pass filter (20 Hz cutoff) was used to remove extraneous noise and a high-pass filter (0.05 Hz cutoff) was used to extract the SCRs. The acclimatization periods were used as baseline (non-task) conditions. A noise floor was estimated by taking the mean value over a window beginning 30 seconds after the start of the acclimatization period, and ending at 4.5 minutes later (participants 1 and 2 had shorter baseline values due to equipment error). Next, the average quantity of SCRs present during the no-stimulus conditions were determined by extracting all SCRs with a peak-to-peak amplitude equal to or greater than 0.03 µS and with a rise time of 1-3 seconds within the acclimatization period. The baseline period was then divided in 7-second long segments, and the average quantity and amplitude of the responses per 7-second time window were calculated; these values became the baseline SCR quantity and amplitude.

The same criteria were used to extract the SCRs from two separate time windows during each trial period. Both windows were seven seconds in duration. The first began 3-seconds after the start of each trial and the second began 1-second after the appearance of the response cue. These epochs are two seconds longer than the 3-5 seconds recommended by other researchers (Boucsein, 2012; Roth et al., 2012; Fowles et al., 1981); however, these lengths are necessary as SCRs can occur at either the onset or the offset of a stimulus. Bi-and tri-phasic responses were counted individually if they met the above temporal and amplitude requirements and produced a recovery limb that extended
below a line tangent to the peak of the response. SCRs, when present, varied greatly across participants (from .03 µS to 3 µS). As such, each participant’s SCR was square root transformed to reduce variance and better approximate a normal distribution.
<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
<th>Sex</th>
<th>Handedness</th>
<th>Eye Color</th>
<th>Corrective Lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>M</td>
<td>Right</td>
<td>Blue</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>M</td>
<td>Right</td>
<td>Hazel</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>F</td>
<td>Right</td>
<td>Hazel</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>F</td>
<td>Right</td>
<td>Brown</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>F</td>
<td>Right</td>
<td>Blue</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>47</td>
<td>M</td>
<td>Right</td>
<td>Blue</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>M</td>
<td>Right</td>
<td>Blue</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>F</td>
<td>Left</td>
<td>Blue</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>F</td>
<td>Right</td>
<td>Blue</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>F</td>
<td>Right</td>
<td>Blue</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>F</td>
<td>Right</td>
<td>Blue/green</td>
<td>No</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>F</td>
<td>Right</td>
<td>Brown</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>19</td>
<td>M</td>
<td>Right</td>
<td>Brown</td>
<td>No</td>
</tr>
<tr>
<td>14</td>
<td>21</td>
<td>F</td>
<td>Right</td>
<td>Blue</td>
<td>No</td>
</tr>
<tr>
<td>15</td>
<td>40</td>
<td>F</td>
<td>Left</td>
<td>Brown</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>F</td>
<td>Right</td>
<td>Hazel</td>
<td>Yes</td>
</tr>
<tr>
<td>17</td>
<td>24</td>
<td>F</td>
<td>Right</td>
<td>Brown</td>
<td>Yes</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>F</td>
<td>Right</td>
<td>Hazel</td>
<td>No</td>
</tr>
<tr>
<td>19</td>
<td>36</td>
<td>M</td>
<td>Right</td>
<td>Blue</td>
<td>Yes</td>
</tr>
<tr>
<td>20</td>
<td>36</td>
<td>F</td>
<td>Right</td>
<td>Blue</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1. Demographic information
Figure 1 Laboratory setup. This setup was used for all three paradigms (the chin rest was only used for pupillometry).
Figure 2 Pupillometry paradigm schematic.
Figure 3 Skin conductance paradigm schematic.
CHAPTER IV
RESULTS

I General Procedures

Planned analyses were conducted using linear mixed-models analyses in SPSS® software. The main justification for this procedure is that it allows researchers to “emphasize patterns of change” and to “emphasize individual differences” (Krueger, 2004). Fixed effects generally included the condition (SNR), session, and their interaction. A random intercept was also included in each model and the covariance type was set to “unstructured.” Several models were generated for the speech perception scores, rating scales, and each paradigm. Gradually, the non-significant fixed effects were removed and the model with the lowest Akaike’s Information Criterion (AIC) was used for analysis. The details of the model used for each paradigm will be described in the subsequent sections.

P-values were not corrected due to the counterbalancing of conditions (SNRs), sentences within the conditions, and testing paradigms. While correcting for multiple comparisons does reduce the chance of committing a type I error (reporting a significant finding when in fact, there was none), it does so at the expense of an increased risk of committing a type II error (failing to report a truly significant finding) (Perneger, 1998). The exploratory nature of this study suggests that a type II error would be more serious than a type I error.

The distributions for all data were examined prior to analysis. Any data that exhibited undesirable skewness (>±1.5) were transformed to allow for a more normal
distribution. Any applied transformation will be described under its corresponding subsection.

Outliers were also examined and removed according to SPSS® criteria (Ghasemi and Zahediasl, 2012):

\[
X \leq (Q1 - 1.5 \times (Q3 - Q1))
\]

\[
X \geq (Q3 + 1.5 \times (Q3 - Q1))
\]

In the equation, Q1 and Q3 are the first and third quartiles (25\textsuperscript{th} and 75\textsuperscript{th} percentiles), respectively. As a result, values that were 1.5 times greater than the interquartile range were removed from analysis.

**II Percent Correct And Perceived Effort**

Speech recognition performance as a function of testing paradigm and SNR is shown in Figure 4; the data from both sessions were combined. The percent correct data were plotted using histograms and analyzed visually for normality. A rationalized arsine transformation (Studebaker, 1985) was applied to the data and resulted in less skew and an improved distribution. The height of each bar represents the mean performance and the error bars represent +/-1 standard error. Each testing paradigm is grouped by experimental condition with the light tan indicating the quiet condition, progressing to a dark red for the -5 dB SNR condition. Performance was at or near ceiling when testing was conducted in quiet and, as expected, declined as the SNR decreased during both sessions.

The final linear model included the following fixed factors: method (pupillometry, dual-task, or skin conductance), session, SNR, and the three-way interaction of method*session*SNR. The effect of SNR was significant (F3,
but the remaining effects of method \((F_{2, 437} = 2.256, p = 0.106)\), session \((F_{1, 437} = 2.494, p = 0.115)\), and the three-way interaction of method*session*SNR \((F_{17, 437} = 0.573, p = 0.912)\), were not significant. Pairwise follow-up testing further indicated that each successively more difficult condition resulted in significantly poorer speech discrimination scores \((p < 0.0001)\). These findings indicate that speech perception performance declined significantly as the SNR decreased, and that the performance across the testing paradigms and sessions was the same.

Participants were also asked to rank their perceived listening effort at the completion of each block of 10 sentences. A value of “0” indicated no effort while a value of “100” indicated maximum effort; no numerical restraints other than these floor and ceiling values were placed on the participants. Subjective listening effort ratings as a function of testing paradigm and SNR are shown in Figure 5; the data from both sessions were combined. The height of each bar represents the mean performance assigned by the listener and the error bars represent +/-1 standard error. The results are grouped by testing paradigm and include the responses for each of the four SNRs. The light tan indicates the values reported for the quiet condition, progressing to a dark red for the values reported during the -5 dB SNR condition.

