ACTIVE NOISE REDUCTION VERSUS PASSIVE DESIGNS IN COMMUNICATION HEADSETS: SPEECH INTELLIGIBILITY AND PILOT PERFORMANCE EFFECTS IN AN INSTRUMENT FLIGHT SIMULATION

by

R. Brian Valimont

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Industrial and Systems Engineering

Advisory Committee:
Dr. John G. Casali, Chairman
Dr. Brian M. Kleiner, Member
Dr. Jeff A. Lancaster, Member
Dr. Thurmon E. Lockhart, Member
Dr. Antonio A. Trani, Member

April 20, 2006
Blacksburg, Virginia

Keywords: Active Noise Reduction, Speech Intelligibility, Pilot, Flight Simulation

ACTIVE NOISE REDUCTION VERSUS PASSIVE DESIGNS IN COMMUNICATION HEADSETS: SPEECH INTELLIGIBILITY AND PILOT PERFORMANCE EFFECTS IN AN INSTRUMENT FLIGHT SIMULATION

by

R. Brian Valimont

Chairman: Dr. John G. Casali

Industrial and Systems Engineering

(ABSTRACT)

Researchers have long known that general aviation (GA) aircraft exhibit some of the most intense and potentially damaging sound environments to a pilot's hearing. Yet, another potentially more ominous result of this noise-intense environment is the masking of the radio communications. Radio communications must remain intelligible, as they are imperative to the safe and efficient functioning of the airspace, especially the airspace surrounding our busiest airports, Class B and Class C. However, the high amplitude, low frequency noise dominating the GA cockpit causes an upward spreading of masking with such inference that it renders radio communications almost totally unintelligible, unless the pilot is wearing a communications headset. Even with a headset, some researchers have stated that the noise and masking effects overcome the headset performance and still threaten the pilot's hearing and overall safety while in the aircraft.

In reaction to this situation, this experiment sought to investigate the effects which active noise reduction (ANR) headsets have on the permissible exposure levels (PELs), speech intelligibility, workload, and ultimately the pilot's performance inside the cockpit. Eight instrument-rated pilot participants flew through different flight tasks of varying levels and types of workload embedded in four 3.5 hour flight scenarios while wearing four different headsets. The 3.5 hours were considered long duration due the instrument conditions, severe weather conditions, difficult flight tasks, and the fatiguing effects of a high intensity noise environment. The noise intensity and spectrum in the simulator facility were specifically calibrated to mimic those of a Cessna 172. Speech intelligibility of radio communications was modified using the Speech Transmission Index (STI), while measures of flight performance and workload were collected to examine any relationships between workload, speech intelligibility, performance, and type of headset.

It is believed that the low frequency attenuation advantages afforded by the ANR headset decreased the signal-to-noise ratio, thereby increasing speech intelligibility for the pilot. This increase may positively affect workload and flight performance. Estimates of subjective preference and comfort were also collected and analyzed for relevant relationships.

The results of the experiment supported the above hypotheses. It was found that headsets which incorporate ANR technology do increase speech intelligibility which has a direct inverse influence on workload. For example, an increase in speech intelligibility is seen with a concomitant decrease in pilot workload across all types and levels of workload. Furthermore, flight task performance results show that the pilot’s headset can facilitate safer flight performance. However, the factors that influence performance are more numerous and complex than those that affect speech intelligibility or workload. Factors such as the operational performance of the communications system in the headset, in addition to the ANR technology, were determined to be highly influential factors in pilot performance.

This study has concluded that the pilot’s headset has received much research and design attention as a noise attenuation device. However, it has been almost completely overlooked as a tool which could be used to facilitate the safety and performance of a general aviation flight. More research should focus on identifying and optimizing the headset components which contribute most to the results demonstrated in this experiment. The pilot’s headset is a component of the aviation system which could economically improve the safety of the entire system.

ACKNOWLEDGEMENTS

I'd like to first express my sincere gratitude to my advisor, Dr. John Casali, for his knowledge, expertise, encouragement, and above all, his guidance through the research project, interaction with the research sponsor, and dissertation process. I consider myself extremely fortunate to have worked on such a great research project with a professional and teacher of Dr. Casali's caliber. The value of this learning experience has been immeasurable.

Another great contributor to this learning experience has been Dr. Jeff Lancaster. I've had the privilege to work with Dr. Lancaster on a daily basis as he was the supervisor of my research project. He provided many answers to my daily questions, in addition to guidance regarding the best direction to consider during my times of uncertainty.

I'd also like to thank the rest of my committee members, Dr. Brian Kleiner, Dr. Thurmon Lockhart, and Dr. Antonio Trani. Their assistance and guidance through the preliminary exams, dissertation proposal, and final dissertation defense, have contributed greatly to the rich learning experience I've derived from this Ph.D. program.

Lastly, but by no means least, I want to thank my wife for her patience and understanding. She sacrificed her new husband two weeks after our wedding and stood by me as I focused my time and energy on the Ph.D. program for the past 3 years. Also to the friends I've made in this program, Navrag, Jeff, Myrna, Linsey, Young Seok, and Ravi, you made the process bearable, and more often, fun. I'll miss you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The aviation system, as a whole, is arguably the largest and most complex system ever devised by mankind, and the flight environment is one of the most dynamic and potentially perilous areas in which a civilian is able to voluntarily be an active and willing participant. The chief component that keeps this system from self-destructing and killing large quantities of travelers is the dissemination of time-critical information through voice communications generally over a designated radio frequency. Speech communications are especially important for teams whose members are physically separated making visual contact impossible, such as the air traffic control (ATC) – aircraft pilot team (Whitaker & Peters, 1993). However, when these communications break down, the results can be catastrophic. Problems in communications, in various forms, have contributed directly and indirectly to some of the costliest aviation disasters recorded (Strother, 1999).

In fact, the worst accident in aviation history was the result of a misinterpreted radio transmission, and a subsequent unintelligible transmission. These simple, common communications errors led to the death of 538 passengers and crewmembers abroad two Boeing 747s, as follows.

The field at Tenerife, Canary Islands, on March 27, 1977, was socked in with thick fog, dropping runway visibility range to less than a quarter of a mile, which permitted only departing airliners to use the active runway. KLM flight 4805 was instructed to backtaxi the active runway, make a 180 degree turn and hold their position awaiting take-off clearance. Meanwhile, Pan Am flight 1736 was cleared to backtaxi the active runway until they reached one of the last runway turn-offs. There, Pan Am 1736 was to exit the runway to allow room for KLM 4805 to initiate its take-off roll. While Pan Am 1736 was backtaxiing on the active runway, the air traffic controller issued KLM only its departure clearance, which KLM correctly readback. The controller then transmitted an additional statement, "Stand by for take-off, I will call you." Tragically, this statement was garbled and presumably unintelligible to the KLM pilots, whom did not reply to the command and most likely believed they were already cleared for take off. Instead of a pilot readback to the previous controller command, the ATC audiotapes picked up the squeal of tires as the KLM Boeing 747 released its brakes and began lumbering towards the Pan Am 747 just approaching their taxiway turn-off. Twenty seconds later, the KLM 747 slammed into the Pan Am 747. The resulting impact forces and conflagration claimed the lives of all crewmembers and passengers save approximately two crewmembers and fifty passengers on the Pan Am 747. All occupants of the KLM 747 perished (Aviation-Safety.net, 1996).

The Tenerife accident and others like it have become some of the most studied crashes in aviation as classic cases of human factors and communications breakdown. Unfortunately, many more general aviation accidents and incidents presently are the result of poor communications and equipment. In 1996, the results of one investigation led to the disturbing conclusion that 78% of reported Emergency Medical Service aviation incidents were attributed to communications difficulties (Connell & Reynard, 1993). Anecdotally, if asked, any pilot would agree, if you fly long enough you will see that poor communications lead to missed call signs, landing on the wrong runway, entering the traffic pattern at the wrong location, and even in some cases the intelligibility of a transmission is of such poor quality that the air traffic controller might use a second aircraft to relay the message in hopes that this aircraft, possibly being closer, will have a better chance of understanding the pilot’s transmission. Noise attenuating technology, such as the active noise reduction communications headset, could potentially facilitate smoother communications, but has yet to be tested in a laboratory environment during simulated flight to see what benefits it may hold for the safety and performance of GA pilots.

HUMAN AUDITION IN THE AVIATION ENVIRONMENT

Physiology of the Human Ear

The human ear has three main purposes carried out in stages in three distinct areas of the ear. The first objective, accomplished by the outer ear, is the modification of an acoustical disturbance which reaches the ear (Henderson & Hamernik, 1995). The pinna, the visible, convoluted, cartilage portion of the ear, is shaped such that certain frequencies are amplified while others are attenuated. Each individual’s pinna is shaped differently giving sounds an “earprint” that is distinctly unique to each individual. Just past the pinna begins the external auditory canal. This structure also modifies the acoustic wave by amplifying those frequencies between 2000 – 4000 Hz, increasing the amplitude by approximately 10-15 dB (Goldstein, 1989). This frequency range is not surprisingly part of the critical bandwidth for human speech, and is also associated with the range that is most susceptible to noise-induced hearing loss (Ward, Royster, & Royster, 2000a).

The second objective is the translation of the airborne acoustical wave into mechanical energy (Henderson & Hamernik, 1995). This process is begun by the pressure fluctuations of the sound wave vibrating the tympanic membrane, more commonly known as the eardrum, of the middle ear. The eardrum transmits the energy through a mechanical chain of small bones, called ossicles. The ossicles are comprised of three tiny bones named the malleus, incus, and stapes. These bones assist the middle ear in its primary function, the amplification of the sound energy before it reaches the entrance of the inner ear. The ossicles assist by providing mechanical advantage in the form of a lever action as the stapes pushes against the oval window, the entrance to the inner ear. The middle ear also employs another concurrent method of amplification. The tympanic membrane is 17 times larger than the oval window, which is transmitted as a proportional increase in the sound pressure as the energy moves from the middle ear to the inner ear. If these functions were not carried out by the middle ear, only 1/1000th of the acoustic energy would be translated from air pressure to fluid pressure. This is approximately a 30 dB loss in sound pressure (Ward, 1986).

The middle ear also protects the inner ear from loud vocalization and impulse noises. It accomplishes this task through the contraction of two muscles in the middle ear, the tensor tympani and the tensor stapedius. These muscles will contract and lock the ossicles together, which will not only prevent the lever action amplification, but alternatively will attenuate the incoming acoustic energy. However, there is an unfortunate small lag in the muscle contractions which renders this method of attenuation as partially ineffective for sudden impulses, such as gunfire. The last function of the middle ear is associated with the Eustachian tube which connects the middle ear to the nasal passages in order to equalize the pressure between the enclosed middle ear and the outside ambient pressure (Ward, Royster, & Royster, 2000a).

The third objective, accomplished by the inner ear, transforms the mechanical movement of the oval window, pressed upon by the stapes, into nerve impulses, which then travel via the auditory nerve to the brain for processing and interpretation (Ward, Royster, & Royster, 2000a). The inner ear consists of two major components, the cochlea, and the auditory nerve, both embedded in the temporal bone of the skull. The energy transformation begins when the stapes pushes on the oval window of the inner ear. This movement stirs the perilymph fluid in the cochlea and causes the basilar membrane to move up and down. Situated on top of the basilar membrane is the Organ of Corti, which houses the hair cells and stereocillia. As the basilar membrane pitches upward and downward, the hair cells and cilia are bent as they are pressed against the tectorial membrane, positioned directly above the Organ of Corti. The deformation of the hair cells causes an electrical nerve impulse to be generated and sent to the brain via the auditory nerve (Goldstein, 1989).

Specific frequencies of sounds are transduced by the inner ear into nerve impulses using two methods. The first method involves the synchronization of the firing of the neural impulses with the frequency stimulating the inner ear. For instance, a 200 Hz tone will cause the neurons to fire 200 times per second. This method is effective only up to the firing limitations of the neurons at approximately 4000 Hz (Ward, Royster, & Royster, 2000a). Above 4000Hz the neurons are unable to fire fast enough to accurately transduce the sound frequency. Therefore, the second method must be relied upon to accurately represent sound frequencies as neural impulses. The neural firing will be localized along the basilar membrane with high frequencies affecting the base of the basilar membrane and the low frequencies affecting the apex. Some sounds within a certain frequency bandwidth (200Hz – 4000Hz) will employ both methods in the frequency transduction (Ward, Royster, & Royster, 2000a). Studying the functions of the human ear makes it quite obvious that the ear is not merely a simple frequency analyzer, but a highly tuned, specialized nonlinear filter which amplifies certain frequency ranges, such as the range of human speech, while attenuating other ranges, such as low frequency sounds (Nobili et al., 1998).

Sensory Neural Hearing Loss

When the structures of the ear become deformed or permanently damaged, the most likely result is a form of hearing loss. There are actually two types of hearing loss, conductive hearing loss, which describes the physical damage of the structures in place in the middle ear, or less commonly, a severe blockage of the external auditory canal in the outer ear. However, this type of hearing loss is quite a rare result of overexposure to extremely high levels of impulsive environmental noise, and therefore will not be covered in further detail in this report (For further detailed coverage of conductive hearing loss, the reader should refer to Newby, 1979).

The second type of hearing loss, sensory neural hearing loss, is the result of overexposure to high levels of any acoustical energy, which may result from two common scenarios. The first scenario involves an impulsive acoustic disturbance which passes through the middle ear before the muscles can contract, and reaches the inner ear with such force that the hair cells are literally broken off at their base. Without these hair cells, neural impulses cannot be generated and sent to the brain. The second scenario involves chronic overexposure to levels of noise not nearly as high as the first scenario, but still high enough to bend the hairs for long periods of time, fatiguing the hair cells. This fatigue from overexposure leads to symptoms such as tinnitus and temporary threshold shifts. If these structures are not given sufficient time to recover, the hair cells will continue to degenerate and the temporarily elevated hearing thresholds will become permanent (Ward, 1986).

Noise-induced hearing loss (NIHL) is often diagnosed using a pure tone audiogram. The audiogram will typically show a distinguishing elevated threshold at 4000 Hz. This classic “notch” in the audiogram is indicative of noise-induced hearing loss, and as the condition continues and worsens, the notch will usually deepen at 4000 Hz and spread to include elevated thresholds at neighboring frequencies (Ward, Royster, & Royster, 2000b).

Pilot Hearing Loss and Cockpit Noise Hazards

Next to vision, audition is arguably the most important sense used during flight. The untapped potential for transferring information, especially time critical information, through the auditory modality is just beginning to be realized (Wagstaff, Tvete, & Ludvigsen, 1999). However, the aviation environment is a perilous environment for a pilot’s hearing. Studies have suggested that aircraft are the source of permanent threshold shifts for pilots regularly exposed to the cockpit noise levels (Tobias, 1972; Gasaway, 1990). One study even stated that aviation communications headsets do not provide adequate protection especially for those pilots, such as flight instructors who typically fly for 10 - 14 hours per day, 6 days per week, weather permitting (Tobias, 1972). Evidence of hearing loss in flight instructors is given by Tobias (1972b), where it was found that 85% of flight instructors and charter pilots have mild to moderate threshold shifts at high frequency ranges. These threshold shift measurements averaged 25 – 60 dB hearing threshold level (HTL). FAA flight inspectors and agricultural pilots fared worse with threshold shifts measuring 40 – 60 dB and 30 – 70 dB, respectively (Tobias, 1968a).

The largest operating category of aircraft, not including scheduled airlines, is the single-engine, light aircraft, such as a Cessna 172 (FAA, 1979). In these aircraft, noise originating from the engine, propeller, muffler, and slipstream, peak well above 105 dB (Figure 1), with the maximum energy concentrated in the general low frequency range of 50-250 Hz. (Tobias, 1968b). At these excessive levels, someone would experience irreversible damage to their hearing with unprotected exposures of as little as 3 – 4 hours per week (Tobias, 1968a).