Figure 5 confirms that for each of the three testing paradigms, the quiet condition required very little effort. Also, subjective effort clearly increases as SNR decreases. This finding validates our choice of SNRs. We argued that as we lowered the SNR, speech recognition would decline and the participants would indicate that the more difficult conditions require greater listening effort.
The data were plotted using histograms and analyzed visually for normality. A rationalized arsine transformation (Studebaker, 1985) was applied to the data and resulted in less skew and an improved distribution. The resulting histograms suggested no significant problems with normality, and there were no outliers present. The final linear model included the following fixed factors: SNR and the three-way interaction of method*session*SNR. The effect of SNR was significant ($F_{3, 437} = 667.812$, $p<0.0001$) but the three-way interaction of method*session*SNR ($F_{20, 437} = 0.818$, $p=0.692$), was not significant. Pairwise follow-up testing further indicated that each successively more difficult condition resulted in significantly greater subjective listening effort ($p<0.0001$). These findings indicate that subjective listening effort increased significantly as the SNR decreased, and that the performance across the testing paradigms was the same.

### III Pupillometry

A number of investigators have used pupillometry to explore the physiologic processes associated with listening effort (Zekveld et al., 2010; Winn et al., 2015; Ohlenforst et al., 2017). We measured peak-to-peak task-evoked pupil dilation while participants listened to sentences in a variety of SNRs. We hypothesized that peak-to-peak pupil dilation would systematically increase as the SNR decreased and our findings were consistent with that hypothesis.

**Figure 6** presents the grand mean change in peak-to-peak pupil diameter as a function of SNR; the data from both sessions were combined. The light tan line represents quiet condition, the brown line the 0 dB SNR condition, and the successively darker red lines the -3 dB and -5 dB SNR conditions, respectively. Previous studies calculated the maximum pupil diameter occurring within a window
beginning -0.5 \textit{before} the offset of the stimulus, and ending 2 seconds \textit{after} the offset of the stimulus. This traditional analysis window in \textbf{Figure 6} is represented by a square consisting of dashed lines. We have departed from these traditional techniques by using a peak-to-peak amplitude calculation. Our analysis window begins approximately 2 seconds before the offset of the stimulus, where the minimum amplitude value is observed. The end of our analysis window occurs in the same location as that used for maximum peak calculations: 2 seconds after the offset of the stimulus. The solid square in \textbf{Figure 6} represents the peak-to-peak analysis window used in the present study.

For all of the SNRs in \textbf{Figure 6}, pupil size remains stable or begins to constrict during the baseline and extending into the first 2-3 seconds of the listening interval. Pupil diameter then begins to increase approximately two seconds before the end of the sentence; this period of dilation continues until approximately one second after the end of the sentence. At this point, pupil diameter begins to constrict for all four of the grand mean tracings. Towards the end of the silent period, the pupil tracing for the quiet condition continues to constrict, although it never fully returns to baseline, before quickly dilating approximately two seconds after the response cue appears. The tracings for the three noise conditions display a different trend: they begin to dilate towards the end of the silent period, 0.5-1 seconds \textit{before} the response cue appears; these tracings do not return to baseline within this analysis window. All four tracings exhibit further dilation 1-2 seconds after the response cue appears.

The data were plotted using histograms and analyzed visually for normality. The resulting histograms suggested no significant problems with normality. Outliers were
assessed and in all, 3 data points (1.89%) were excluded from the analysis. The final linear model included two fixed factors: SNR and session. The effect of SNR was significant \((F_{3,137.028}=5.920, p=0.001)\), but session \((F_{1,137.031}=2.927, p=0.089)\) was not. The interaction term was not included in the model as its effects were not significant \((F_{1,137.021}=.166, p=0.919)\) and resulted in a higher AIC value, meaning that the fit of the linear model was poorer with the interaction term included. The non-significant main effect of session indicates that there was no significant overall increase or decrease in peak-to-peak pupil diameter across the two sessions. Follow-up testing revealed the following significant pairwise differences in peak-to-peak amplitudes: quiet and -3 dB SNR \((p=.007)\), quiet and -5 dB SNR \((p<.0001)\), 0 dB SNR and -3 dB SNR \((p=.028)\), and 0 dB SNR and -5 dB SNR \((p=.003)\).

**Figure 7** depicts the same data as **Figure 6**, but includes a ribbon extending +/- 1 standard error about the mean. The light tan ribbon represents the quiet condition, which is compared to the results from each of the three noise conditions. The dark brown, red, and dark red tracings represent the three successively more challenging noise conditions: 0 dB SNR, -3 dB SNR, and -5 dB SNR. There is a large degree of overlap between the quiet and 0 dB SNR conditions, with greater separation between the quiet condition and both the -3 dB SNR and -5 dB SNR conditions.

**Figure 8** shows the grand mean peak-to-peak change in pupil dilation values calculated from **Figure 6**. The height of each bar represents the mean peak-to-peak pupil dilation for each of the experimental conditions and the vertical lines represent +/-1 standard error. As the SNR decreases, more listening effort is required and the mean peak-to-peak increases. Again, the significant differences were between the following
SNRs: quiet and -3dB SNR (p=0.007), quiet and -5 dB SNR (p<0.0001), 0 dB SNR and -3 dB SNR (p=0.028), and 0 dB SNR and -5 dB SNR (p=0.003).

**IV Dual-Task**

The dual-task paradigm is a behavioral task that has also been used to assess listening effort (Wu et al., 2016; Hornsby, 2013; Picou et al., 2013). This dual-task paradigm consists of a primary task—usually speech recognition—and a simultaneous secondary task. We used speech recognition as our primary task and a visually cued button press as our secondary task. Single-task (cued button press only) and dual-task (speech recognition plus cued button) values were obtained for each session. During the dual-task, the participants listened to sentences and were cued to press a button at random during the sentence presentation. The average median cued reaction time value was calculated for each participant for the baseline condition and at each SNR. The baseline value was then subtracted from the value at each SNR to obtain a “change in reaction from baseline” metric. The grand mean value for each SNR was then obtained by averaging across participants. A longer reaction time indicates greater listening effort.

The data were plotted using histograms and analyzed visually for normality. The resulting histograms suggested no significant problems with normality. Outliers were assessed and in all 12 data points (7.5%) were identified and excluded from further analysis. The final linear model included two fixed factors: SNR and session. The fixed effects of SNR ($F_{3,140}=3.148$, p<0.027) and session ($F_{1,140}=30.210$, p<0.0001) were significant. The interaction of SNR*session was not included in the final linear model as it was not significant ($F_{3,140}=0.385$, p=0.764) and resulted in a higher AIC value (indicating a poorer linear model). Pairwise testing indicated a significant difference
between the quiet and -3 dB SNR conditions (p=0.028) as well as between the quiet and -5 dB SNR conditions (p=0.004)). Longer reaction times indicate increased listening effort, but the only significant differences observed were between the quiet and two most challenging conditions. Additionally, reaction time decreased across all conditions from the first session to the second session by an average of 42.963 ms. As there was no significant interaction, this constant is strong evidence for a practice effect.

Figure 9 shows the change in reaction time relative to the baseline for each SNR. As stated previously, there was no interaction between session and SNR, so the data were analyzed together. The height of each bar represents how the grand mean reaction time for each SNR increases as the task is made more challenging; the error bars represent +/-1 standard error. The light tan represents the quiet condition, while the dark brown, red, and dark red the 0, -3, and -5 dB SNR conditions, respectively.

Figure 10 splits the combined data into their respective sessions to illustrate the significant main effect (decrease in overall reaction time) that occurred from session one to session two.

V Skin Conductance

Skin conductance responses were measured using electrodes attached to the thenar and hypothenar eminences of the non-dominant hand. First, a 5-minute baseline recording was obtained during which time the participant was not engaged in any task. Then, participants listened to sentences presented in quiet, 0 dB SNR, -3 dB SNR, and -5 dB SNR while SCRs were measured concurrently. Responses in each analysis window were counted as present if they were \( \geq 0.03 \, \mu S \). SCRs rarely occur synchronously within
an analysis window and as such, grand mean amplitude values can obscure true trends (Boucsein, 2012; Roth et al., 2012). Thus, SCR amplitude values were calculated separately for each response, and then averaged together. There is no known upper limit to SCR amplitude or SCR quantity as they pertain to listening effort, so no large or frequent responses were excluded as long as they met the specifications laid out in our methodology (occurring 1-5 seconds after the beginning or end of a stimulus; minimum amplitude of $\geq 0.03 \mu S$; rise time of 1-3 seconds). The average quantity of SCRs $\geq 0.03 \mu S$ within each 7-second analysis window was also calculated. The mean participant values were then averaged together to obtain both a grand mean SCR amplitude and grand mean SCR quantity.

Ambient environmental temperatures can impact thermoregulatory sweating, which in turn can confound skin conductance measures. We measured the temperature in the sound booth at the beginning and end of each session for all but the first three participants. Figure 11 shows a plot encompassing the range of ambient room temperatures for both sessions. The top of each bar represent the mean temperature, while the error bars represent +/-1 standard error; the light-colored boxes signify the average starting temperature for each session while the dark red boxes represent the average ending temperature for each session. The dashed lines represent the ideal range of ambient temperatures (71.6-75.2ºF) for recording skin conductance responses (Vertungo et al. 2003).

A repeated measures analysis of variance was conducted on the ambient temperature data. There was a significant effect of starting versus ending temperature ($F_{1,16}=67.102, p<0.0001$); there was no significant effect of session ($F_{1,16}=1.073, p<0.316$).
and the interaction between session*starting versus ending temperature was not significant \( (F_{1, 16}=0.790, p<0.387) \). Pairwise testing indicated that overall, the average ending temperature was significantly \((1.815^{\circ}F)\) warmer than the starting temperature. Even so, the mean temperatures \((71.003-74.694^{\circ}F)\) generally stayed within the recommended range.

Figures 12-14 show skin conductance responses from three individual participants. These figures display area plots obtained by combining responses across all of the experimental conditions. These three participants were chosen to demonstrate the range of skin conductance responses that occurred within and across sessions. The top figure in each group displays the results from the first session, while the bottom figure displays the results from the second session. Each area plot is lightly colored and, when layered on top of each other, some areas become darker while other areas remain lightly colored. Thus, areas of darker color indicate greater SCR activity, while areas of lighter color indicate sparser SCR activity. The start of the stimulus, silent period, and response intervals are indicated on the \(x\)-axis. Both analysis windows are labeled and indicated by the boxes consisting of dashed lines. The figures contain the responses from all trials for a given participant.

The top portion of Figure 12 depicts SCRs from the first session of one participant. The dark areas indicate a greater quantity of SCRs, while the light areas indicate fewer SCRs. The responses in the bottom portion of Figure 12 are from the second session. Note that the first analysis window is more lightly shaded than the analysis window from the first session. This indicates fewer SCRs occurred during the second session, but only during the first analysis window.
Figures 13 and 14 show results from participants that had very few responses during the first session and significantly more and larger responses during the second session. In Figure 13, there are few SCRs present during the first session, and most of them lie within the second analysis window. During the second session, the responses are both greater in quantity and larger in amplitude, but the area plots are lightly colored, indicating fewer SCRs within the analysis windows. Figure 14 shows very few measureable SCRs during the first session, and few–but substantially larger–SCRs during the second session.

The group SCR amplitude and quantity data were plotted using histograms and analyzed visually for normality. A square root transformation was applied to the SCR amplitude data and resulted in a normal distribution (Fowles et al., 1981; Boucsein, 2012; Roth et al., 2012); no gross violations of normality were observed for the SCR quantity data. Outliers were determined and excluded from the analysis. In all, 16 data points (4%) were excluded from the amplitude analysis and six (1.5%) points were excluded from the quantity analysis.

The final linear model for SCR amplitude included three fixed effects and one interaction: session, window, SNR, and session*SNR. All four terms were significant ($F_{1, 362.218} = 4.332, p=0.038$; $F_{1, 362} = 7.877, p=0.005$; $F_{4, 362.17} = 3.010, p=0.018$; $F_{4, 362.042} = 4.181, p=0.003$, respectively). The significant interaction suggests that the grand mean SCR amplitude for at least one SNR changes significantly from session 1 to session 2. Pairwise testing indicates that the 0 dB SNR condition was significantly smaller than the baseline amplitude ($p=0.003$) and the quiet ($p=0.004$), the -3 dB SNR ($p=0.002$), and -5 dB SNR ($p<0.0001$) conditions during the first session. Pairwise testing also indicates
that the baseline SCR amplitude was significantly smaller than all four of the SNRs during the second session (p=0.022, 0.017, 0.007, 0.029 for each progressively more challenging condition). The significant effect of session indicates that, across all conditions, the average SCR amplitude was significantly larger during the second session. The effect of window demonstrates that that on average, the responses that occurred during the second analysis window (1-second after the response cue appeared) were larger in amplitude than those that occurred during the first analysis window.

**Figure 15 and 16** display the grand mean SCR amplitude for each session as a function of SNR. **Figure 15** displays the results from the first session while **Figure 16** displays the results from the second session. The height of each bar indicates the mean, square root transformed SCR amplitude. The error bars represent +/-1 standard error. The first bar in each graph indicates the average SCR amplitude during the baseline condition; the remaining bars represent each successively more challenging experimental condition.

**Figure 17** shows the grand mean quantity of SCRs for the baseline condition and each of the experimental conditions; the data from both sessions were combined. The height of each bar indicates the mean SCR quantity. The error bars represent +/-1 standard error. The bars on the left-hand side of each pair represent the quantity of SCRs from the first analysis window, while the bars on the right-hand side represent the quantity of SCRs from the second analysis window. The quantity of SCRs \( \geq 0.03 \) \( \mu \)S is similar across the baseline condition and all experimental conditions during the first analysis window. The quantity of SCRs during the second analysis window is greater during each of the experimental SNRs than the baseline condition; there is no
difference in SCR quantity across each of the four SNRs during either analysis window.

The final linear model for SCR quantity included the fixed effect of SNR, which was not significant ($F_{4, 373.994} = 1.362, p=0.247$) and analysis window, which was significant ($F_{4, 374.033} = 48.973, p<0.0001$); there was also a significant interaction of analysis window*SNR ($F_{4, 373.988} = 3.113, p=0.015$). The main effect of session was not included in the final model because it was non-significant and resulted in a larger AIC value. Pairwise testing indicated that there were approximately 3.522 more SCRs in the second analysis window than the first ($p<0.0001$); the main effect of SNR is best interpreted with its interaction term. Pairwise testing indicated no significant differences in SCR quantity between any of the conditions within the first analysis window. There were significantly fewer SCRs during the baseline condition compared to the second analysis window for the quiet ($p=0.012$), 0 dB SNR ($p=0.023$), -3 dB SNR ($p=.008$), and -5 dB SNR conditions ($p<0.0001$). There were no statistically significant findings for SCR quantity across the four SNRs within each analysis window.

Previous studies have shown that highly arousing images and sounds cause changes in SCR amplitude (Tranel, 2004; Bradley and Lang, 1999). So, it is reasonable to suggest that a participant may become more anxious or aroused after improperly hearing a sentence or before expressing an uncertain response. So, to examine this effect further, we segregated SCR tracings based on performance: one category for tracings associated with 100% correct speech perception performance or tracings associated with any score less than 100% correct. The reason for this post-hoc analysis was to further explore the possible differences of SCR quantity and SCR amplitude between conditions that were
likely low-arousal (perfect speech perception performance) and likely high-arousal (imperfect speech perception performance). Only the data from the 0, -3, and -5 dB SNR conditions was segregated in this manner; the quiet condition was not assessed because performance was generally at ceiling and perceived effort was ranked low. The responses from both analysis windows were combined. We expected to see differences in SCR quantity and amplitude between the perfect and imperfect speech perception performance for each experimental condition.