This environment can often cause critical situations where noise is intensified and communications are essential to maintain safety, such as takeoff with full throttle where tower transmissions must be heard to avoid other air traffic (e.g. KLM 747 accident at Tenerife; Tobias, 1972a). Episodes such as this, set in the takeoff or landing flight phases, may be even more destructive to the pilot's hearing than the cruise phase of flight where most audiological investigative measurements have been taken (Gasaway, 1986). The likelihood of auditory injury is higher because the ambient noise level is increased due to the need for maximum horsepower during takeoff from the engine. Therefore, the pilot must raise the volume on his or her communication headset to overcome the noise levels and comprehend the incoming radio messages. The increased volume under headsets can easily become hazardous and is considered to be a potential threat to the pilot's hearing (Gasaway, 1986).

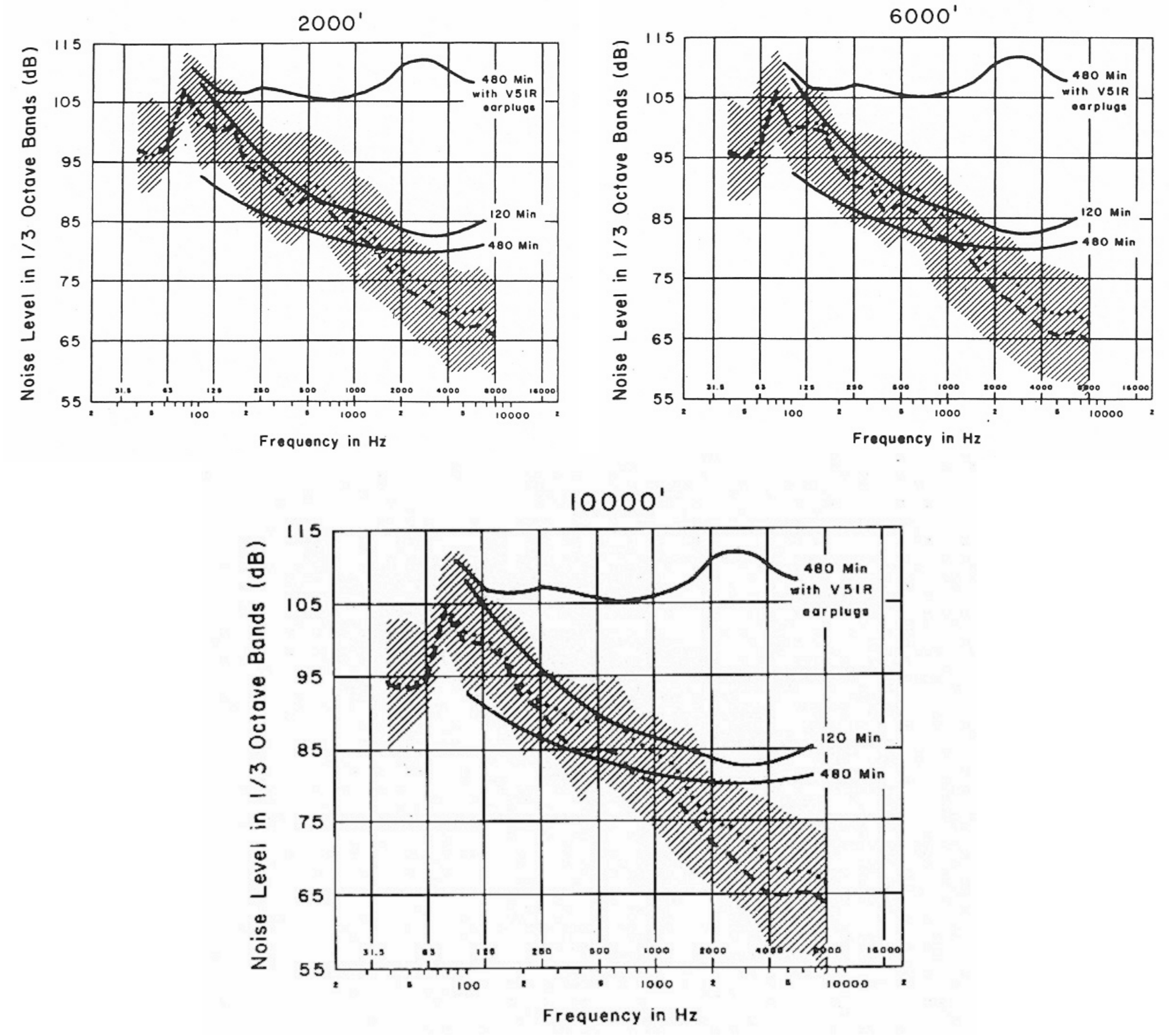


Figure 1. Cockpit spectral noise levels measured in GA single engine aircraft, at 3 different altitudes (from Tobias, 1968a).

These hazardous sound levels also create an additional problem. The ambient noise levels exist at frequencies which passive aviation communication headsets do not effectively attenuate. Low frequency noise passes quite easily through the construction materials of the headset and the bones of the skull, while higher speaker volumes set by the pilot on the headset increase the sound levels in the frequency range (speech) where the ear structures are most vulnerable to damage (Gasaway, 1986).

The Masking Effect of Noise

Although high noise levels and pilot hearing loss is a real problem, pilot NIHL is not as dangerous as the communication difficulties caused by the masking effect of high noise levels on the speech carried through radio communications. Masking is defined as the increase in the threshold of audibility of one sound, the masked sound (in this case, speech), caused by the presence of another sound, the masking sound (in this case, ambient noise; Robinson & Casali, 2000). The adverse effects of masking rears its ugly head as the significant increase in masked threshold for human speech. A masked threshold is the sound pressure level (SPL) at which an auditory stimulus, such as speech, is just audible 50% of the time in the presence of a specified type of noise. Typical masking sounds in the cockpit emitted by the engine, propeller, or slipstream will increase the masked threshold for a pilot well above their normal threshold of audibility.

To guard against the adverse effects of masking, engineers generally try to reduce the background noise level. If this is neither feasible nor practical, then a common trick is to change the spectrum of the signal to contrast with the noise and thereby increase the likelihood of signal detection (Robinson & Casali, 2000). However, when dealing with speech in intense low frequency noise, the problem is more complex. Masking effects are further exacerbated for a pilot in the cockpit of a light general aviation aircraft due to the very physical nature of the high intensity, low frequency noise. These high intensities cause the upward spread of the masking effect, where the masking effect spreads out above and below the frequencies of the masking noise, with the masking effect being greater at the frequencies above the noise rather than below (Figure 2; Pickett, 1959). The upward spread of masking is the force behind a low frequency noise's ability to mask the critical speech intelligibility bandwidth of 600 – 4000 Hz (Robinson & Casali, 2000; Stevens, Miller, & Truscott, 1946). The high sound levels are also theorized to physiologically overload the auditory structures of the ear resulting in what is known as cochlear distortion (Stevens, Miller, & Truscott, 1946). Cochlear distortion occurs when very high levels of noise overload the cochlea to a point where the cochlea is no longer able to accurately transduce the acoustical energy reaching it (Robinson & Casali, 2000). In the operational setting, this translates into the engine noise's ability to easily mask speech in the form of radio communications, leaving the radio communications unintelligible, if even audible.

Cognitive and Physiological Effects of Noise

It has been demonstrated that noise irrelevant to the task being performed by an individual not only creates a distracting effect, but at high enough levels, can lead to the

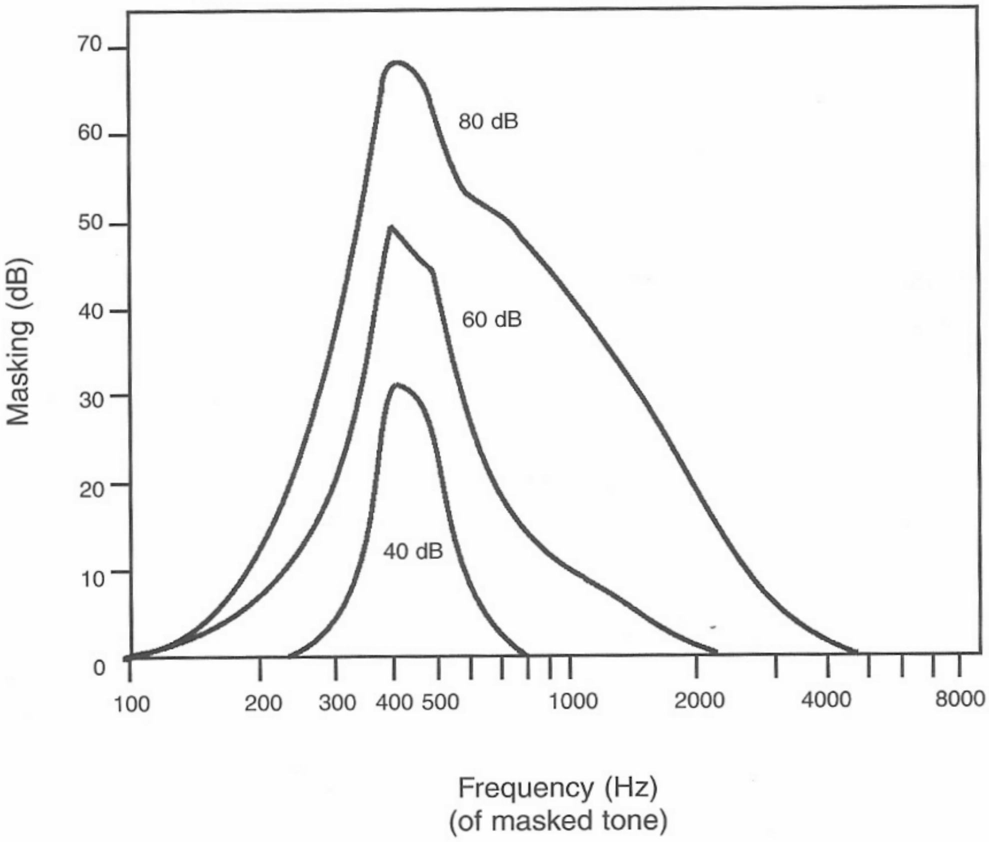


Figure 2. The effect of noise intensity level on the upward spread of masking (from Robinson & Casali, 2000).

deterioration of short term verbal memory. Gomes et al. (1999) found that workers exposed to the same type of noise as GA pilots are exposed to, high pressure amplitude (≥ 90 dB SPL) and low frequency (≤ 500 Hz), had significantly lower memory quotients using the Wechsler memory scale, and significantly lower immediate verbal memory. Furthermore, studies concerning vibroacoustic disorder (VAD), a disorder that arises from long-term exposure to high amplitude, low frequency noise, support these findings and add that long term exposure to the aforementioned type of noise leads to permanent cognitive and physiological maladies, such as vascular changes, poor memory retention, and low performance on attention tests (Albuquerque, da Gama, & Macedo,1991; Pais, Araujo, & Ribeiro, 1996).

It has also been suggested that long term exposure to noise contributes to the onset of cardiovascular disorders, most notably hypertension, much in the same fashion as everyday stress does. Noise stimulation causes the blood vessels to constrict and, in turn, the arterial blood pressure shall increase. Additionally, an acceleration in heart rate accompanies the vasoconstriction. As daily noise exposure continues over long periods of time, the body will habituate, resulting in the body's stabilization at an elevated heart rate and blood pressure, the symptoms of hypertension (Greifahn & Di Nisi, 1992). Peterson et al (1984) supported this theory with the use of monkeys which were exposed to 80 dBA noise levels for several months. Blood pressure was noted to increase every time the noise was introduced, and after a few months, the monkeys' blood pressure was noted to persist at an elevated level even after the termination of the noise stimulus.

Other studies conducted with human subjects at lower noise levels have elicited similar cardiovascular responses. The heart rate pattern showed two distinct acceleration periods (Eves & Gruzelier, 1984). Directly following the onset of noise, the first acceleration reaches a maximum heart rate at 3 – 4 seconds, and is thought to be an instinctive mechanism, preparing the body for fight or flight, independent of stimulus duration or state of consciousness (Di Nisi et al., 1990; Di Nisi, Muzet, & Weber, 1987). This initial reaction lasts 8 – 9 seconds and is followed by a second heart rate acceleration period. However, this episode is more ominous than the former. It lasts the entire duration of the noise stimulus and is believed to be responsible for the habituation response. Vasoconstriction is thought to coincide with the period of the second heart rate acceleration (Griefahn & Di Nisi, 1992). The onset of vasoconstriction occurs approximately 3 seconds after the onset of the noise stimulus and reaches its maximum response at 8 – 9 seconds following onset. Following the termination of the noise, the blood vessels dilate and usually return to baseline 11 –12 seconds later. As heart rate and vasoconstriction is controlled by the sympathetic branch of the autonomic nervous system, the combined reactions to noise indicate these reactions are the symptoms of an elevated sympathetic response to noise (Di Nisi et al., 1990; Di Nisi, Muzet, & Weber, 1987).

The type and intensity of the noise is also important concerning noise-induced cardiovascular disorders. In previous studies, constant low frequency noise, such as pink noise, has produced the largest sympathetic reactions, even above those of an impulse nature, such as gunfire (Kryter & Poza, 1980). This differentiation is thought to be caused by the difference in bandwidth between the two types of noise. Kryter and Poza (1980) found a positive relation between bandwidth of noise and the extent of vasoconstriction, such that as the bandwidth increases, as the case is for pink noise, vasoconstriction will become more severe. Conversely, in the case of the gunshot, the bandwidth is so small that the vasoconstriction is barely measurable.

Also, surprisingly low sound intensities exceed the threshold to initiate this response. Accelerated heart rates and vasoconstriction have been reported in workers at noise levels of 74 and 80 dBA (DeJoy, 1984). These noise levels, though deemed safe by OSHA standards for hearing conservation, may lead to cardiac disorders. Furthermore, the two aforementioned cardiovascular responses have been shown to continue even in workers who have been aurally habituated to the offending noise. Therefore, researchers have concluded that "enhanced and pathogenic responses" might occur in certain situations where long-term exposure is combined with "distinct noises, such as aircraft noise" (Greifahn and Di Nisi, 1992).

Noise not only plagues the physiological functions of the body, but also the psychomotor actions. Early studies have shown that performance decrements are more likely to occur under high-intensity, intermittent noise, than low-intensity, constant noise while conducting a target tracking task on an oscilloscope (Plutchnik, 1959). In 1971, Eschenrenner employed a more complex task to test this theory. Manual image motion compensation was the task used. This was basically an orbital simulator where a high-resolution photograph of the Earth's surface is viewed by participants through a telescope and a gimbaled mirror to simulate an orbital pass over the pictured area of the planet. The participants were allowed to move the photograph on two axes, and were instructed to stay at or below a velocity criterion of 40 microradians per second. The results of this study supported the noise – psychomotor performance degradation theory, finding an individual's ability to perform a far more complex task (image motion compensation) was significantly impaired by the temporal pattern and the intensity of noise, where high intensity intermittent patterned noise has the greatest adverse impact on performance.

However, there is some controversy concerning this topic. Some studies have called these findings into question, showing no performance decrements during high-intensity intermittent noise involving an individual's ability to track a moving target (Hack, Robinson, & Lathrop, 1965; Plutchnik, 1961). Data from these experiments clearly show an initial decrement in tracking performance at the onset of the auditory stimulus, and a subsequent gradual improvement in performance concomitant to the individual's adaptation to the auditory stimulus. The discrepancy between these conflicting results seems to lie in the difference in complexity of the task.

Performance of easy tasks is theorized to improve in the presence of noise, where noise is an arousal factor in an otherwise underloaded situation (Scott, 1962). Therefore, tasks such as manual image motion compensation, or piloting an aircraft, which are very complex cognitive and psychomotor tasks, requiring continuous sensory information processing, decision-making, and action sequences, are extremely susceptible to extraneous distracters, such as noise (Eschenbrenner, 1971). Other more simplistic tasks are more resistant to the adverse effects of noise.

SPEECH INTELLIGIBILITY

Humans communicate with each other through vocalization in every setting and situation where two or more are gathered (Whitaker, Peters, & Garinther, 1989). This vocalization is a very important and intuitive means of conveying information when the message is short, interactive, and need not be referred to later; just the type of message teams, such as a pilot and an air traffic controller, rely upon to accomplish tasks in the flight environment. Speech communication is especially important when time is limited and a rapid exchange of information cannot be accomplished by typing a message for display to another team member (Whitaker & Peters, 1993). Unfortunately, as imperative as speech communication is to many situations and settings, it can be easily thwarted by poor intelligibility. Thus, a number of objective and subjective measures have been developed to estimate the intelligibility of speech in a variety of different applications and environmental factors.