**Figures 18** shows SCR amplitude as a function of perfect versus imperfect sentence recognition; the data from both sessions and both analysis windows were assessed together. The height of each bar represents the mean square root SCR amplitude for each condition and the error bars represent +/-1 standard error. A square root transformation was applied to the SCRs and the data was examined and outliers were removed in the same fashion as described during earlier results. In all, 3 (2.5%) tracings were excluded from the analysis. The final linear model for the amplitude analysis included the fixed effect of correct vs. incorrect sentence recognition, which was significant ($F_{1,95.454} = 8.562, p=.004$). This finding indicates that on average, a participant’s SCR amplitude was significantly larger when they did not understand the sentence than when they did.

**Figure 19** shows SCR quantity as a function of perfect versus imperfect sentence recognition; the data from both sessions and both analysis windows were assessed together. SCR quantity data were plotted and assessed for violations of normality. After removing four outliers, the data exhibited an approximately normal distribution. The final linear model included the main effects of SNR ($F_{2,95.87} = 3.511, p<0.034$) and correct
versus incorrect performance ($F_{1, 96.03} = 27.894, p<0.0001$), as well as the interaction of SNR*correct vs. incorrect speech recognition performance ($F_{2, 96.027} = 30.472, p<0.0001$). All three were highly significant indicating that the quantity of SCRs was different when the participant achieved perfect sentence recognition than when they did not fully understand the sentence and this effect was significant at each level of SNR. Despite these significant pairwise findings, the data are shown collapsed across SNR to simplify interpretation. Even with this simplification, this figure demonstrates that participants had significantly more SCRs when they did not understand the sentence compared to when they did, thus paralleling the SCR amplitude findings of Figure 18.

A central tenant of this study was that SCR quantity and amplitude would increase as SNR decreased. Post Hoc testing was performed on the data to determine the relationship between SCR quantity and SCR amplitude. 16 data points (5%) were determined to be outliers and were excluded from the analysis. Figure 20 shows this relationship for both analysis windows, collapsed across both sessions. Pearson correlation coefficients were calculated for each of these figures and a strong, positive linear relationship exists between the SCR quantity and SCR amplitude between both time windows ($r=.892$ and $r=.674$, respectively). These relationships demonstrate that fewer responses correlate with smaller amplitudes, and a greater quantity of responses correlate with larger SCR amplitudes. Simply put, individuals who had more responses generally had larger amplitudes while those with fewer responses generally had smaller amplitudes.
VI Individual Linear Trends

The remaining section of the analysis departs from analyzing group trends to focus on individual trends. To our knowledge, no previous study has reported individual trends for any of the measures used in this study. Single participant values were analyzed separately to examine how the three measurements of listening effort and the subjective rating scales change on an individual basis. We made the A-priori assumption that the individual effects would be largest between the quiet condition and the -5 dB SNR condition or, as in the case of the skin conductance responses, between the baseline (no stimulus) condition and the -5 dB SNR condition. These assumptions were validated by the fact that all participants subjectively rated the -5 dB SNR condition as more challenging than the quiet (or baseline) condition. We normalized each participant’s starting values by dividing them by itself, resulting in a value of “1.” We then normalized the values in the challenging condition by dividing them by the values from the quiet condition. By doing so, we emphasized the linear change across these two SNRs rather than emphasizing the raw, measured responses. We then fit the change across these conditions with a first-order (linear) polynomial to assess the trend of each individual slope. As previous statistical testing showed that there were no significant differences across sessions, and for the sake of simplicity, we combined the data from both sessions. Slopes that were ≥0 were considered to have a positive slope, while values <0 were considered to have a negative slope. A positive slope represents increasing effort while a negative slope represents decreasing effort. All participants start at the same point and diverge. Again, the main reason for this was to demonstrate, at the individual level, how many participants’
subjective and objective results behaved in the anticipated manner; that is, increased as the task became more challenging.

**Figure 21** displays the normalized linear relationship for perceived effort between the quiet condition and the -5 dB SNR for pupillometry, **Figure 22** is for the dual-task paradigm, and **Figure 23** is for skin conductance. In each of the figures, the first 10 subjects are represented by unique colors and symbols but with solid lines; the remaining subjects are represented by the same colors and symbols but with dashed lines. In all instances, 100% of the participants displayed the expected trend: that is, a positive slope between the easiest and the most difficult task.

**Figure 24** displays the normalized trend between the quiet condition and the -5 dB SNR condition for the normalized change in peak-to-peak pupil diameter. The combination of line (solid or dashed), color, and symbol represent a unique participant. The first 10 subjects are represented by unique colors and symbols but with solid lines; the remaining subjects are represented by the same colors and symbols but with dashed lines. A value that lies within the light green box, when present, indicates a negative trend between the easy and difficult conditions, which was not anticipated. The average subjective rating for the 0 dB SNR and -5 dB SNR conditions are given at the bottom of the chart. For this paradigm, 75% of participants displayed the expected positive trend.

**Figure 25** displays the normalized trend between the quiet condition and the -5 dB SNR condition for the normalized change in secondary task performance for the dual-task paradigm. The first 10 subjects are represented by unique colors and symbols but with solid lines; the remaining subjects are represented by the same
colors and symbols but with dashed lines. The average subjective rating for the quiet and -5 dB SNR conditions are given at the bottom of the chart. For this paradigm, 80% of participants displayed the expected positive trend.

Figure 26 displays the normalized trend between the baseline (no stimulus) condition and the -5 dB SNR condition for the normalized square root change in SCR amplitude. The first 10 subjects are represented by unique colors and symbols but with solid lines; the remaining subjects are represented by the same colors and symbols but with dashed lines. The average subjective rating for the 0 dB SNR and -5 dB SNR conditions are given at the bottom of the chart. For this paradigm, 50% of participants displayed the expected positive trend.

Figure 27 displays the normalized trend between the baseline (no stimulus) condition and the -5 dB SNR condition for the normalized change in SCR quantity. The first 10 subjects are represented by unique colors and symbols but with solid lines; the remaining subjects are represented by the same colors and symbols but with dashed lines. The average subjective rating for the 0 dB SNR and -5 dB SNR conditions are given at the bottom of the chart. For this paradigm, 55% of participants displayed the expected positive trend.

VII Test-Retest Reliability

Figure 28 displays the test-retest reliability for the subjective rating scales and the three objective measures during the most difficult condition (-5 dB SNR). The different paradigms are presented superimposed and are thus shown without units. In both figures, the dashed gray line represents the theoretical value where Session I=Session II. In the larger figure, the black line represents pupillometry, the dark
brown line represents the subjective rating scales, the red line represents the dual-task paradigm, and the light blue line represents SCR quantity. The data are presented in arbitrary units. Pearson correlation coefficients showed a significant test-retest reliability between all of the measures for the -5 dB SNR condition (r=.79, p<0.0001 for the rating scales; r=.79, p<0.0001 for pupillometry; r=.79, p<0.0001 for the dual-task; r=.79, p<0.0001 for SCR quantity). The inset shows the Pearson correlation coefficients for SCR amplitude during the -5 dB SNR condition. The solid black line and symbols represent all of the SCR amplitude data and the red circle represents an outlier in the data. Before the outlier was removed, the test-retest reliability was not significant (r=.19, p<0.431). After removing the outlier (yellow line and symbols), the reliability became significant (r=0.62, p=0.005).
Figure 4 Speech perception ability as a function of experimental condition and grouped by testing paradigm. As the testing condition becomes more challenging, speech perception abilities decline significantly. The error bars represent +/-1 standard error. These data are collapsed between sessions. Each successively poorer condition is significantly different than the previous (p<0.0001).
Subjective listening effort as a function of experimental condition and grouped by testing paradigm. As the testing condition becomes more challenging, the participants rank their experience as requiring significantly more listening effort. The error bars represent +/-1 standard error. There are no significant findings across the paradigms. Each successively poorer condition is significantly different from the previous (p<0.0001).