Measures of Speech Intelligibility

French and Steinberg (1947) began the application of what is now known as critical band theory to the objective measurement and prediction of speech intelligibility by incorporating the contributions of various physical parameters of a transmission channel into a single index. Earlier, Fletcher (1940 as cited in Robinson & Casali, 2000) developed critical band theory with his observations that mechanically, the ear acts like a series of acoustic filters, with the bandwidth of each adjacent filter proportional to the filter's center frequency. When this theory is applied to masking by broadband noise, only a narrow critical bandwidth of noise positioned at the center frequency is effective as a masker. K.D. Kryter further refined the theory and application of critical bands throughout the 1940's, 1950's, and 1960's.

Articulation Index. In 1969, the American National Standards Institute (ANSI) adopted their "method for the calculation of the articulation index" (ANSI S3.5 – 1969). This index, built on previous work from French and Steinberg, and K.D. Kryter, whom had refined application techniques to transduce a speech sound pressure into an electric signal, then were able to channel it through a set of adjacent bandpass filters. Taking the resultant voltage levels for each frequency of interest, mathematical calculations can derive different measures such as the root mean square (rms) bandpass pressure (Kryter, 1962).

The Articulation Index built upon this principle with two calculation methods to determine a numerical representation of the quality of speech intelligibility based on the spectra of the noise and speech through a transmission channel. The first method is the 20 band method. This method separates the speech and noise spectra, extending from 200 Hz to 6100 Hz, into 20 contiguous bandwidths for calculations. The second method is the 1/3 octave band or octave band method. This method breaks the noise and speech spectra up into 1/3 octave bandwidths or octave bandwidths. For obvious reasons, the 1/3 or octave band methods are not nearly as precise as the 20 bandwidth method, and therefore will not discussed in this report (If the reader is interested in these methods, he or she is directed to ANSI S3.5 – 1969 for a detailed account).

A summarized version of the AI calculations will be presented (Again, if the reader desires greater detail they are directed to the ANSI S3.5 – 1969 standard):

Step 1: Plot the known or estimated spectrum level of the speech peaks. These can be estimated by adding the frequency response characteristics of the system and the difference between the overall long-term speech rms and 65 dB. If necessary, also subtract a correction factor if speech is presented over a loudspeaker in a reverberant or semi-reverberant room. This step will yield the effective spectrum of speech.

Step 2: Plot the corrected spectrum level of the steady-state noise reaching the ear of the listener. Correct the noise spectrum for increased masking effectiveness at any center frequencies of the noise that exceed 80 dB.

Step 3: Plot the effective masking spectrum of noise. Choose the largest of the noise spectrum at each center frequency from the noise spectrum, the corrected noise spectrum, or the spread of masking spectrum.

Step 4: Calculate the difference, in dB, between the spectrum level of the speech peaks and the effective masking spectrum for each center frequency of the 20 bands.

Step 5: Add all the difference results together and divide by 600. The resultant is the Articulation Index measurement (Figure 3 shows an example worksheet with calculations).

Speech Intelligibility Index. In 1997, ANSI updated the AI with the adoption of ANSI S3.5 – 1997. This standard illustrates the calculation of the Speech Intelligibility Index (SII), which now supersedes the Articulation Index. However, the calculations for the SII will not be covered in this report as the SII has some major drawbacks concerning its generalizability for various applications. First, the standard states that it does not account for distortions or the upward spread of masking effects, it only accounts for threshold changes while the listener is wearing a hearing protection device (HPD). Secondly, and probably most important to speech intelligibility, the new standard includes no procedure to convert the calculated SII score to a psycho-acoustic measure of percent words correct, or some other psycho-acoustic measure based on the proportion of a speech message comprehended by a human listener (Robinson & Casali, 2000). This oversight is quite unfortunate as the only direct measure of human speech intelligibility is to measure the amount of speech a human listener comprehends. Therefore, the SII is an indirect measure of speech intelligibility, which for accuracy and validity should be compared to a psycho-acoustic measure.

Speech Transmission Index. A second objective method of speech intelligibility came along in the early 1970's. Steeneken and Houtgast (1972) developed a method which utilizes a speech-spectrum-like artificial signal played over a communication channel to measure the speech intelligibility (A summarization is given here, for further

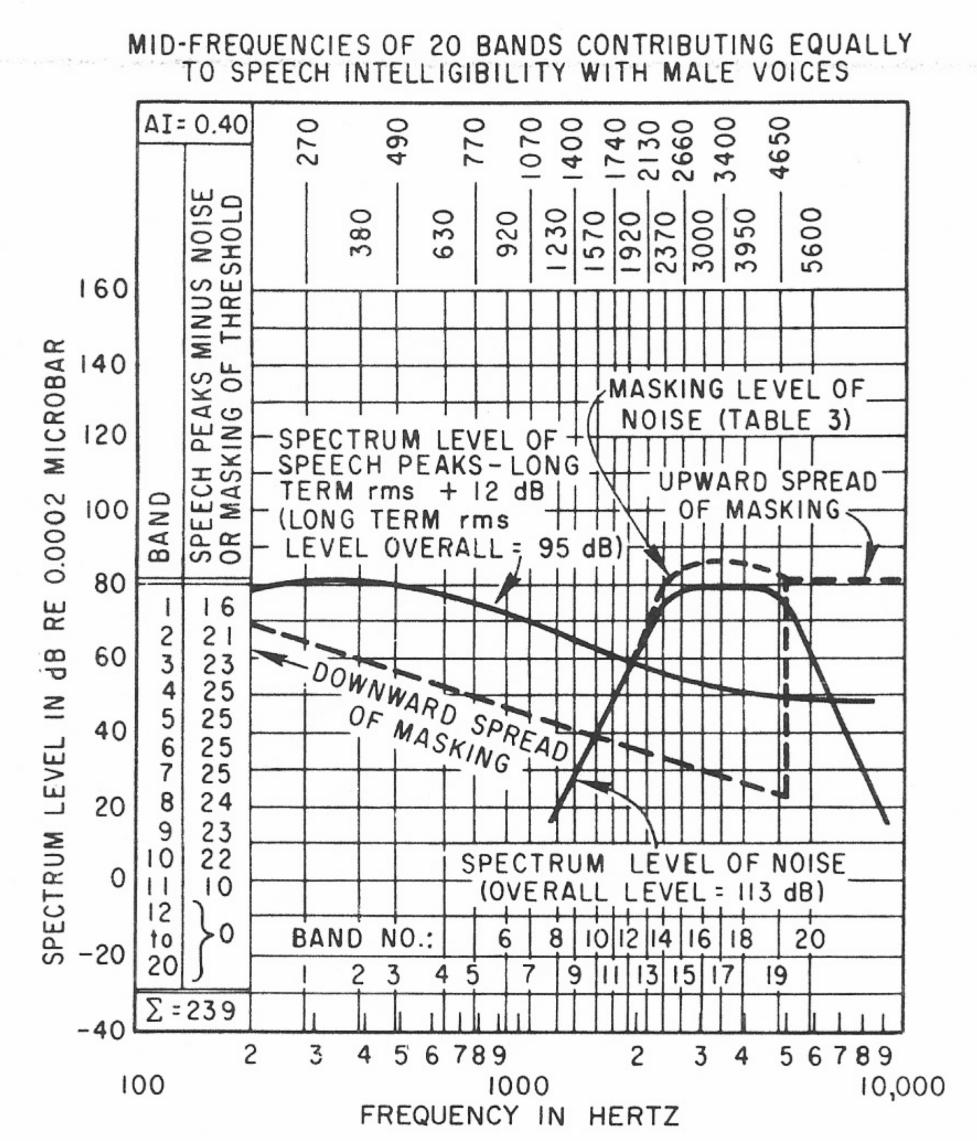


Figure 3. Sample AI worksheet with calculations (from ANSI S3.5-1969).

detail and clarification refer to Steeneken & Houtgast, 1972). This method is known as the Speech Transmission Index (STI).

Steeneken and Houtgast began the development of the STI by realizing that it would be much simpler to measure intelligibility if they could identify a single representative physical feature of speech for use as a test signal. They found this in the speech and noise frequency spectra, or more specifically the preservation of dissimilarities between the noise and speech frequency spectra as they are transmitted along the communication channel (Steeneken & Houtgast, 1971). Thus, the first calculation of the STI is a computation of the dissimilarities between the frequency spectra. These differences are defined for two sounds, i & j, as:

$$D_{i,j} = \sum_{n=1}^{N} \left| L_{i,n} - L_{j,n} \right| \quad ^{p}$$

where, $L_{i,n}$ and $L_{j,n}$ are sound pressure levels (SPL), in dB, for sounds i & j at the nth band filter (based on 1/3 octave band analysis).

This equation holds true for measuring differences at the talker's end and the listener's end of the transmission. The first stage comparison between the talker and listener is defined by the next equation, and designated the transmission index (TI). This ratio is an indication of how well the dissimilarities are transmitted.

$$TI_{i,j} = {D'_{i,j}} \big/ {D_{i,j}}$$

where, $D'_{i,j}$ is the dissimilarity at the listener's end.

The TI will become the STI for two critical sound spectra representative of the entire amount of preservation of dissimilarities between the speech sounds in general. These two representative sounds must be related to the SPL of speech normally applied to the transmission channel. To do this, first a speech reference level (SRL) is taken with a SPL-meter. Sound one will be set equal to the SRL, while sound two will be the SRL minus ΔL, where ΔL is derived by plotting the two sound spectra. Finally, an octave band weighting and an alternating rate is applied to account for time domain distortion. The final equation which yields the STI measurement is:

$$STI = {}^{1}\!/_{\alpha} \sum_{n=1}^{5} \alpha_n \left({}^{\Delta L'_n}\!/_{\Delta L} \right)^p$$

where, ΔL (dB) = the initial SPL difference
α_n = octave band weighting factor

Later revisions and extensions of the STI have taken into account nonlinear distortions, such as peak clipping. The resulting revised STI and the original AI are arguably the most widely used objective measures and predictors of speech intelligibility through a communication channel (ANSI, 1969, 1997; Elhilali, Chi, & Shamma, 2003).

Modified Rhyme Test. The most commonly known subjective, direct speech intelligibility test was developed by House et al. (1965). The Modified Rhyme Test (MRT) is used to measure the speech intelligibility levels of the spoken word, whether through a communication channel, or in varying environmental factors, such as aircraft engine noise. The test consists of 50 English words. The words are presented one at a time by either a speaker or pre-recorded taping. The task of listener is to choose the correct word spoken from a closed set of six rhyming possibilities. The index of speech intelligibility is the percentage of spoken words identified correctly by the listener. This method has been utilized very successfully in investigations extending from its development up through present day (Lancaster, 2004; Lancaster, 2005; Urquhart, 2002; Whitaker & Peters, 1993; Whitaker, 1991). Due to the reliability, success, and popularity of this speech intelligibility measure, a definition of speech intelligibility has evolved based on its methods. Speech intelligibility is the count of the percentage of words that the listener heard correctly (Kling & Riggs, 1972).

Revised Speech Transmission Index. For many years after the development of the AI, STI, and MRT, they were used without extensive modification. However, in recent years, speech intelligibility measurement has received renewed attention with a number of new techniques being developed, and extensions of established methods surfacing. Steeneken & Houtgast (2002) have revised their Speech Transmission Index to better accommodate intelligibility of the spoken word, more specifically, consonant-vowel-consonant (CVC) word scores. Previous STI methods have been mainly focused on the potential intelligibility of a signal through a communication channel. This revised Speech Transmission Index (STI_r) includes the parameters of the original index (frequency domain distortion, nonlinear distortion, time domain distortion), in addition to the new parameters of automatic gain control, waveform coding, redundancy correction for adjacent frequency bands, the prediction of specific groups of phonemes, and the redefining of the frequency weighting functions.

Even with the decades of revisions to the STI method, it is still mainly used with test signals to ensure high accuracy and good repeatability. Researchers argue that running speech applications are limited, as accuracy with the STI measurement is compromised in naturalistic vocalizations (Li & Cox, 2003). The STI can roughly estimate the transmitted and received speech envelope spectrum using the following equation (Steneeken & Houtgast, 1980),

$$\mathrm{MTF}(F) \approx \mathrm{Ey}(F) / \mathrm{Ex}(F)$$

where Ex(F) = the envelope of transmitted speech,

Ey(F) = the envelope of received speech.

This relationship would be precise if envelopes of natural speech were periodic and the spectra of speech were white with constant power per unit bandwidth. Unfortunately this is not the case. Speech is a highly complex, stochastic process which only approximately conforms to Steneeken and Houtgast's equation for the STI measurement. It is for this reason that the STI measurement can use speech signals in its calculation, but the modulation transfer function, at the core of its revised calculations, becomes highly inaccurate as opposed to the use of an artificial test signal (Elhilali, Chi, & Shamma, 2002).

Artificial Neural Network. Therefore, with the ever-growing popularity of neural networks and their exceptional ability to accurately carry out nonlinear computations, the STI technique has been transformed into an artificial neural network (ANN) in order to ensure more flexibility to accommodate the natural spoken word. The basic principle of the STI, speech must retain its original envelope to be intelligible, is still the foundation for this method. Therefore, as the spectrum of speech is increasingly modified, the intelligibility of the speech will become increasingly poor. The modification of the speech spectrum stems from contaminates such as ambient noise, reverberation, peak clipping, etc. The ANN calculates the sound spectra for the input sound, in this case running speech, and the output sound. If the difference between the two spectra can be obtained, then the ANN can make the nonlinear and linear calculations to obtain the STI for running speech intelligibility.

The neural network architecture consists of a nonlinear, multilayer, feed-forward, back propagation network, with a size of 40 – 20 – 10 – 1. The numbers in the network size relate to the number of neurons present at each layer of the ANN. Additional specifications of the network include back propagation, which is used in this case as a very powerful network training algorithm. The downside of this method is the consumption of great amounts of time and training data sets to properly train the neural network and reduce error production down to an acceptable level. Error, an ANN output, is reduced through the input of speech files with known STI values, then the ANN calculates an STI value and the two values are compared. The difference in the two values is the error term, which is fed back through the ANN (back propagation) to iterate all equations and weightings, modifying them slightly with the information the machine has "learned." A preprocessor is also used in the network architecture for data reduction purposes. If the preprocessor were not incorporated the number of input neurons needed to effectively process the amount of data in a speech envelope would make the ANN unwieldy (Li & Cox, 2003).

An ANN has been tested with both broadband and octave band speech signals in room acoustics and provides measurement accuracy on the same scale as the original STI measurement when using artificial test signals, typically a standard deviation of less than 0.2 has been found (Li & Cox, 2003). However, no testing has yet been done over a communications channel. Speech reproductions would need to be made and included in the training sets during back propagation training.

Spectro – Temporal Modulation Index. Another extension based on the STI method was proposed by Elhilali, Chi, and Shamma (2003). They theorize that any manipulation of speech that does not significantly disrupt the integrity of its spectro-temporal modulation is harmless to its intelligibility. Elhilali et al. concede that existing objective measures, such as the STI, effectively analyze the effects of noise, reverberation, and time domain distortion, but criticize that they do not assess the effects of a joint spectro-temporal modulation present in vocalization, which significantly changes the envelope of speech. Therefore, the Spectro-Temporal Modulation Index (STMI) was devised to quantify the degradation of the spectral and temporal modulations due to noise.