Figure 5 Subjective rating as a Function of Paradigm and SNR
Figure 6 Grand mean change in peak-to-peak pupil diameter from baseline condition. The performance is significantly different between the quiet and -3 dB SNR, the quiet and -5 dB SNR, and the 0 dB SNR and -5 dB SNR conditions. During the quiet condition the pupil tracing generally remains flat until approximately two seconds before the end of the sentence, at which time dilation occurs. The three noise conditions show pupil constriction prior to pupil dilation.
Figure 7 Grand mean change in peak-to-peak pupil diameter. The tan area represents +/-1 standard error about the mean of the quiet responses. The brown, red, and dark red areas represent +/-1 standard error about the mean for the three noise conditions.
Figure 8 Change in peak-to-peak pupil diameter from baseline, collapsed across sessions. Significance differences exist across the quiet and -3 dB SNR, quiet and -5 dB SNR, 0 dB SNR and -3 dB SNR, and the 0 dB SNR and -5 dB SNR conditions. The error bars represent +/-1 standard error.
Figure 9 Grand mean change in reaction time (from the single-task only condition) for each of the four experimental conditions; the values are collapsed across both sessions. There was a significant difference between the quiet and the -3 and -5 dB SNR conditions. The error bars represent +/-1 standard error.
Figure 10 Change in reaction time from the baseline condition (single-task only) for both sessions. The error bars represent +/-1 standard error. Examining both sessions separately effectively illustrates the learning effect (overall improvement in reaction time) that occurred across sessions.
Figure 11 Ambient changes in room temperature. The ending temperature was always significantly warmer than the starting temperature. The dashed lines represent the ideal ambient temperature proposed by Vertungo et al., 2003.
Figure 12 Skin conductance response amplitudes for Participant 15, all conditions. Areas with darker shading represent a greater quantity of overlapping responses. Responses during the first session occurred in the anticipated windows, whereas in the second session very few responses occurred in the first analysis window.
Figure 13  Skin conductance response amplitudes for Participant 3, all conditions. Areas with darker shading represent a greater quantity of overlapping responses. Very few responses occurred during the first session. Responses were greater in quantity and amplitude during the second session. The occurrence of the responses appears asynchronous.
Figure 14 Skin conductance response amplitudes for Participant 12, all conditions. Areas with darker shading represent a greater quantity of overlapping responses. Few responses of small amplitude occurred during the first session. Responses were infrequent but greater in amplitude during the second session. The occurrence of the responses appears highly synchronous.
Figure 15 Grand mean skin conductance response amplitude as a function of SNR and collapsed across analysis window, Session I. The average amplitude of the responses during the 0 dB SNR condition are significantly smaller than the baseline condition and the mean amplitudes for the three remaining experimental conditions. The error bars represent +/-1 standard error.
Figure 16 Grand mean skin conductance response amplitude as a function of SNR and collapsed across analysis window, Session II. The average amplitude of the responses during the baseline condition are significantly smaller than the means for the four experimental conditions. The error bars represent +/-1 standard error.
Figure 17 Grand mean skin conductance response quantity as a function of testing condition. The bars on the left-hand side of each pair represent the quantity of SCRs from the first analysis window, while the bars on the right-hand side represent the quantity of SCRs from the second analysis window. The quantity of responses is significantly greater during second analysis window than the first analysis window and the baseline condition. The error bars represent +/- 1 standard error.
Figure 18 Grand mean skin conductance response amplitude based on sentence recognition, averaged across the 0, -3, and -5 dB SNR conditions. On average, the amplitude of a SCR occurring during an *incorrect* response was greater than the amplitude of a SCR occurring during a *correct* response. The error bars represent +/-1 standard error.
Figure 19 G Grand mean skin conductance response quantity based on sentence recognition, averaged across the 0, -3, and -5 dB SNR conditions. On average, the number of SCRs occurring during an incorrect response was greater than the number of SCRs occurring during a correct response. The error bars represent +/-1 standard error.
Figure 20 Correlation of SCR quantity and amplitude. Red symbols represent the first analysis window and blue symbols the second analysis window.
Figure 21 Grand mean (collapsed across sessions) normalized linear relationship of perceived listening effort during pupillometry. 100% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR).
Figure 22  Grand mean (collapsed across sessions) normalized linear relationship of perceived listening effort during dual-task testing. 100% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR).
Figure 23 Grand mean (collapsed across sessions) normalized linear relationship of perceived listening effort during skin conductance testing. 100% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR).
Figure 24 Grand mean (collapsed across sessions) normalized linear relationship of the change in peak-to-peak pupil diameter. The points within the shaded area represent a decrease in pupil diameter from the least to the most challenging situation, which was not anticipated. 75% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR). The grand mean subjective ratings for both conditions are provided for reference.
Figure 25 Grand mean (collapsed across session) normalized linear relationship of the change in reaction time from baseline. The points within the shaded area represent a decrease in reaction time from the least to the most challenging situation, which was not anticipated. 80% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR). The grand mean subjective ratings for both conditions are provided for reference.
Figure 26 Grand mean (collapsed across session) normalized linear relationship of the change in SCR amplitude. The points within the shaded area represent a decrease in SCR amplitude from the least to the most challenging situation, which was not anticipated. 50% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR). The grand mean subjective ratings for both conditions are provided for reference.
Figure 27  Grand mean (collapsed across session) normalized linear relationship of the change in SCR quantity. The points within the shaded area represent a decrease in pupil diameter from the least to most challenging situation, which was not anticipated. The three topmost lines are truncated to allow for better examination of the remaining data. 55% of participants displayed the expected trend; that is, a positive slope from the least challenging condition (quiet) to the most challenging condition (-5 dB SNR). The grand mean subjective ratings for both conditions are provided for reference.
Figure 28 Test-retest reliability. The most difficult condition (-5 dB SNR) was examined for test-retest reliability. Session I values are on the X-axis while Session II values are on the Y-axis. The dashed gray line indicates the theoretical values where Session I=Session II. The data for all paradigms are presented superimposed, without units. The top figure shows the reliability for the rating scales, pupillometry, dual-task, and SCR quantity. All four measures had significant reliability. The inset shows the reliability for the SCR amplitude. The reliability of the measure increased dramatically after removing one data point, which is contained within the red circle in the top-left portion of the figure.
CHAPTER V
DISCUSSION

In recent years, the topic of "listening effort" has gained considerable attention amongst auditory researchers; however, there has yet to be a clear definition of what is being measured and how to measure it. There are simple, subjective ways—such as asking a participant how difficult a listening task was—and there are objective ways—such as the dual-task paradigm, pupillometry, and skin conductance. At times, simple methods are not appropriate because some individuals (e.g., children, and some adults) have difficulty describing their experience. For this reason, it is useful to consider some of the many possible objective methods to assess listening effort. This study made within participant comparisons of one behavioral and two physiologic measures of listening effort: a dual-task paradigm, pupillometry, and skin conductance.

All of the participants were adults with normal hearing acuity (see Table 1 for demographic information). They listened to IEEE sentences in quiet and in continuous speech-shaped noise. The participants were instructed to repeat back as much of the sentence as possible while we varied the difficulty of the task by manipulating the SNR. We also asked each participant to subjectively rate their degree of listening effort on a scale from “0-100” following each block of 10 sentences. The rating scales were used to verify our assumptions about how SNR impacted perceived effort; namely, as the SNR decreased listening effort increased. We reported the results from each of the measures and also examined their intersession reliability. We described the group results as well as the individual trends for each measure. The following list is a summary of our findings:

1) Group mean speech perception performance decreased significantly as the SNR decreased.
2) Group mean subjective listening effort ratings increased significantly as the SNR decreased.

3) When listening effort was measured using the dual-task paradigm, group mean reaction time values for the dual-task increased as the SNR decreased, with a statistically significant difference between the quiet and the -3 and -5 dB SNR conditions. An unexpected but statistically significant overall decrease in reaction time was observed between the two sessions. We surmised that this difference was evidence of a practice effect.

4) When listening effort was measured using pupillometry, group mean measures of pupil dilation increased as the SNR decreased. Statistically significant differences were found between the quiet and -3 dB SNR, quiet and -5 dB SNR, 0 and -3 dB SNR conditions, and 0 and -5 dB SNR conditions. The resulted obtained using pupillometry were consistent across sessions.

5) When listening effort was measured using skin conductance, trends in the data showed that mean SCR amplitude and quantity values changed for most participants as the SNR decreased. During the first session, group mean SCR amplitude values for the 0 dB SNR condition was significantly smaller than the baseline and the other three conditions (quiet, -3 dB, and -5 dB SNR); during the second session, the baseline SCR amplitude value was significantly smaller than each of the test conditions (quiet, 0 dB, -3 dB, and -5 dB SNR).
6) Group mean SCR quantity during the second analysis window were significantly greater than the baseline condition and each of the test conditions (quiet, 0 dB, -3 dB, and -5 dB SNR).

7) Group mean SCR amplitude was larger for incorrectly understood sentences than correctly understood sentences.

8) Group mean SCR quantity was greater for incorrectly understood sentences than correctly understood sentences.

9) The individual trends were more varied. 100% of participants rated the -5 dB SNR condition as more effortful than the quiet condition. 80% of the participants’ reaction time during the dual-task increased from the quiet to the -5 dB SNR condition. 75% of the participants’ pupils dilated more during the -5 dB SNR condition than during the quiet condition. 50% of participants’ SCR amplitude was greater in the -5 dB SNR condition than the quiet condition. 55% of participants’ SCR quantity was greater in the -5 dB SNR condition than in the quiet condition.

10) Test-retest reliability was strong across many of the measures. Pupillometry exhibited the strongest reliability of the three objective measures.

Clearly, no single method was perfect and currently, all would be difficult to implement in a clinical setting. The following sections compare and contrast each of the findings in greater detail, and relate the results to current knowledge in the listening effort literature.
I Pupillometry

We hypothesized that pupil diameter would increase as SNR decreased and this hypothesis was supported by our findings. Previous research using speech-shaped noise (Zekveld et al. 2010; Koelewijn et al. 2012; Zekveld and Kramer 2014; Ohlenforst et al., 2017) showed that pupil diameter increases as SNR decreases. For example, Zekveld et al. (2010) found an average maximum pupil diameter change from baseline of .17 mm when speech perception abilities were driven to 50% correct. Our study found a similar mean peak-to-peak change in pupil diameter of .18 mm (Figure 8) at a -5 dB SNR; this condition resulted in an average speech perception score of 63% words correct (Figure 4). While small, the differences between the Zekveld study and the current study might be explained by how the results were quantified. We calculated peak-to-peak pupil diameter within a broad analysis window while Zekveld and colleagues found the maximum pupil diameter within a narrow analysis window; the difference between these analysis windows can be seen in Figure 6. The box consisting of dashed lines represents the analysis window for the peak-to-peak analysis employed in the current study while the box consisting of solid lines represents the more traditional maximum amplitude window (as in Zekveld et al., 2010 and Winn et al., 2015). Had we used this traditional analysis method, the pupil diameter values for the -3 and -5 dB SNR conditions would have been essentially the same. Further studies looking at various analysis techniques are needed.

Previous pupillometry studies have used a variety of techniques to make speech perception more challenging and increase listening effort. Zekveld et al.
(2010) used speech-weighted, stationary noise; Koelewijn et al. (2012) compared stationary noise, fluctuating noise, and single-talker maskers; Zekveld and Kramer (2014) presented sentences in single-talker babble; Ohlenforst et al., 2017 used both stationary noise and a single-talker masker; Winn et al. (2015) varied the number of channels in a vocoder simulation. All of these studies found that pupil diameter increased as listening effort increased; the same was the case for the present study (Figure 6, Figure 8). Additionally, four previous studies (Zekveld et al. 2010; Koelewijn et al. 2012; Zekveld and Kramer 2014; Ohlenforst et al., 2017) reported changes (from baseline) in maximum task-evoked pupil dilation ranging from 0.17-0.23 mm. But Winn et al. (2015), exhibited markedly different findings. They reported a change (from baseline) in maximum task-evoked pupil dilation of approximately .51 mm for a 4-channel vocoder. The average speech perception score during this condition was approximately 50% (sentences). Their findings were more than double the next largest finding (.23 mm by Ohlenforst et al., 2017) and nearly three times larger than the results from the current study (.18 mm, see Figure 8). We did not specifically compare different methods used to evoke changes in pupil diameter, but there seems to be something unique about the methods employed by Winn and colleagues. We posit that the way in which sentence intelligibility is manipulated may have a differential effect on pupil diameter. Vocoder simulations distort speech, but they do so in a way that is novel to most listeners. Many people are familiar with listening to speech in the presence of broad-band noise or babble, but few people are familiar with listening to vocoded sentences. Thus, the novelty of the vocoder task, combined with poor speech perception abilities, may have a synergistic
effect on changes in pupil diameter. It would be valuable for a future study to examine how these different methods may impact pupillometry measures.

Another methodologic difference between the present study and previous studies was how the noise was presented. The previous studies typically gated the noise on a number of seconds prior to the onset of the speech perception material. We played speech-shaped noise continuously throughout each block of 10 sentences (see Figure 2). This meant that the listener was constantly immersed in noise for several minutes at a time rather than having interludes of silence. Extended periods of low-level noise can have negative psychophysiological impacts on people, especially children (Evans et al., 1998). It stands to reason that extended periods of noise may also impact changes in pupil diameter. We came to similar conclusions as the previous studies: poorer SNRs result in increased listening effort which in turn result in greater changes in pupil diameter. It seems that how the noise is presented—gated versus continuous—may not significantly impact pupil dilation, but it would be helpful for another study to examine the duration of noise exposure on the task-evoked pupil dilation more closely.

To date, only one other study (Ohlenforst et al., 2017) has used fixed SNRs, rather than fixed performance, to evaluate listening effort. Our findings mirror those of Ohlenforst and colleagues: poorer SNRs increase listening effort and this effort is reflected physiologically as an increase in pupil diameter. This is relevant because fixed SNRs—due to their speed and universal applicability—are faster to administer than an individualized metric such as the SNR50. Pupillometry measures using fixed SNRs may save time over more individualized measures, but more research will be
needed to better understand the tradeoffs between using a fixed performance versus a fixed condition. Furthermore, it is unlikely that the same SNRs will apply to people with hearing loss—specifically, cochlear implant users—so similar studies will also need to be conducted in a variety of clinical populations.