The need to quantify these modulation decrements is based on the theory that reliable, accurate representations of spectro-temporal modulation are needed for human perception. This theory is well grounded in the neurophysiological data from mammalian auditory cortexes and earlier stages in auditory processing (Dau, Puschel, & Kohlrausch, 1996; Drullman, Festen, & Plomp, 1994). STMI begins with a computational auditory model separated into two basic stages. The early stage models the transformation of the acoustic wave into a three dimensional representation, termed an auditory spectrogram

(Figure 4), are based on the impulses allowed through various filters used to simulate hair cell nerve impulses. The auditory spectrogram is then channeled into the central auditory stage. Here, the spectral and temporal modulation content of the spectrogram are estimated and the magnitude of the response as a function of frequency at each predetermined time instance. When the STMI was compared against human perception, there was a very strong positive linear correlation between the STMI result and the percent correct recognition scores for CVC words by human participants. This experiment was performed in the presence of the less than ideal conditions of white noise and reverberation. STMI was found to reflect the deterioration of intelligibility as well as the STI and subjective human listener tests for conditions containing combined noise and reverberation. However, the STMI further detects the distortions that are inseparable along the temporal and spectral dimensions, not reported by other speech intelligibility measures (Elhilali, Chi, & Shamma, 2003).

One of those distortions that occur regularly is the condition known as phase jitter. This distortion is commonly associated with telephone channels, caused by the fluctuations in power supply voltage (Lee & Messerschmidt, 1994). Phase jitter destroys the carrier of the speech signal, but leaves its envelope relative untouched. In other words, the time dynamics of the speech are left alone, but the spectral modulations are affected. As phase jitter (a) increases, speech intelligibility decreases until phase jitter reaches full strength (a = 1), at which point speech is heard as white noise. Phase jitter only affects the spectral modulation; it does not affect the modulation amplitude of the narrow band carriers used by the STI measurement. Therefore, even though speech

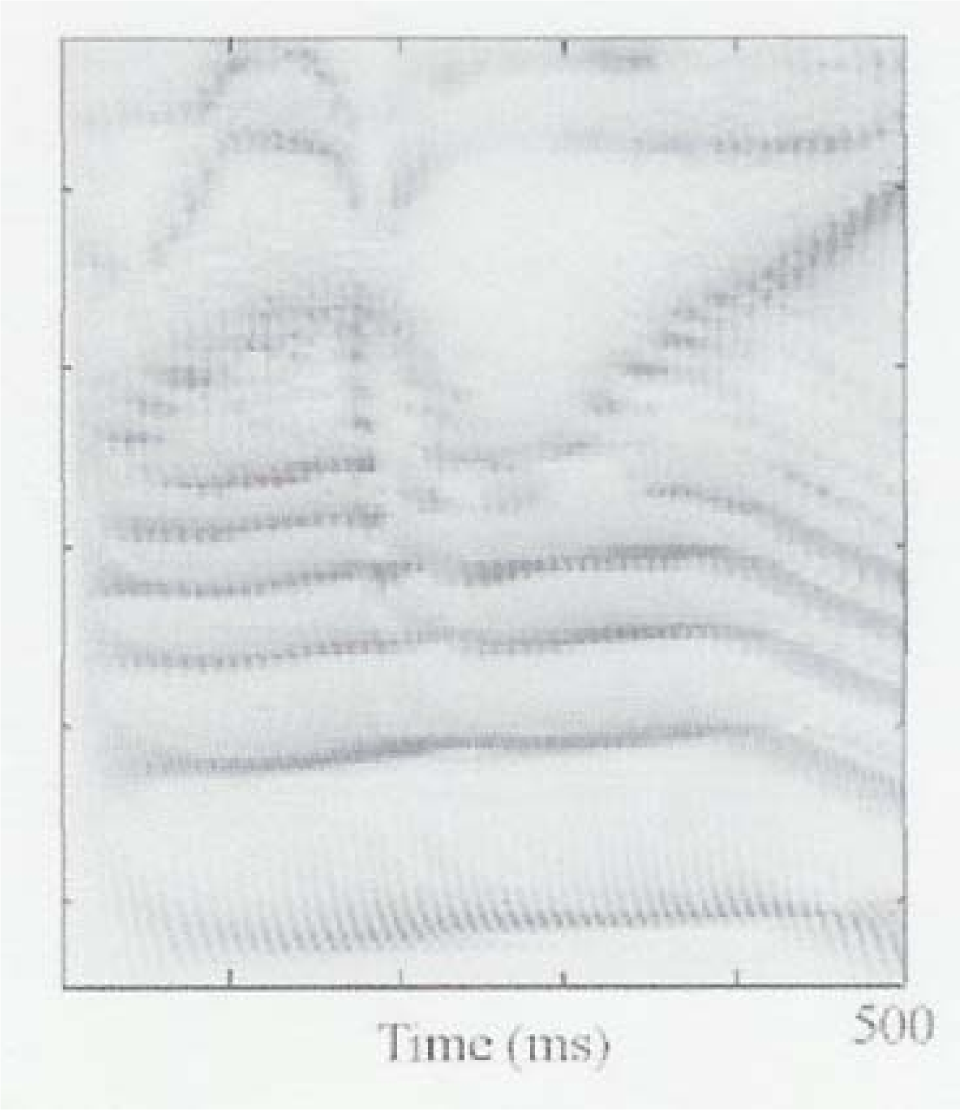


Figure 4. Sample 3-D auditory spectrogram (from Elhilali, Chi, & Shamma, 2003).

intelligibility is degraded, the STI will measure no difference as it is completely insensitive to phase jitter. However, during experimental trials incorporating phase jitter into speech lists, the STMI measurement, and psycho-acoustic word score show a marked decrease in speech intelligibility. When the severity of phase jitter reaches a = 0.75, the STMI and the pycho-acoustic test measure speech intelligibility to be 20% (Figure 5), while the STI test still measures speech intelligibility for the same speech trial as 100%(Elhilali, Chi, & Shamma, 2003).

The previously mentioned, and well known speech intelligibility measures incorporate a constant signal-to-noise ratio per trial throughout the entire test. However, a few adaptive procedures have also appeared in the literature. In the past, adaptive procedures have been mainly used to measure sensory thresholds, but this technique could also be appropriately utilized for the measurement of speech intelligibility (Dirks et al., 1982; Levitt, 1978). An adaptive technique is defined for this purpose when the signal-to-noise (S/N) ratio is determined by the subject's response to preceding test items. The main advantage offered by this method is a quick way to estimate the S/N ratio corresponding to a given percent-correct point on an individual's speech intelligibility response curve. The most common adaptive paradigm, and most widely used for sensory threshold measurements, is the up/down procedure with a fixed step size, such as 1 down/1 up, 2 down/1 up, and 3 down/1 up steps (Levitt, 1971; Levitt & Rabiner, 1967). Though these techniques have proven to be accurate and reliable, there are deficiencies inherent to their design. Due to the step being a fixed integer value, the adaptive procedures can be used only to estimate certain percent-correct values. To obtain other

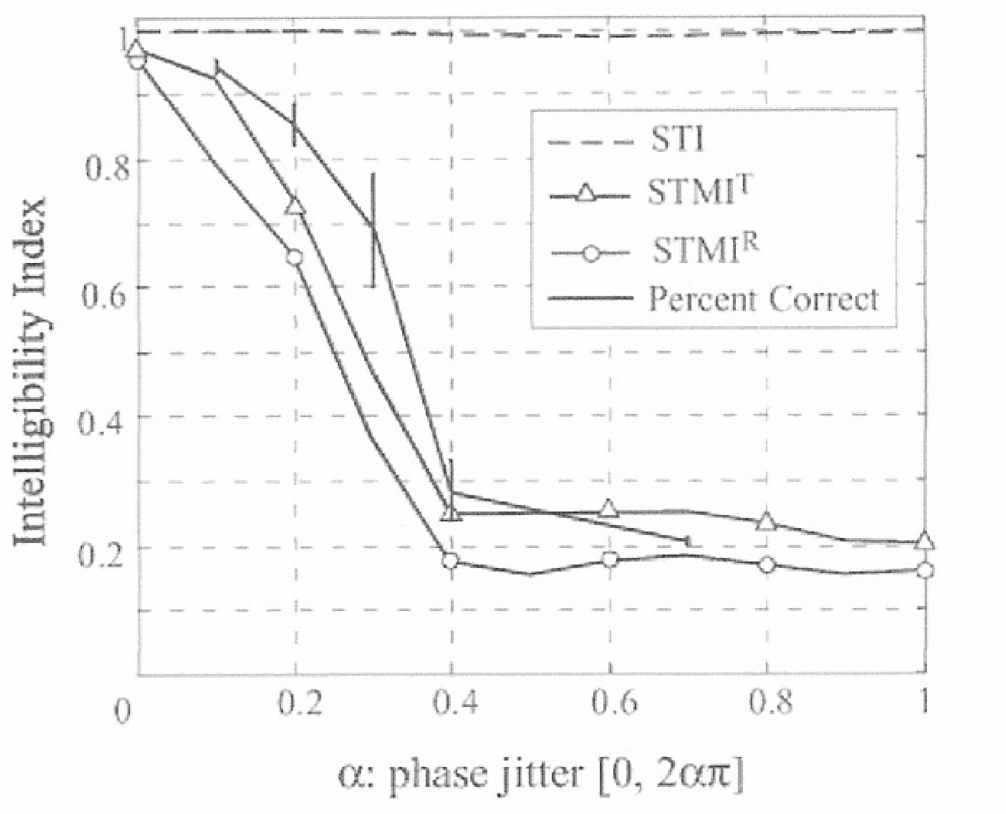


Figure 5. Phase jitter effects on the STI, STMI, & percent correct tests (from Elhilali, Chi, & Shamma, 2003).

percent values, a complex decision rule and a large number of responses at each testing level must be used. The large number of participants combined with the long trials and a complex statistical decision rule makes for a method with low efficiency and decreased practical value (Zera, 2004).

Maximum Likelihood Procedure. Due to the inefficiency of the stepwise adaptive technique, Zera (2004) has modified an adaptive maximum likelihood procedure (MLP) for use in measuring speech intelligibility. MLP has often been noted in the literature as the most efficient adaptive procedure developed to date (Green, 1993; Macmillian & Creelman, 1991). Certain experiments have shown adequate threshold estimates after as little as 12 participant responses (Green, 1993).

The cornerstone that MLP is based upon is the assumption that the shape of the participant's response curve, associated with a certain task, is known in advance and is invariant when expressed as a function of a certain stimulus intensity. Therefore, MLP can be used to determine the S/N ratio associated with a given percent-correct speech score. Zera also postulates that the MLP is not phoneme specific, and can therefore be applied to various speech intelligibility tests to create an adaptive version of that test. In the validation study, the MLP procedure was applied to the Modified Rhyme Test because the MRT test is so well known and so highly recommended, it would serve as the perfect platform and baseline with which to test the new procedure. Zera makes the additional claim that the MLP is also generalizable to sentence-based speech intelligibility, but it should be noted this claim has yet to be empirically substantiated.

The MRT words were presented in the presence of pink noise to test the adaptive MLP. Setting target word scores (TWS) for speech intelligibility of 58%, the MLP resulted in an average S/N ratio of –9.3 dB. When this S/N ratio was then retested utilizing a human listener in the MRT paradigm, the results showed a 59% average correct word score in speech intelligibility. Similar results were collected in subsequent trials utilizing higher percent-correct speech intelligibility levels. At a 75% correct MLP level, the S/N ratio was found to be –4.6 dB, which, when tested, translated into an average MRT percent-correct of 74.3%, and at the 90% correct MLP level, the S/N ratio was found to be –2.0 dB, which resulted in an average MRT percent-correct of 84.9% (Figure 6).

The results do show a systematic error bias, but Zera is quick to note that this may be due to statistical variability caused by a small number of data points collected. The data do seem to support this claim showing that the difference between data points and the normalized response curve of the MRT test (House et al., 1965) do not exceed the standard errors associated with the data points. Those standard errors reported by Zera (2004) were 0.5 – 0.8 for the point level estimates, and 2.5% - 3.5% for the MRT percent-correct word scores.

The study is concluded with a comparison of the MLP derived response curve and the MRT response curve as reported in House et al. (1965), and a Monte Carlo simulation. The slope of House et al.'s normalized response curve is 6.8 dB. The obtained function from the MRT-MLP has a slope of 6.34 dB. Therefore, the data seems again to be in agreement with the normalized response curve reported by House et al. (1965). The Monte Carlo simulations were conducted to conclude the number of test items that must be presented to reduce the standard deviation of the resulting S/N ratios to the level of the normalized data from the House et al. (1965) study. It was found that 25 test items are

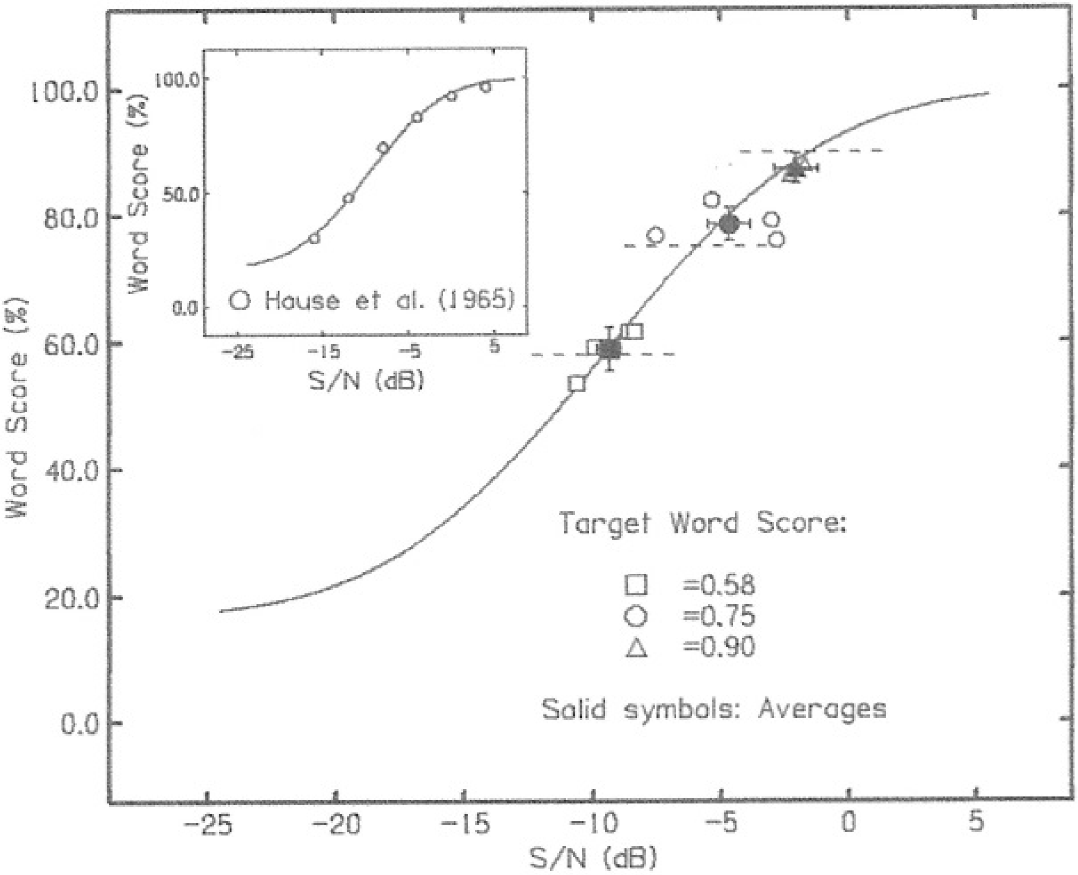


Figure 6. Response curve of the MLP compared to that of the MRT (from Zera, 2004).

needed to reach the normalized standard deviation level, but that the data converges on the target level in as few as 10 – 15 responses (Zera, 2004).

***Coordinated Response Measure*.** The Coordinated Response Measure (CRM) for speech intelligibility was originally developed by Moore (1981), and associate researchers at the Air Force Research Laboratory. Their goal was to provide the armed services with a speech intelligibility test which possessed greater relevance and superior external validity regarding military communications than the presently utilized MRT. The CRM incorporated sentences of the form "Ready (assigned call sign), go to (color) (number) now." These sentences were spoken by 8 talkers and received by 8 listeners with 8 assigned call signs, each working at a control panel consisting of two dials, one with 4 colors, and the other with 8 numbers.