To our knowledge, no other study has examined the validity of pupillometry as a means to assess listening effort on an *individual* basis. *A priori*, we expected to see individual pupil diameter increase (exhibit a positive slope) from the quiet condition to the -5 dB SNR condition. We normalized all of the participants’ values in order to place them on a similar scale. Altogether, 75% of participants displayed the expected trend (Figure 25). Furthermore, pupillometry was the most reliable of the three objective measures during the most difficult SNR with a Pearson correlation coefficient of 0.72 (Figure 28), indicating that the individual measures were consistent across the two sessions for this condition. Certainly, there is room for improvement but with further refinements to both study design and analysis techniques, pupillometry may soon gain traction as an individual assessment of listening effort in the clinic.

In summary, many studies have shown that pupillometry consistently demonstrates changes in listening effort. Results are relatively simple to obtain and the measurement has a good degree of validity and reliability, potentially making it useful as an individual assessment of listening effort. The current study also demonstrated that pupillometry is illustrative of individual trends in listening effort. Despite these promising findings, pupillometry still has several disadvantages. Perhaps the largest barrier is the light reflex and the strict control of the


environmental lighting that is required to accurately assess task related changes in pupil diameter. The participants from the current study were confined to a chin rest and were clearly limited in their ability to adjust their position—this is a significant weakness, particularly for children and adults with mobility concerns. There is also a continued concern about cost of the equipment, which is valid, as well as a concern about how to analyze and interpret the data. However, both of these concerns will become diminished as pupillometry becomes more widely studied and the technologic and methodologic hurdles are overcome.

II Dual-Task

We hypothesized that secondary task performance (reaction time via a cued button press) would increase as listening effort increased. Previous studies have also used a dual-task paradigm and showed that increased listening effort results in a decline in secondary task performance (Sarampalis et al. 2009; Hornsby, 2013; Picou et al., 2013; Wu et al., 2016). As in previous studies, we found that reaction time gradually increased as SNR decreased, with significant differences in reaction time occurring between the quiet and the -3 and -5 dB SNR conditions (Figure 9). Unique to the current study was the fact that we used fixed SNRs, rather than fixed performance (e.g., SNR50) and allotted fewer practice trials to the dual-task paradigm than may be necessary. Some studies, such as that by Wu et al., 2016, implemented a minimum of 60 practice trials before beginning the actual trials. Other researchers were not as detailed about their practice sessions, simply stating that the participants “required extensive practice sessions” (Hornsby, 2013; Picou et al., 2013). One study by Sarampalis et al. (2009), was even more nebulous with their practice guidelines,
stating that they had participants practice "until they reported being comfortable with" the dual-task. We allotted 20 practice trials for the single-task alone and an additional 10 practice trials for the dual-task. This was done to keep the number of practice trials for the dual-task commensurate with the practice trials for the physiologic paradigms. This meant that participants in the current study may have received far fewer practice trials than participants from other studies. We found a significant overall decrease in reaction time from the first session to the second session (Figure 10), which has been interpreted as an overall practice effect. Currently, there are no clear recommendations regarding the appropriate number of practice trials for a dual-task listening effort study. Clearly, a study that develops standardized practice procedures would be of value.

The choice of the secondary task measure in a dual-task paradigm is also important. Previous studies have used word recall (Hornsby et al., 2013), performance in a driving simulator (Wu et al., 2014), and a simple, cued button press (Piccou et al., 2011, 2013; Wu et al., 2016); in the present study, we employed a simple cued button press. This was done in an effort to select a secondary task that was in a distinct modality from the primary task, thus minimizing any unwanted confounding effect on speech perception abilities (Ninio and Kahneman, 1974). Studies that employed a secondary task that was too similar to the primary task (as in Hornsby et al., 2013) may have inadvertently confounded their results. A recent study by Wu et al. (2016) illustrated exactly this point: a difficult secondary task that incorporates more linguistic processing demands (i.e. the Stroop task), resulted in greater (i.e. poorer) reaction times than a simple secondary task (i.e. cued button press). What is needed is a study that looks at the
effect of several secondary tasks on the same primary task and makes recommendations accordingly.

The dual-task paradigm may be a valuable tool to assess individual trends in listening effort. Despite the overall decrease in secondary task performance (learning effect) from the first session to the second session, 80% of participants’ reaction time was poorer during the -5 dB SNR condition than the quiet condition. Furthermore, the dual-task paradigm exhibited strong test-retest reliability for the most difficult SNR with a Pearson correlation coefficient of 0.62 (Figure 28), indicating that the individual measures were consistent across the two sessions for this condition. This, combined with its straightforward implementation, indicates that the dual-task may have value as a clinical assessment, as long as concise methodologic recommendations are first established.

To summarize, the dual-task has been an important tool for assessing listening effort. It is inexpensive, easy to administer, and is reliable as an individual metric. However, it was not as sensitive to as many contrasts as pupillometry (Figure 8 versus Figure 9) and appeared to be subject to a practice effect. Clearly, more research is needed to better understand practice effects, as well as the most appropriate form of secondary task, before the dual-task can be implemented in more broad-reaching settings, such as the clinic.

III Skin Conductance

We hypothesized that skin conductance response amplitude and quantity would increase as listening effort increased. Previous studies have quantified listening effort by analyzing tonic changes to skin conductance, known as the skin conductance
level (Mackersie and Cones, 2011; Mackersie et al., 2015). This paper is unique in that we used skin conductance responses (SCRs), a phasic measure of autonomic nervous system activity, to assess changes in listening effort. Skin conductance level was not analyzed in the present work because it is difficult to quantify and can obscure the SCRs, which are smaller changes related to the onset or offset of a stimulus (Roth et al., 2012). We successfully demonstrated that such responses can be obtained in most participants while they were engaged in a listening effort task. Generally, participants showed larger SCR amplitudes when they misunderstood a sentence (Figure 18); the number of SCRs were also increased when they misunderstood a sentence (Figure 19). SCR was not a sensitive metric of listening effort during this study, but due to the fact that it has not yet received much attention in the listening effort literature, the following discussion will address our findings at greater length.

We hypothesized that SCR amplitude would increase as the SNR decreased. We found that during the first session, the SCR amplitude for the 0 dB SNR condition was significantly lower than the baseline amplitude and the amplitude of the three remaining conditions (quiet, -3 dB, and -5 dB SNR) (Figure 15). The amplitude of the 0 dB SNR condition, but none of the other test conditions, significantly increased from the first session to the second session (Figure 15, Figure 16). It is unlikely that this finding is related to an order effect, as all SNRs were counterbalanced, but ultimately the implications for this finding are unclear.

The baseline SCR amplitude (recorded during the no-stimulus trial) obtained during the second session was significantly smaller than the SCR amplitude for all
four of the experimental SNRs (Figure 16). Rather than offer any insight into listening effort, this finding may simply suggest that the participants were more relaxed, and thus less aroused, during the no-stimulus trial of the second session than they were during the four experimental conditions.

We also found that the average SCR amplitude was larger during the second analysis window (after the response prompt) than during the first analysis window. A potential concern regarding these findings is that these responses may reflect vocalization artifacts, not listening effort. While we did not specifically test for such artifacts, the effects of vocalization artifacts were controlled for by our study design. For one, it would be unusual for muscle movements from the head and neck area to contaminate a distal, palmar electrode recording site. Moreover, the presence of SCRs is reflective of central nervous system activity—not motor activity—and displays distinct traits (Roth et al., 2012). Motor activity, say from an individual clenching their fingers, results in a rapid ascent to a sharp peak, followed by a rapid decent when the individual unclenches their figures. These artifacts were specifically controlled for in the design of the study (by requesting the participants to refrain from movement) and during the analysis (by examining the rise time of the response—see page 21 for details), making them an unlikely contributor to any increases in SCR amplitude or quantity. In the future, these concerns could be better addressed by assessing respiration rate, muscle contraction, and skin conductance simultaneously. By doing so, any contamination of the skin conductance record could be removed by subtracting the unwanted responses from the skin conductance responses.
Ambient room temperature may have also impacted our SCR observations. Cooler temperatures in the sound booth may have resulted in a premature drying of the corneum (Boucsein, 2012; Roth et al., 2012), resulting in fast absorption of the conductive paste. In turn, this may have led to smaller amplitude SCRs. In contrast, high temperatures may have increased thermoregulatory sweating and resulted in greater perspiration, lower skin resistance, and increased SCR amplitudes. Our findings indicated that the testing environment was statistically warmer at the end of the session than the start of the session; however, these temperatures were still within the recommended range (Vertungo et al., 2008). A study in an environment where the temperature can be held constant would be able to address the impact of ambient temperature on listening effort. Much like EMG and respiratory measurements, the addition of concurrent skin temperature measurements could also help statistically control for the potential confounding effect of ambient room temperature.