The resultant speech intelligibility score was defined as the percent of correct numbers and colors identified from utterances associated with a listener's assigned call sign. A comparison study was conducted evaluating the CRM and MRT in a variety of communication channel jamming conditions. Results of this study supported the conclusion that the CRM test was less sensitive to interfering noise than the MRT, but that the overall performances between the two measurements were highly correlated (Moore, 1981).

Following the Moore (1981) account of the CRM speech intelligibility method, little information was reported until Brungart (2001) devised a normative evaluation of intelligibility using the CRM method to measure the signal-to-noise ratio with a speech spectrum shaped masking noise. During the experiment, each of the eight listeners heard 120 sentences taken from a 2048 sentence corpus and spoken by eight talkers in random order with masking noises randomly varied from 64 – 70 dB in 1 dB per trial steps. Speech signals during the experimental trials were scaled to produce signal-to-noise ratios of –18 dB to 15 dB in 3dB steps (Brungart, 2001). Brungart then attempted to relate CRM results to the well known Articulation Index measurement.

Not surprisingly, the CRM test outcomes showed performance results which fit the typical S-curve that is characteristic of most speech intelligibility measures conducted in the presence of noise. However, quite surprisingly, the data showed that the correct identifications of color and number in each trial were independent. The probability of a correct overall identification differed from the product of the independent identification probabilities by only 1.1% (Brungart, 2001). Therefore, it was concluded that coarticulation played only a minor role, if any, in the intelligibility scores.

The next comparison made in the CRM validation process was a comparison between the CRM results and predictions made utilizing the well validated Articulation Index. This comparison proved problematic. There seems to be an inherent and fundamental difference between the two techniques. While the AI is based on the use of the spectral properties of phonetically balanced words of speech to determine intelligibility, the CRM procedure operates utilizing a very restricted vocabulary with words that are not phonetically balanced. Therefore, Coordinate Response Measure is not directly comparable with the Articulation Index (Moore, 1981).

To circumvent this problem, Brungart attempted to use Moore's (1981) rough estimation of the AI response, which is based upon the MRT versus AI curve provided by Kryter (1962). These rough estimations based on earlier estimations did not make for sound scientific calculations and conclusions. It was later concluded that the CRM procedure was not appropriate to determine the AI of speech intelligibility, especially in the masking of speech (Brungart, 2001). A note of additional interest was found in Brungart's conclusions. Contrary to the results found in Moore (1981), Brungart (2001) found that the CRM procedure was quite sensitive, even more so than the MRT, to small intelligibility changes in difficult listening situations, such as very noisy work environments or the jamming of a communications channel to near inoperability. Further research will be needed to settle this discrepancy in the literature.

The CRM testing procedure does contain two major advantages. First, the test procedure can be conducted with multiple talkers and listeners carrying on simultaneously. This provides the experimenter with a rudimentary tool to test speech intelligibility in a multiple speaker, party line type scenario. The second major advantage lies in the simplicity of the intelligibility commands. These simple color-number commands lend themselves well to translation into many different languages, as most, if not all, languages contain terms for colors and numbers. On the other hand, the simplicity, a limitation to the same 32 possible color-number responses, limits the applicability of the technique when compared to the MRT or sentence-based speech intelligibility measures (Brungart, 2001).

Speech Reception Threshold. The previously mentioned subjective speech intelligibility tests share one fundamental flaw common to many designs. They are each subject to inherent floor and ceiling effects represented by the S-shaped performance curves characteristic of these tests. However, there is a method that is not subject to the same floor and ceiling performance effects, the Speech Reception Threshold (SRT) test developed by Carhart (1946). The SRT is defined as a step-wise adaptive test where the stimulus presentation level is increased or decreased depending on the listener's ability to correctly repeat the uttered material or not (Brungart, 2001). Specifically applied to the SRT test, this means that the presentation level is varied as necessary for a listener to recognize the speech materials correctly a preset, specified percentage of time. This percentage is generally set at 50% to find the true reception threshold, but can be changed according to the need of the experimenter. Levitt (1978) has described this adaptive approach as offering effective placement of the presentation level with "concomitant methodological improvements in efficiency and accuracy during estimation."

To derive the SRT, a stimulus, in this case speech in noise, is presented to a listener. The speech intensity level is increased in each subsequent test item after an incorrect response from the listener, or the speech stimulus shall be decreased in intensity in accordance with a correct listener response. The SRT is then estimated from the average of presentation levels in a preselected latter portion of the experimental segment. Implicitly important in this process is the need for speech material that is not only different but also of equal difficulty. Intertwined in this necessity is the need for either a familiarity of the material on the part of the participants, such as the use of their native spoken language, or extensive training to breed this level of familiarity (Nilsson, Soli, & Sullivan, 1994). These necessities must be designed into the experimental framework to counteract the severe damage that learning effects could have on the experimental results if subjects are allowed to familiarize themselves with the material spoken during testing. Using the SRT, the S/N ratio at the threshold can be derived to facilitate comparison of threshold measurements at different speech and noise presentation levels.

Dirks et al. (1982) tested speech intelligibility SRTs incorporating sentence speech material against measurements utilizing performance intensity functions, and concluded the SRT procedure provided the same information as the lengthier performance intensity function procedures. The use of sentences, instead of single word utterances, in the SRT method is thought to increase external validity. Isolated utterances or carrier phrases may not represent the normal spectral weighting, fluctuations, intonations, and pauses, as they would be found in naturalistic conversational speech (Dirks et al., 1982). Furthermore, sentences allow measures to be effectively collected utilizing certain auditory systems in noise where isolated utterances may not allow adequate duration to engage any dynamic processing characteristics of the auditory system. This underscores the basic need for sentence length measurements (Nilsson, Soli, & Sullivan, 1994).

Researchers of speech intelligibility metrics have called into question the efficiency of using longer length spoken material in testing (Brungart, 2001). However, according to the results of multiple experiments, sentences retain almost the same efficiency as those tests that incorporate singular utterances, and bear no difference in efficiency to those word tests which are embedded in carrier phrases (Dubno et al., 1984; Hagerman, 1982; Hagerman, 1984; Plomp & Mimpen, 1979). The testing length of a typical Hearing in Noise Test (HINT) incorporating 1 list with 10 – 12 sentences lasts approximately two minutes (Nilsson, Silo, & Sullivan, 1994). In fact, sentence length material used in the SRT format has been found not only to be efficient, but also to be highly reliable in results (Nilsson, Soli, & Sullivan, 1994).

This is not to say that the sentence SRT (sSRT) is without drawbacks. Significant learning effects have been found to plague the procedure because of the repeated use of the same words, even if their positions within and between sentences are changed between trials (Hagerman, 1984). Also, the method of scoring the sentences has been troublesome. Dubno et al. (1984) scored the final word in each sentence, but the limited amount of data points produced from this method significantly reduced the efficiency and reliability of SRT measures. Later, Gelfand et al. (1988) scored the entire sentence utilizing a more complex experimental procedure during vocalization. The results of this later experiment were significantly more reliable with an average within subject difference score of less than 1 dB signal-to-noise ratio (S/N ratio).

Unfortunately for the United States, the majority of the research concerning SRT and sSRT was done with speech material developed in Dutch, German, and British English, each incorporating the languages' own idioms and dialects. To fill in this gap, Nilsson, Silo, and Sullivan (1994) developed the HINT test. They adapted a British SRT sentence corpus to American English and tested the material for sSRT use. They found that the mean presentation level of the speech stimulus stabilizes after the fifth sentence response. This was determined based on the standard deviation fluctuations diminishing and finally changing very little from the fifth sentence through the entire list. Therefore, all measurements in Nilsson, Silo and Sullivan's study were obtained using the average of the fifth and all subsequent sentences, regardless of list length. The results also have shown that the sSRT procedure is reliable and sensitive enough to detect a threshold difference of 2.21 dB for one 12 sentence list, or a 1.12 dB threshold difference for three

12 sentence lists. The average reported threshold for speech in 72 dBA noise was 69.08 dBA, or a -2.92 dB signal-to-noise ratio (Nilsson, Soli, & Sullivan, 1994).

Reliability and repeatability of the sSRT measures and consequently the standard error of the results can be estimated from the standard deviation of the differences in subject responses between repeated measure trials (Plomp & Mimpen, 1979). The standard deviations of the HINT sentence lists associated with each overall list mean was quite small, fluctuating within 1 dB of the mean. For further support an analysis of variance found no significant effect for list type F(2,149) = 1.97, p > .01. Confidence interval widths for the sSRT are not noise level dependent as the SRT is linearly related to noise level once the noise level crosses to suprathreshold (Nilsson, Soli, & Sullivan, 1994). The width of a confidence interval is rather determined by the number of sentences per list and the number of lists incorporated into the experimental trial.

Another question regarding the use of sSRT that must be addressed is the effects that the stimulus bandwidth may have on the reliability of the test. The Speech Reception Threshold, in noise, is known to increase as the bandwidth of the speech signal is reduced (Bronkhorst & Plomp, 1989). This situation was also tested for the sentence Speech Reception Threshold. sSRTs in a reduced speech bandwidth condition were found to be significantly higher than the full bandwidth condition. An increase of 3 dB was common as the upper octave bands were eliminated and the bandwidth was reduced to approximately 2 – 2.5 kHz. Bandwidth reduction from 2 kHz to 1 kHz produced an even higher sSRT increase of an additional 7 – 9 dB. These findings are in agreement with Plomp (1986) calling the reliability of the sSRT measure into serious question as the testing bandwidth drops below 2.5 kHz (Nilsson, Silo, & Sullivan, 1994).

Speech Intelligibility under Communication Headsets

The easiest and most effective method to improve speech intelligibility is to simply raise the signal level of speech. This will, in turn, increase the most crucial variable in speech intelligibility, the signal-to-noise ratio, thereby increasing speech intelligibility (Kryter, 1985). This is easier said than done. With such high intensities of cockpit noise threatening to permanently damage the pilot's hearing as it is, raising the speech signal level to overcome the background noise would only exacerbate an aurally dangerous situation.

Though speech intelligibility is unfortunately readily affected by the presence of background noise, there are avenues other then raising the speech signal level that may be taken to intervene and improve speech intelligibility at the listener's end. Studies conducted on hearing protection devices (HPDs) have resulted in improved speech intelligibility for normal hearing listeners in high noise level environments. It's theorized that the improvement is due to an overall reduction in sound level at the cochlea, which can then respond without distortion (Berger, 2000).

The effect of HPDs on speech intelligibility is a highly complex subject influenced by many different variables, such as a person's hearing sensitivity, the absolute noise and signal levels, and the signal-to-noise ratio. Data are available which demonstrate an improvement in speech intelligibility in the presence of high noise levels, above 80 – 85 dBA, when the ears are occluded by a passive attenuating hearing protector (Casali & Horylev, 1987; Townsend, 1978). Further studies have uncovered a decrease in speech intelligibility below the 80 dBA level (Howell & Martin, 1975; Suter, 1992), which has led to the general rule that a noisy environment should be attenuated below 85 dBA, but above 70 dBA (ISO 1996).

As previously mentioned, a person's hearing sensitivity influences the HPD's effect on speech intelligibility. This influence can be strong when dealing with hearing impaired listeners. HPDs can potentially attenuate sounds below the individual's threshold of audibility rendering speech not only incomprehensible, but inaudible (Berger et al., 2000). Unfortunately, there have been no definitive studies to set the threshold between improving and degrading speech intelligibility for HPD use by the hearing impaired. However, a rough estimate has been proposed at a hearing threshold level of 35 dB when averaged across the frequencies 2000, 3000, and 4000 Hz, based on the work of Lindeman (1976).

Additional variables arise in the real world that significantly affect speech intelligibility while using HPDs, such as visual cues, the context of the message, and the experience of the listener in high noise level environments (Berger, 2000). Rink (1979) confirmed the influence of visual cues on speech intelligibility when he demonstrated that both normal hearing and hearing impaired subjects maintained speech comprehension regardless of which HPD was donned as long as visual cues were also presented. Moreover, Acton (1970) showed that subjects experienced with listening in noise were better able to discriminate speech than a non-experienced control group of equal hearing sensitivity.

In aviation, the majority of pilots wear a communication headset, which is basically a passive attenuation device, much like a HPD, but with an integrated communication system (boom microphone and earcup loudspeaker). David Clark, one of the most well known and effective passive attenuation aviation communication headsets, was measured to afford 92% speech intelligibility in light aircraft noise (Townsend, 1978). Townsend inferred that the high level of speech intelligibility was realized due to the higher attenuation provided by the headset, as compared to other passive headsets, which decreased the signal-to-noise ratio to a negative ratio at the speech reception threshold in the experimental situation. In other words, a reduction in noise received at the pilot's ear lowered the S/N ratio at threshold increasing speech intelligibility. Furthermore, it was suggested that for pilots to realize the best possible comprehension of radio and intercom communications in light aircraft cockpits, they should wear a headset which is capable of highly attenuating the environmental noise (Robertson & Williams, 1975

Further work by Townsend and Olsen (1979), in attempts to improve speech intelligibility under the aviation communications headset, was focused on the manipulation of binaural speech phase and its contributions toward eliminating masking effects. These two researchers attempted to identify a masking level difference (MLD). Earlier work by Tobias (1970) discussed an improvement in speech as phase was manipulated. Speech intelligibility scores were increased by 35% when masked speech was presented out of phase by two sound sources on opposite sides of a listener's head (i.e., dichotic listening). Tobias noted that this effect was only seen when the out of phase speech signals were presented concurrently with constant in phase noise, also presented through the two sound sources.

Townsend and Olsen (1979) were not as fortunate in their experiments investigating out of phase speech in the light aircraft cockpit. Their results showed a mean MLD of 0.9 dB, showing that the phase manipulation had afforded no means of escape from the masking effects in the cockpit. Further experiments were taken out of the aircraft cockpit and conducted in a laboratory setting using a tape recording of the aircraft noise held in phase and transmitted along with the speech signal over a communications headset. In this case, the MLD increased significantly to 6.6 dB. Therefore, it was concluded that the noise present in the light aircraft cockpit is of a random phase, which is known to nullify MLD measurements (Tobias, 1970; Townsend & Olsen, 1979).

Another variable that can adversely affect speech intelligibility, even if the pilot is wearing a sound attenuating headset is headset leakage. Headset leakage occurs when the circumaural seal between the headset and the skull is broken. A broken seal could be the result of improper fit or hair, but in the case of the pilot, it's generally the result of wearing sunglasses which hook around the ear. The arms extending to the lenses break the circumaural seals, allowing noise, especially low frequency noise, to enter under the headset. Low frequency noise, that which dominates the cockpit environment, enters through the small leak so readily because the long wavelengths of these noise frequencies, below 1000 Hz, are far less susceptible to obstruction than short wavelength, high frequency noises (Wagstaff, Tvete, & Ludvigsen, 1996). Wagstaff, Tvete, and Ludvigsen found that headset leakage can cause an average decrease in speech intelligibility of 39% with monosyllabic speech intelligibility tests. Furthermore, speech intelligibility may decrease as the bank of test items increases because better S/N ratios are needed to maintain the same speech intelligibility level for increasing vocal complexity (Nilsson, Soli, & Sullivan, 1994).

Although research into speech intelligibility has produced a wealth of knowledge concerning the variables that can positively and negatively affect speech intelligibility, Whitaker and Peters (1993) present an important point. Past studies have completely overlooked the impact that the degradation of speech intelligibility can have on operator effectiveness while performing a task. In their investigation, participants performed a simulated tank gunnery task while speech intelligibility was measured utilizing the MRT, and mental workload was measured using SWAT. Results of the testing showed that degraded speech intelligibility interferes with task performance and increases mental workload (Figure 7). Mental workload ratings begin to increase at the point speech intelligibility decreased from 100% to 75% as the tank crew perceives the onset of communication difficulties. Actual task performance is, however, more robust with no significant deficiencies until the speech intelligibility degradation reached 50%. Therefore, it was concluded in this first experiment that mental workload may be a valid forewarning of impending overload and task performance degradation due to compromised speech intelligibility (Whitaker, Peters, & Garinther, 1989).