Another factor that may have impacted our SCR amplitude findings are external factors such as: sleep deprivation, prescription medicine, caffeine intake, stress level, or some other unknown element. Any of these factors can affect an individual’s state of arousal, which can in turn either positively or negatively impact SCR amplitude (Boucsein, 2012, Roth et al., 2012; Fowles et al., 1981). While a researcher cannot ethically request a participant refrain from prescription medicine, greater care could be taken to document each participants’ medicine routine. One could also improve upon our findings by requesting that participants refrain from, or state their consumption of, other substances known to affect skin conductance, such as alcohol or caffeine, for 24 hours prior to testing.
When we first analyzed SCR amplitude, correct and incorrect speech perception results were grouped together, and our findings were largely non-significant. We then reanalyzed our findings but separated the SCR tracings for correct speech perception responses from SCR tracings for incorrect speech perception responses. After doing so, we found that on average the SCRs during incorrect speech recognition performance were significantly larger than the SCRs during correct speech recognition performance (Figure 18); we also demonstrated that SCR quantity for incorrect speech recognition is greater than for correct speech recognition (Figure 19). These findings suggest that for participants who have normal hearing, SCR amplitude and quantity may better reflect the uncertainty surrounding their speech perception abilities. If so, such measures may be useful to implement within a pediatric cohort, whose reliability is not always consistent. If our findings hold, SCR quantity for incorrectly perceived sentences should increase systematically as the task is made more difficult; we would also expect to see larger amplitude responses during incorrectly perceived sentences. A departure from this trend may suggest that the participants’ abilities (or inabilities) may be tied to some other factor, such as inattention, rather than strictly their perception abilities.

Skin conductance was highly variable and thus not a reliable indicator of listening effort on an individual basis (Figures 26 and 27). Despite these findings, the test-retest reliability was fair during the most difficult condition for SCR quantity with a Pearson correlation coefficient of 0.45 (Figure 28). This indicates that the individual SCR quantity measures were mostly consistent across the two sessions for this condition. On the other hand, the test-retest reliability for SCR amplitude was
poor during the most difficult condition ($r=0.19$; Figure 28-inset, black line and symbols) but the reliability quickly became significant after the largest outlier was removed ($r=0.62$; Figure 28-inset, gray line and symbols).

In summary, our results for SCR amplitude and SCR quantity during a challenging listening task were sensitive to the contrasts between correctly and incorrectly understood sentences, but they were not sensitive to changes in listening effort. Despite the relative ease of administration and potential portability of skin conductance measures, its use solely as an indicator of listening effort are not recommended.

### IV Comparisons Of The Proposed Measures

A primary goal of this study was to assess several techniques that have been suggested as viable means for measuring listening effort. We hypothesized that all of the tools would reflect changes in listening effort, but some were more effective than others. It seems clear from the current findings that if the individual is able, simply asking them about their perceived effort is the most appropriate tool. This assertion is supported by the fact that as a group, perceived effort increased significantly as the task was made more challenging. Moreover, 100% of individuals rated the most difficult condition as requiring more effort than the easiest condition and the test-retest reliability of the subjective scales during the most challenging condition was high ($r=0.79$, Figure 28). However, a subjective rating scale does not lend insight into the physiologic processes involved in effortful listening, and as such remains contraindicated in certain clinical populations.
Of the objective measures assessed during this study, pupillometry was the best tool for assessing listening effort. It can be used in the absence of a reliable participant, as well as if there is a strong desire to understand the physiology behind listening effort. The sensitivity of pupillometry is greater than the dual-task paradigm at the group level (Figure 8 and Figure 9), had good sensitivity on an individual basis (Figure 24 and Figure 25), and exhibited strong test-retest reliability during the most challenging SNR (Figure 28). Pupillometry also requires little to no training of the participant, which is an obvious advantage in a clinical setting.

Yet another advantage of pupillometry is its potential future use in real-world listening environments such as restaurants. In fact, there are already companies that have designed compact eyeglass systems to track eye gaze in order to learn more about human behavior (e.g., marketing). Smartphones have also become more sophisticated, and it seems plausible that eye tracking will someday be a realizable add-on feature for researchers who use smartphones. In fact, pupillometry would function well paired with a subjective measure such as the smartphone assessments employed by Wu et al. (2015). Such a combination would be valuable to researchers as they would be able to synchronize physiologic and subjective assessments, as well as audio recordings, to a particular real-world event and location—all recorded by the smartphone. This may give researchers the ability to obtain a more complete picture of effortful listening as it unfolds in real-time. However, the major limiting factor to this becoming a reality is the fact that as of yet, there is no way to extract task-evoked pupil responses from the much larger changes due to the light reflex. Until the light
reflex can be controlled, either through signal processing or statistically, pupillometry is limited to a well-controlled environment, such as a laboratory or clinical setting.

The dual-task paradigm with a cued button press as the secondary task, while not as sensitive as pupillometry, is still a recommended tool for assessing listening effort. The design is simple, inexpensive, and has good validity on an individual basis with 80% of participants exhibiting increased reaction time from the easiest task to the most difficult task. Two significant weaknesses of the dual-task are that it is only sensitive to large contrasts in listening effort (Figure 9) and it is susceptible to practice effects (Figure 10). Additionally, it is time-intensive, there is no clear recommendation as to the most appropriate secondary task, and much like the subjective scales, there are clinical populations that will not be able to perform a dual-task.

In the present study, SCR amplitude and quantity were not sensitive to discrete changes in listening effort at either the group or individual level. Thus, their utility as a viable measure of listening effort is called into question. However, we did find that SCR amplitude was larger and SCR quantity was greater for sentences that were incorrectly understood than for sentences that were correctly understood (Figure 18 and Figure 19). On a physiologic level, these findings suggest a heightened sense of arousal during incorrect understanding than correct understanding. Thus, SCRs appear to be a physiologic reflection of performance. Such a tool may more fully reflect speech understanding in a group of young children or other individuals where fatigue and lack of participation are a concern. Admittedly, there is a substantial void in the
literature that needs to be filled before this application can be realized, but it is intriguing nonetheless.

V Conclusions

Listening in noisy environments is challenging and requires greater effort than listening in quiet. As the SNR decreases, a person must exert more effort to overcome the acoustic deficits produced by the noise if they wish to maintain a high degree of speech understanding. This change in effort can be measured behaviorally and physiologically. The present study demonstrated that pupillometry and a dual-task paradigm were valid and reliable measures of listening effort; of these two paradigms, pupillometry was the most sensitive to discrete changes in listening effort. Skin conductance responses were also measured, but the amplitude and quantity of the responses were not sensitive to changes in listening effort. However, the quantity and the amplitude of responses measured increased when the participants repeated the sentences incorrectly compared to sentences that were repeated correctly. Taken together, these findings suggest that a clinical or other real-world application of listening effort deserves further attention.
REFERENCES