In the second investigation conducted by Whitaker (1991), the interaction of speech intelligibility and task difficulty, and their effects on task performance, were studied. A task in the study was defined as more difficult if it was more speech-intensive to carry out correctly. Speech intelligibility was theorized to be affected by three parameters, size of vocabulary, variability of protocol, and expectancies of communications (Miller, Heise, & Lichten, 1951). The results concurred with the last

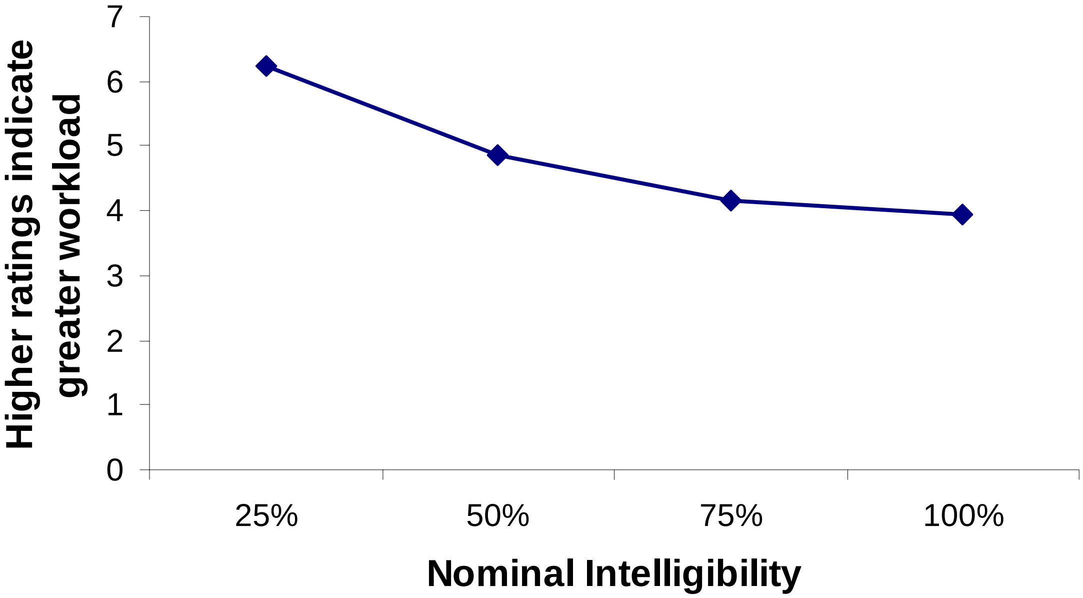


Figure 7. Relationship of workload and speech intelligibility (Whitaker and Peters, 1993).

study. SWAT, again, showed an increase in perceived task difficulty at moderately high levels of speech intelligibility, even though operational performance measures showed no decrease. Performance deteriorated more drastically than the previous study for the more complex speech intensive tasks (Figure 8). Performance metrics associated with speech intensive tasks are more adversely affected than those associated with less intensive ones. These data were thought to have resulted because the vocabulary and uncertainty of the communications were greater, and communications were coordinated between more crew members, which were summarized in Whitaker's performance versus intelligibility curves. It was theorized that the underlying variable in this investigation is uncertainty. Loss of speech intelligibility results in uncertainty because of the loss of information contained in the degraded communications. Therefore, mental workload increases and performance decreases due to the resources being removed from task performance and reallocated to communications tasks (Whitaker, 1991).

A final study of the performance-speech intelligibility relationship was conducted by Whitaker and Peters in 1993. Previous results were again supported. Whitaker and Peters drew several conclusion from this and prior studies. First, performance is a definite, direct function of speech intelligibility. Therefore, as speech intelligibility declines, so will task performance. Workload rating decrements generally precede performance decreases and may indicate that individuals have less reserve mental resources to cope with the degraded levels of intelligibility.

Data collected during this investigation showed that performance is quite robust to intelligibility decreases, as long as communication is a simple diad, and the task requires only standard operating procedures. However, as tasks become more complex,

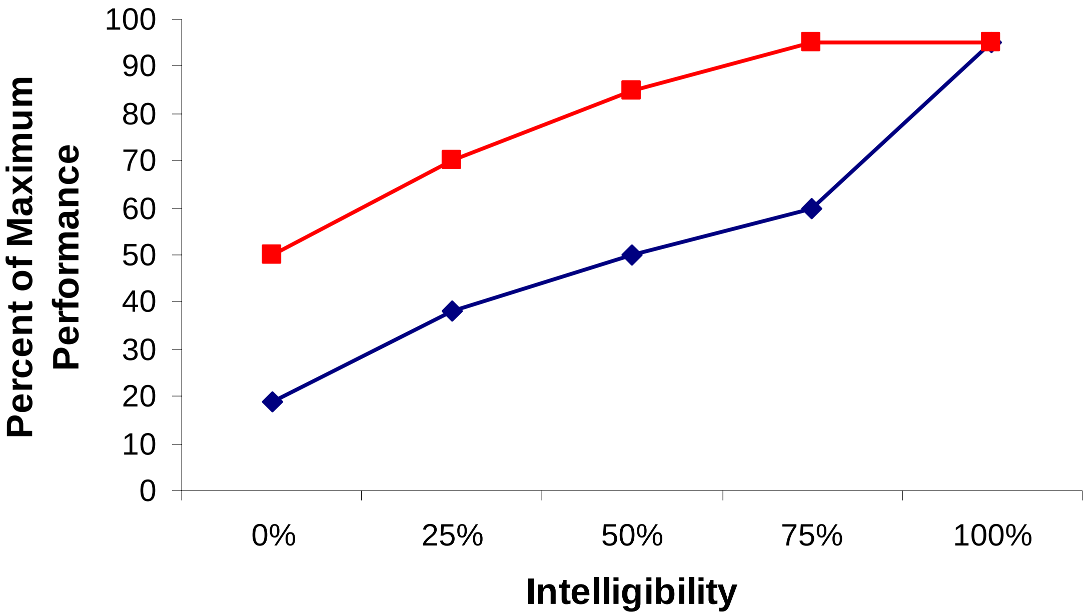


Figure 8. Relationship of performance and speech intelligibility at various task difficulties (blue line is simpler gunnery task, red line is more complex navigation task; Whitaker, 1991).

the adverse effects of the losses in speech intelligibility become more costly in terms of time, errors, and workload. In complex situations team performance dropped when intelligibility decreased from 100% to 75%. Even smaller decreases from 100% intelligibility could result in adverse performance effects preventing tank crews from achieving mission success (Fig. 10, Whitaker & Peters, 1993).

Rationale for Chosen Speech Intelligibility Experimental Method

Presently, the subjective measures of speech intelligibility (e.g. MRT, sSRT, etc.) are highly accurate, easily to administer, and the only known methods of directly, empirically measuring human speech intelligibility. Unfortunately, the methodological designs of these investigative procedures do not readily fit into the experimental design of the flight scenarios because actual communication is part and parcel of any such scenario. This experiment will manipulate three preset speech intelligibility levels throughout each flight scenario. Therefore, there will be no need to measure percent words correct during the simulation. This will already have been done in a preceding pilot study. Furthermore, the monosyllabic word tests, and sentence – based speech intelligibility tests, such as the MRT and sSRT, do not reflect natural running speech or the aviation language, and intermingling the words and sentence lists into the radio communications would destroy the high level of realism which this study is intending to investigate.

Therefore, the objective measure of Speech Transmission Index was chosen to control the speech intelligibility levels in this experiment. The STI has undergone major

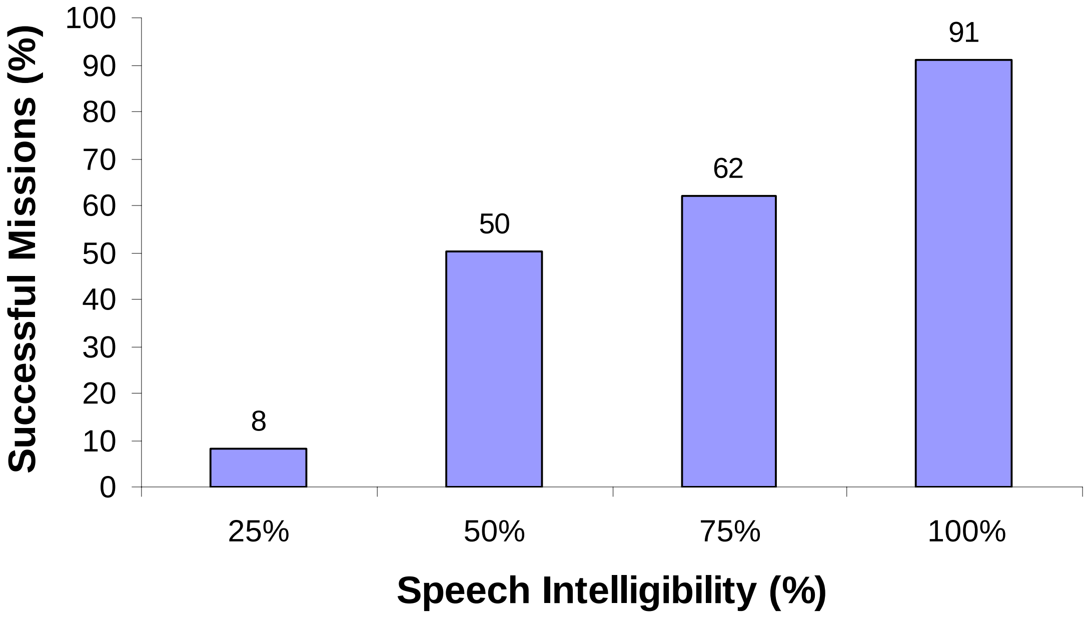


Figure 9. Missions successfully completed at varying levels of speech intelligibility (Whitaker & Peters, 1993).

revisions over the years to correct deficiencies when applying it to running speech. Presently, the STI is a widely used measure to predict speech intelligibility. This predictive nature makes it the perfect choice to control the three levels of speech intelligibility used in this study.

ACTIVE NOISE REDUCTION

In 1930 the French engineer, H. Coanda, documented and subsequently patented the idea of canceling a bothersome, or unwanted sound by adding a sound wave to the environment which is identical in every way to the first sound wave, except its phase is 180 degrees out of phase with the bothersome sound. This suppression of sound energies came to be known as destructive interference. Three years later, a German physicist, P. Leug found an application for the idea. Leug theorized using active noise reduction (ANR) as an alternative to the circa 1930's passive controls for low frequency noise in ducting. Leug designed a simple system, which would measure the primary disturbance and subsequently introduce a secondary disturbance utilizing a transducer to cancel out the primary disturbance. The major disadvantage of Leug's system design was that it made no allowances in the design for the active noise reduction system to adapt to any changes in the environment, or the equipment itself. However, this is a rather moot point, as neither Coanda nor Leug ever demonstrated a successful operational system. Therefore, the theory and design of an active noise reduction system slipped from the consciousness of researchers for a couple decades (Tokhi & Leitch, 1992).

The technology was revived by H.F. Olson during his investigations into the possibilities of using ANR in small rooms, ducting, and personal headsets during the 1950's. Olson was successful in constructing an operating system, but because of the limitations in the available electronic hardware and the state-of-the-art in control theory, his system provided limited attenuation over a small frequency range. Around the same time, another researcher employed by General Electric, demonstrated another active noise cancellation system applied specifically to reduce electrical transformer noise. Although this system operated successfully, it was not practical for two major reasons. First, the system only reduced noise over a small angle subtended between the loudspeaker and the measurement microphone, making the system relatively ineffective at reducing overall transformer noise. Secondly, and most impractical, the system required the ANR controller to be adjusted manually. Therefore, an operator had to be present at all times to listen to the noise levels and adjust the controller appropriately to optimize the active noise reduction performed by the system.

In 1957, ANR technology was first adapted to earmuffs. Meeker (1957) proposed what he called the “Active Ear Defender Systems” for the United States Air Force and developed two viable systems incorporating active attenuation technology. One system (System I) utilized open-loop architecture, while the second system (System II) utilized a closed-loop design. A third system was designed, but proved not to be an effective noise reduction system. Following development, the first two systems were tested at Wright-Patterson Air Force Base. System I successfully attenuated sounds by 10 dB at frequencies between 10 and 1000 Hz. System II attenuated sounds by 20 dB from 50 to 500 Hz.

Since the 1950’s the sciences of Electronics, Control, Signal Processing, Acoustics, and Vibration have made giant leaps forward in knowledge and applications, and with these leaps, active noise reduction has become a more viable consumer product, especially when dealing with one-dimensional sound environments such as air conditioning ducting or aviation communications headsets.

How Active Noise Reduction Works

Theoretically, the ideal ANR system would create an entire secondary sound field 180 degrees out of phase with the primary source, changing the radiation impedance, or the environmental resistance to the sound wave propagation of the primary source and leaving a listener in wonder of what happened to all the sound energy. The secondary sound field is actually “unloading”, or suppressing the sound power of the primary source, again by changing the environmental resistance. To accomplish this effectively the ANR control source must be large enough and located at a distance such that it is able to produce the required radiation impedance to the primary source. To illustrate the ANR process an electrical analogy will be used. A standard 110 V electrical wall socket sends 0 Watts (W) while nothing is plugged in because the environmental resistance allows no electricity through; that is, until, a 60 W light is plugged into the socket. At that point, the socket sends 60 watts, but if a 1500W radiator is plugged in it will send 1500 watts. Therefore, the electrical power depends on the electrical impedance that the power point experiences. In like manner, the acoustic radiation impedance of the primary source can be altered by the introduction of a secondary acoustic signal. Of course, total active sound cancellation of even perfect periodic noise is idealized. In reality, changes in the environment, harmonic distortion, and even construction and geometric differences between the primary and control sources prevent the achievement of total silence; rather a reduced level of noise is generally the goal. Some noise, such as non-periodic and completely random, cannot even be controlled utilizing active noise reduction due to the fact that the sound characteristics cannot be predicted in advance (Hansen, 2001).

Basic Control System Structures

There are two types of control systems used to analyze the noise and generate the correct inverse signal of the offending sound. For the purposes of this review, non-adaptive ANR systems will not be discussed as very small changes in the sound or the environment can render non-adaptive systems utterly useless. Most modern active sound control systems are self-tuning, or adaptive, affording them the capability to respond appropriately to sound or environmental changes, in addition to changes in the system itself, such as loudspeaker wear.

The first type of active sound control system is the most widely used presently in existence. The adaptive feedforward system is named because a reference sensor samples the noise and feeds it forward to the control system where the noise is filtered by an electronic controller. Then the signal is analyzed and the controller responds by sending the appropriate signal to an output source. Further down the sound's path of travel an error sensor again samples the residual sound pressure and provides a signal to the control algorithm to measure controller effectiveness and appropriately adjust the output to obtain the minimization of error, or, physically, the minimization of any residual sound pressure. Crucial to the effectiveness of a feedforward system is the signal processing time from reference microphone sampling to the output, which must be less than, or equal to, the time sound needs to propagate from the reference sensor to the control output location (Bartholomae & Stein, 1990).

The second type of control system is the active feedback control system. The difference between the two systems is how the control signal is derived. Where the feedforward system samples the sound first, the sound has already passed the electronic controller when it’s sampled by the error microphone in a feedback system and then the signal is sent back to the electronic controller to be analyzed. Therefore, due to the very nature of a feedback system, the sound it is attempting to attenuate must be predictable for the system to perform effectively. Feedback systems are generally limited to sounds of a constant or periodic nature. Although this is an ideal system for situations where it is not possible to sample the sound early enough to effectively utilize feed-forward technology, the performance of feedback systems are not nearly as good as feed-forward (Hansen, 2001).

To construct an effective feedback system, several important decisions must be considered. First, the feedback system must be designed by considering the physical system and the electronic control system as one system. This must be done because any change in the sound, environment, or equipment can severely limit the attenuating capabilities of the active control system if overlooked. The delay between the error sensor input and the controller output must be considered as it is the limiting factor for the effective bandwidth of frequencies that the feedback controller will be able to attenuate. In fact, the bandwidth of effective control is proportional to the reciprocal of the delay (Hansen, 2001). Lastly, one of the major disadvantages of using the feedback system is that impulsive and/or high frequency disturbances can cause the controller to issue positive feedback, thereby undesirably increasing noise levels at certain frequencies, usually around the band of 1000 Hz to 3000 Hz (Bartholomae & Stein, 1990). However, this characteristic can be controlled by installing low pass filters to attenuate high frequency signals to an amplitude which will not interfere with the functioning of the active feedback control system.

Active Noise Reduction in Headsets

Besides the system architecture, active noise reduction designs for headsets can be classified according to another dichotomy, open or closed back devices. Open back ANR headsets can easily be recognized by the supra-aural design, where the ANR electronics are surrounded by foam pads, which are situated on the pinnae. In a siren-canceling study, open-back ANR headsets were effective noise attenuation devices. However, their performance did vary according the siren type ranging from 8 to 22 dB at the 800 Hz peak frequency, and 15 dB attenuation at 4000 Hz (Casali & Robinson, 1994).

Open-back ANR attenuation performance is generally good, but the major drawback to the design is the total lack of passive attenuation capabilities. Due to compact design, which only leaves room for the ANR electronics, there is no appreciable passive attenuation. Therefore, if the electronics of the ANR system fail, no hearing protection is provided to the individual. In the high noise levels of a GA cockpit, the lack of backup passive attenuation makes the open-back system undesirable. Therefore, the present study will focus solely on closed back, circumaural headset designs.

In addition to the closed back design, ANR aviation headsets generally utilize the active feedback or a very simplified active feedforward system design (See Figure 10 for an active feedback schematic with an integrated communications circuit; Steeneken & Verhave, 1996; Hansen, 2001). The previously discussed problems with system instability are the major reason active systems are generally only effective for situations where the offending sound exists in the low frequency range (below about 1000 Hz). At higher

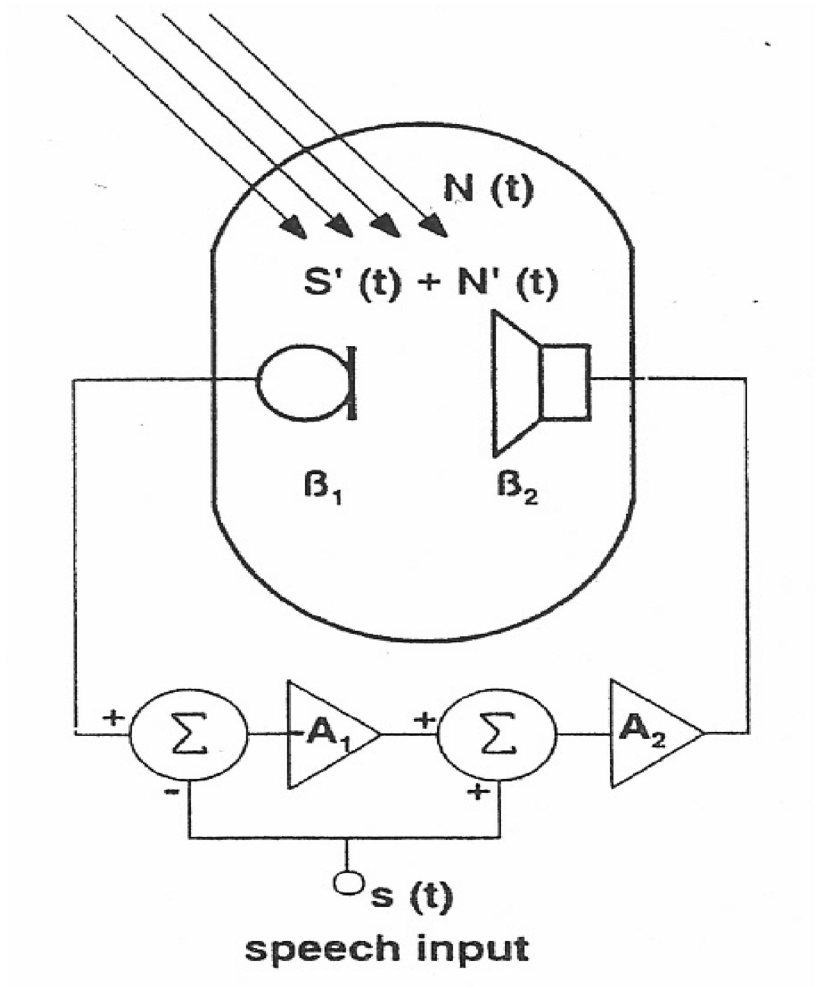


Figure 10. ANR headset with integrated communications circuit where, N(t) = primary noise signal, N'(t) = resulting noise signal, S(t) = primary noise signal, S'(t) = resulting speech signal, B_1 = frequency transfer for microphone, B_2 = frequency transfer for loudspeaker, A_1 = gain and frequency transfer of correction amplifier, A_2 = gain and frequency transfer of loudspeaker (from Steneeken & Verhave, 1996).

Frequencies, passive systems are less problematic, more effective, and much less costly, making them the prudent choice.

Conversely, the 15 – 30 dB attenuation common to conventional passive hearing protection devices and communications headsets may not be effective in protecting the wearer from low frequency noise, such as the noise found in the cockpit of GA aircraft or military vehicles. Furthermore, the attenuation of low frequency noise maybe inadequate to reduce at ear noise levels so speech may remain intelligible (Gower & Casali, 1994). This is due to the fact that the short wavelengths of midrange and high frequencies can not penetrate the materials and construction of the passive headset as effectively as the long wavelengths associated with low frequency noise, which easily pass through the headset or travel by bone conduction through the skull to the cochlea of the inner ear. Low frequency noise causes a HPD wearer to hear sounds in the environment as attenuated, muffled sounds. A general warning found in the literature cautions that light aircraft noises are of such a high intensity level that the very limits of earmuff attenuation may be exceeded (Casali, 1989; Gower & Casali, 1994; Wagstaff, Tvete, & Ludvigsen, 1996; Van Wijngaarden &Rots, 2001; Wagstaff & Woxen, 2001). Earmuffs alone cannot ensure total protection against noise-induced hearing loss. In reaction to environments with very high levels of noise, it is common to incorporate earplugs underneath the earmuffs in attempt to augment the attenuation characteristics and gain better relief from noise. However, when speech is involved, the double hearing protection will necessitate even higher volume levels set on aviation headsets to make speech audible. This situation has been demonstrated to lead to highly distorted speech, and dramatically reduced speech intelligibility in the operational environment (Wagstaff & Woxen, 2001).

For many situations and environments, passive hearing protection or headsets are adequate. Although due the inherent attenuation design, the passive HPD can have adverse effects on the quality of the sound underneath the headset, and the auditory performance of the wearer. First, passive HPDs unbiasedly attenuate all sounds coming to the ears of the listener altering the listener’s interpretation of those sounds. This means that sounds such as signals, alarms and speech, which are essential means of communication, are equally attenuated with the unwanted noise.

For present-day ANR technology, active attenuation of 20 – 25 dB (Figure 11) is common but limited to noise below 1000 to 2000 Hz (Casali, 1992; Gower & Casali, 1994), and can be used for intensely high noise levels up to 160 dB SPL (Steneeken & Verhave, 1996). Additionally, because of the equal attenuation of all sounds, the signal to noise ratio does not improve, which is the most important factor in achieving adequate and reliable signal or speech intelligibility (Casali, 1992). In reality, the signal-to-noise ratio may decrease quite dramatically under the passive headset for high intensity low frequency noise due to the upward masking effect. The bias towards attenuating midrange and high frequency noise exhibited by passive HPDs coupled with the upward masking perpetrated by the low frequency noise creates a situation where signals and speech above 2000 Hz are those usually missed by wearers, especially those with a preexisting permanent hearing loss. Therefore, communication using standard communication headsets in noisy environments is very difficult.

The assessments of active noise control versus passive noise control have been sparse. Presently, ANR headsets cannot be sold or approved as hearing protection devices

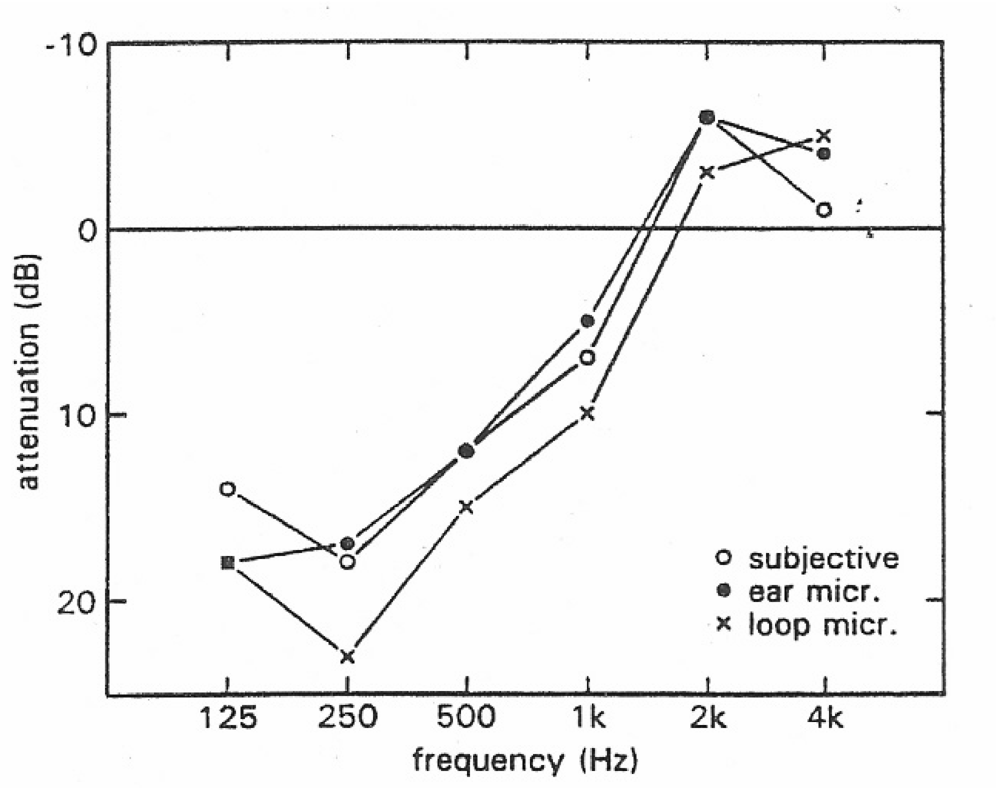


Figure 11. ANR attenuation curve (from Steeneken & Verhave, 1996).

under EPA regulations and do not receive an noise reduction rating (NRR) according to the standard measuring method (EPA, 1979; ANSI S3.19 – 1974; ISO 4869-1) because the noise introduced by the electronic systems of active control contaminate the results of the threshold of hearing metrics used for the rating of passive devices (Steneeken & Verhave, 1996). However, the few studies that have measured the attenuation performance and speech intelligibility of active attenuation headsets in comparison to passive attenuation headsets have reported mixed results, at best. Nixon et al. (1992) studied three ANR headsets, comparing the active mode speech intelligibility versus the passive mode. The results showed no practical differences in the two modes, leading the researcher to comment that the ANR functions "were not impressive." Using the same methodology, Wheeler & Halliday (1981) reported a 26% improvement in speech intelligibility with the ANR electronics were turned on, versus when they were turned off. The results of other studies using the same methodology have reported results which fall somewhere in the middle of a continuum between Nixon et al. and Wheeler & Halliday (Chan & Simpson, 1990; as cited in Gower & Casali, 1994).

Gower and Casali (1994) were the first to conduct an in-depth study into the attenuation performance and speech intelligibility characteristics of an active attenuating headset compared to a purely passive attenuating headset using two high quality aviation communications headset, a Bose ANR headset, and a David Clark headset. They reported the Bose ANR outperformed the David Clark in attenuation by up to 22 dBA (depending upon frequency). Average attenuations of the 106 dBA tank noise resulted in 80.9 dBA (28% OSHA noise dose) for the Bose ANR headset and 94.4 dBA (75% OSHA noise dose) for the David Clark headset. It should be noted that for any noise dose above 50%, OSHA regulations require a hearing conservation program to be instituted at the place of business.

The speech intelligibility testing used lower intensity pink noise, instead of the higher intensity tank noise, because even though the pink noise was of lower intensity, it contained higher energy in the critical bandwidth of speech than did the tank noise, and therefore masked speech more effectively. Results of the speech intelligibility testing between the two headsets show that although the Bose ANR headset clearly displayed better attenuation characteristics, the David Clark faired better during the speech intelligibility measures. In two experiments, the Bose ANR headset required a significantly higher S/N ratio, on average of 12 dB, for 70% speech intelligibility than the David Clark headset required. It seemed that David Clark overcame its lower attenuation characteristics with a better frequency response characteristic in the critical bandwidth of speech (Gower & Casali, 1994).

A later study (1997) conducted for the U.S. Army evaluating the effectiveness of a new model of ANR armor crew headsets concurs with Gower & Casali concerning the attenuation properties of ANR headsets, but disagrees concerning the speech intelligibility characteristics of the active headset. For armored vehicles, such as the Bradley Fighting Vehicle, where noise levels are routinely 114 dBA with peaks that exceed 128 dBA, ANR headsets reduce exposure level to 83 dBA and extend daily exposure times from 20 minutes to 12 hours. Additionally, speech intelligibility scores using the previous passive military headset versus the new ANR headset show an increase of 21% in speech intelligibility as measured by the MRT (Anderson &

Garinther, 1997). 68% intelligibility was achieved using the previous headset, while 89% intelligibility was achieved using the new ANR headset.

The most recent evaluation of ANR technology was also conducted for the Army. Urquhart (2002) evaluated two different ANR systems developed for the operators of the Bradley Fighting Vehicle along the variables of attenuation performance, effects on speech intelligibility, cognitive performance, and mental workload. The attenuation performance of the ANR systems on the 114 dBA Bradley noise environment is was a noise reduction of nearly 30 dBA to 83 dBA, or 90 dBA with the communication system turned on. Moreover, while the author made some other fascinating comparisons between the microphones used in the Bradley ANR systems, what is most interesting to this study is some of the overall conclusions Urquhart drew regarding ANR. He found that the ANR system with the best low frequency attenuation afforded the highest speech intelligibility scores and at these high speech intelligibility levels, subject performance on a cognitive assessment battery was also highest. Therefore, it was concluded that ANR was able to indirectly raise speech intelligibility across the communication system, and the more intelligible speech allowed the participants to reallocate attention and mental resources to the cognitive assessment battery, resulting the higher scores.

The results of Urquhart's study support the theory which provides the foundation for this dissertation study. ANR will attenuate the low frequency noise enhancing the speech intelligibility of communications in a high level, low frequency dominant noise environment. This increase in speech intelligibility will decrease the mental resources required for effective communications and the freed mental resources are able to be reallocated to another concurrently performed task, raising the performance levels on this task.

ANR headsets have the potential to enhance hearing protection and speech intelligibility at low frequencies (usually below 1500 Hz). ANR technology is especially effective for this function at the ear because of the sound propagation under a headset is one dimensional, just as it is in ducting which has already seen good results with the use of only one channel (Hansen, 2001). Furthermore, when the low frequency attenuation advantages of active noise attenuation are combined with the high frequency attenuation advantages of passive noise attenuation, the best results are usually achieved (Figure 12, Wagstaff, Woxen, & Andersen, 1998).

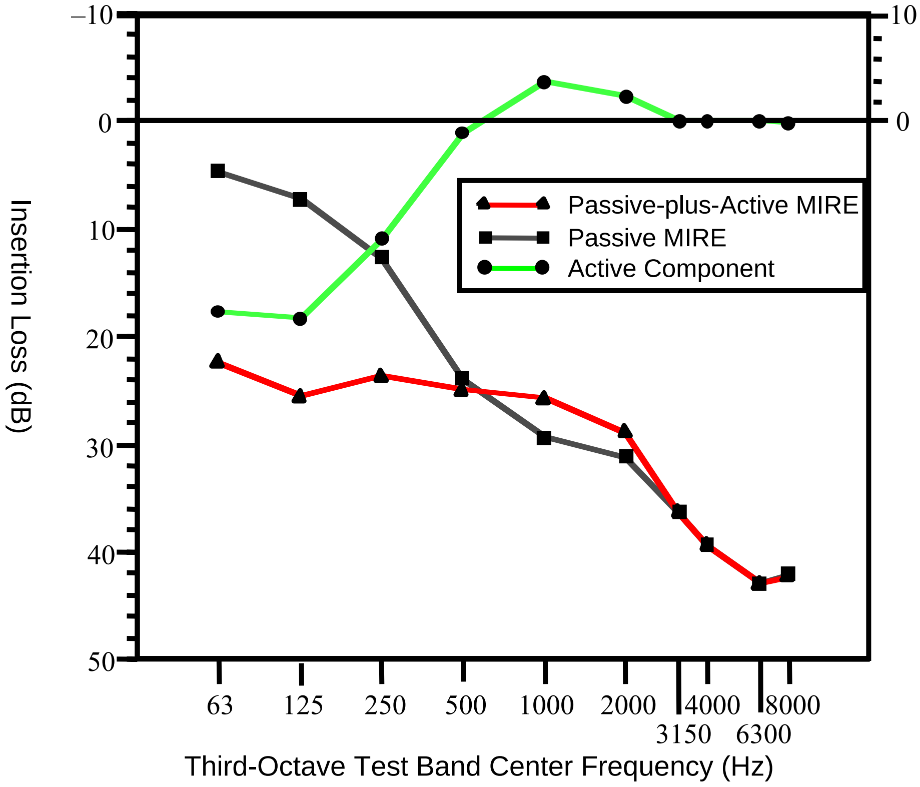


Figure 12. Active, passive, and active + passive attenuation curves for the NCT PA-3000 Closed-back headset (from Robinson & Casali, 1995).

MENTAL WORKLOAD

An issue in the late 1970's brought mental workload to the forefront of researcher's attention. The issue concerned the elimination of the flight engineer position from all medium-ranged commercial airliners (Lerner, 1983). Since that time, research in the area has amassed into a voluminous knowledge base, which has spurred the realization that modern systems, such as the aircraft, and aviation in general, have forced the pilot to be less reliant on physical skill and far more reliant on mental capabilities.

Modern aviation systems require a pilot to make safe, intelligent, and timely systems management decisions, often under the pressures of crucial time constraints (Hankins & Wilson, 1998). A system, or environment, like the previously mentioned one, which constantly and continually imposes multiple concurrent task demands, has the potential to unknowingly exceed the pilot's available resources (Stark et al., 2000). Therefore, many techniques have been devised to quantify mental workload in an attempt to prevent overload before it leads to a tragic accident. However, before a description of the most prevalent techniques is given, a definition of mental workload and specific criteria regarding its measurement must be put forth.

A Definition of Mental Workload

Mental workload has been defined in many different ways; however, every definition contains the commonality of a description of an operator-system relationship, where the operator performs certain tasks to achieve a specific goal through the system. Mental workload is imposed by the system in the form of task demands. The operator responds to the task demands by allocating a certain amount of mental resources from the operator's finite resource supply. Two other relationships have been derived from this operator-system relationship which have proved valuable to researchers and system designers. There is an inverse relationship between workload and operator reserve resource capacity, and an inverse relationship between workload and task performance (Wickens & Hollands, 2000). These relationships, though useful, do not provide absolute values regarding what is an acceptable versus what is an unacceptable workload demand. Absolute values concerning workload do not exist. Rather the comparison of relative values is the common practice in this field (Wierwille & Eggemeier, 1993).

Criteria for Evaluating Mental Workload Metrics

When evaluating the scientific literature to determine the best mental workload metrics for use in a particular project, many researchers have compiled and iterated a list of criteria which can be used as a scale against which to measure the array of mental workload indices (O' Donnell & Eggemeier, 1986; Wierwille & Eggemeier, 1993; Humphrey & Kramer, 1994; Wickens & Hollands, 2000).

- Sensitivity: The degree to which a measurement can distinguish between changes in performance, task difficulty, resource demands, or levels of workload. Those measurements that display the capability to reflect variations in a number of different factors which affect mental workload are known to demonstrate a broad bandwidth of sensitivity known as global sensitivity.

- Diagnosticity: The ability to discern the specific type or cause of the level of workload. It is also associated with the ability to attribute mental workload to a certain operator task, or aspect of the operator task. This criterion is often associated with multiple dimension scales based on Wickens' (1991) multiple resource model.

- Transferability: Often associated with global sensitivity, metrics demonstrating this criterion posses the ability to be used to measure different variables in a variety of applications.

- Selectivity: A measurement which is sensitive only to variables which influence the resource expenditure imposed by mental workload during information processing. If a technique displays sensitivity to certain variables, or only in certain types of applications, the technique could prove very useful as a diagnostic aid.

- Intrusiveness: This is a very undesirable property where the index artificially contaminates the mental workload results by interrupting or influencing the operator's performance of the task. The researcher should be particularly wary of intrusiveness when using a secondary task to gauge mental workload performance.

- Bandwidth and Reliability: This criterion is especially important when measuring workload over time, such as a measurement for peaks in workload over a minute, hour, or day. In this case, the metric must offer reliable measurement rapidly enough to gauge any changes that occur over the specified amount of time.

Classification of Mental Workload Metrics

Mental workload measurements are classified into three general categories, performance-based measures, physiological measures, and subjective measures. Each of these measures possesses their own advantages and disadvantages, which are generally unique to the application they are incorporated within.

Performance-based measures. Performance-based measures focus on evaluating the task performance of an operator, be it system-related tasks or extraneous tasks imposed while the operator performs system-related tasks. When dealing with measurements concerning the operator performance of system-related tasks, these are termed measures of the primary task. On the other hand, when dealing with measurements concerning extraneous tasks to the system, which have been imposed upon the operator solely for the purpose of measurement, these are termed measures of the secondary task.

Primary task measures are based on the assumption that, in general, speed and/or accuracy of system-related task performance will decrease as workload increases above some critical threshold (Eggemeier & Wilson, 1991). Primary task measures can prove to be very valuable tools, demonstrating high levels of sensitivity to operator resource demand, if they are specific to the task demand and workload is high enough to consume available resources (Wierwille & Eggemeier, 1993). Additionally, due to the intuitiveness and high degree of face validity, Wierwille & Eggemeier urge that primary task measures be included in all investigations of mental workload.

Primary task measures can also find themselves at the other extreme of the sensitivity spectrum. Even if multiple tasks are included, if the primary tasks are too easy and do not surpass the operator's critical threshold, the low to moderate workload demands will not trigger a performance decrement, illustrating the potential these measures have to be extremely insensitive (Wierwille & Eggemeier, 1993; Wickens & Hollands, 2000). Insensitivity can also occur when the operator posses the ability to expend additional resources to meet task demands and maintain the status quo performance, even in the face of increased workload. This scenario is especially prevalent with well-trained, skilled operators, such as pilots. Lastly, Wickens and Hollands (2000) point out that primary task measures can easily fail because decision-making in the face of increased mental workload may impose enormous cognitive demands upon an operator. Therefore, the simple psychomotor performance actions are an unrepresentative measure of all the entailed mental operations. Therefore, researchers have turned to other techniques to more directly measure the investments into the primary performance, or the reserve capacity still available.

In response to the aforementioned concerns regarding reserve capacity and primary task measures, secondary task measures were developed. Secondary task measures are based on the theory that the secondary performance is inversely proportional to the primary task demands (Wickens & Hollands, 2000). Therefore, if the primary task is easy, demanding low amounts of resources, the secondary task performance will have a lot of resources available to draw on, and secondary performance measures should be high. If the primary task difficulty increases, so will the resource demand, leaving few available resources for the secondary task, whose performance will proportionally decrease (Sanders & McCormick, 1993). Common secondary tasks may be as simple as mental arithmetic or a tracking task. The most commonly used secondary tasks are listed below.

- Time estimation: the participant is instructed to mentally gauge and report a certain length of time while performing the primary task (Casali & Wierwille, 1983). This technique has proved to be more reliable than retrospective time interval estimation (Hart, 1975). Past studies have indicated that time estimation is highly sensitive to perceptual demands, but quite insensitive, and often intrusive, to communications and mediational demands (Casali & Wierwille, 1984).

- Rhythmic tapping: the participant is instructed to produce finger or foot taps at a constant rate. The tapping variability increases as the primary task workload increases (Michon, 1966).

- Random number generation: the participant is instructed to produce a series of random numbers. As primary task workload increases, the participant begins to generate more repetitive, less random, series of numbers (Wickens & Hollands, 2000).

- Probe reaction time: reaction time of a secondary task is measured. The assumption governing this technique is that increased primary task workload will delay secondary task reactions (Wierwille & Eggemeier, 1993).

The choice of a secondary task requires first an analysis of the primary task to uncover its resource demand characteristics. Then a secondary task can be chosen so that the resource demands between the primary and secondary tasks overlap, yielding the highest secondary measurement sensitivity (Casali & Wierwille, 1984).

Two variations of the secondary task technique are available to researchers for use as workload measures. The first variation is a loading task. Whereas, the primary task was the priority when using a secondary task, the loading task is now the task of priority. Participants are instructed to devote any necessary resources to the loading task, and the degree to which the loading task intrudes on the performance of the primary task is the measurement of workload. The second variation is an embedded secondary task. When intrusion of the secondary task is a confound that must be removed from the experimental design, an ideal alternative is to make the secondary task a legitimate system task, but at a lower priority than the primary task (Wickens & Hollands, 2000). The secondary task can be performed concurrently with the task whose workload is to be assessed. This method removes the artificiality of the secondary task design making the whole design completely transparent to the participant. For example, if the primary task of a flight simulation is flying an ILS approach with the localizer and glideslope needles centered, than an embedded secondary task could be ATC radio communications. Therefore, performance on the radio would indicate the amount of resources available without the pilot realizing there is an extraneous workload task involved in the simulation.

Physiological measures. Another method researchers have been using to circumvent intrusion problems in workload measurements is through the use of physiological measures involving the autonomic and central nervous systems (Kramer, Sirevag, & Braune, 1987). Though these measures are not as sensitive or diagnostic as performance-based measures or subjective measures, they are appealing because of their unobtrusiveness and the ability to record and store continuous, running workload estimates over long periods of time. Many different physiological parameters can be found in the literature. Unfortunately, there are very few instances where workload metrics are systematically compared, the one major exception is the series of instrument flight simulator-based experiments run by Casali & Wierwille (Casali & Wierwille, 1983; Wierwille & Connor, 1983; Casali & Wierwille, 1984; Wierwille, Rahimi, & Casali, 1985).

In the first series of four studies, physiological measures did not fair well in the sensitivity analysis. In Wierwille & Connor (1983) workload measures were compared while performing a psychomotor task. The only significantly sensitive physiological measure was heart rate, $F(2,10) = 8.89$, $p = 0.006$. The next study investigated workload emphasizing communications tasks, where again, only one physiological measure reached significant sensitivity, pupil diameter, $F(2,10) = 5.90$, $p = 0.0203$. Pupil diameter was found to differentiate low workload from moderate workload, and low workload from high workload, but not moderate workload from high workload (Casali & Wierwille, 1983). The third study in the series investigated workload during a perceptually-natured task. Again, only one physiological measure reached significant sensitivity, respiration rate, F(2,10) = 8.02, $p < 0.008$ (Casali & Wierwille, 1984). In the final study of the series, workload was measured as it affects mediational/central processing activities. Eyeblink (F(2,10) = 4.58, p = 0.0388) and fixation fraction (F(2,10) = 11.33, p = 0.0027) were the only physiological measures which were significantly sensitive to workload. Eyeblink was found to decrease as workload increased from low to moderate, then the relationship reversed and eyeblink increased as the level of workload increased from moderate to high workload. As for the other measure, fixation times increased as workload increased. However, fixation times were only sensitive enough to discriminate low workload from high workload and medium workload from high workload (Wierwille, Rahimi, & Casali, 1985). Though these studies have demonstrated less than desirable results concerning physiological workload measures, many other investigations have reported quite successful results utilizing various physiological indices of workload.

One of the most well known and practical physiological measures of workload is heart rate, and its derivatives, such as heart rate variability and heart period. Heart rate measures provide an overall index of arousal due to task demands. This measure has proven sensitive to workload in many reports over the past three decades (Hancock & Caird, 1993). Critics of this have leveled claims that daily fluctuations in metabolic demands and individual emotions confound the use of heart measures as indices of workload. Claims concerning the increase in heart rate due to increased metabolic demands were negated by measuring oxygen consumption rates and correlating changes expected due to physical demand (Blix et al., 1974). However, heart rates measures are susceptible to fluctuations in emotion, stress, anxiety, task difficulty, or practically any other factor that would raise an individual's arousal level (Sammer, 1998). To circumvent these confounds, a more sensitive measure utilizing heart rate was developed. Heart rate variability (HRV) is based on the fact that the degree of irregularity in normal heart rhythm decreases as mental workload, or the difficulty of a task, reportedly increases (Luczak & Laurig, 1973).

Traditionally HRV is measured by spectral analysis and quantification of the heart's inter-beat interval (Jorna, 1992). The spectrum is then broken up into three frequency bandwidths: 0.01 – 0.06 Hz, 0.07 – 0.14 Hz, and 0.15 -0.50 Hz (Mulder, 1992). The most important of these three being the middle band, as research has shown that a small peak at 0.1 Hz of the spectrum indicates an increase in mental workload (Sammer, 1998). As mental workload continues to increase and begins to overtax an individual's capabilities, as was demonstrated in a multi-task setting, a decrease in spectral power at the 0.1 Hz frequency is observed (Jorna, 1992).

There are problems that have been noted concerning the internal validity of the use of spectral analysis and the measurement of heart rate variability. There is a strong confounding dependence of HRV on heart rate and respiration. As previously mentioned, heart rate is sensitive to any factor which can raise an individual's arousal level. Therefore, indirectly, heart rate variability is also confounded by these same factors (Sammer, 1998). Heart rate and heart rate variability also covary strongly with respiration rate. Studies have shown that breath holding immediately initiates a reflex that reduces heart rate and increases heart rate variability. Therefore, most studies involving heart rate or HRV also collect respiration activity data, as a control variable for the spectral analysis (Backs, 1998). A further problem relating the application of the two measures to flight tasks. Spectral analysis of heart rate data requires at least 5 minutes of constant, consistent task performance for averaging, and other calculations (Jorna, 1992). In practical terms, this means a pilot would have to perform a turn, an emergency procedure, or an instrument approach for at least 5 minutes. This is entirely unrealistic. Most events which unfold in aviation begin and end in a matter of seconds, unless an external variable such as weather is introduced.

In reaction to the shortcomings of the heart rate and heart rate variability measures, two extensions to these measures have been introduced, event-related heart response (Jorna, 1992), and heart period (Veltman & Gaillard, 1996). Event-related heart response is very simply a recording of an individual's heart rate while performing a flight simulation with a co-recorded, synchronized mission events timeline, including pilot inputs and occurrence of external problems, which is then compared to a previously recorded baseline heart rate. One study that utilized this method regarding a flight simulation and an ATC datalink condition showed the heart rate slowed at the uplink occurrence. Past studies have shown that this decrease is associated with information uptake, and is usually followed by a subsequent increase in heart rate as information is Assimilated into short-term memory. The event tagging of the heart rate plot (Figure 13**)** is said to facilitate comprehension and interpretation of the heart rate data as a measure of workload (Jorna, 1992).

The second method, heart period, has gained increasing popularity in the literature (Sammer, 1998; Veltman, & Gaillard, 1998; Backs, Lenneman, & Sicard, 1999). The major contributing factor to the popularity of using heart period over heart rate, or heart rate variability, is the reliability of heart period, due in most part to its indifference to changes in respiration (Veltman & Gaillard, 1998). Although studies may disagree on the sensitivity of the measure, with some studies concluding that heart period is highly sensitive to small changes in mental workload (e.g. Veltman & Gaillard, 1998), while others studies show that the sensitivity is no better than heart rate which can only discriminate between low workload and high workload (Backs, Lenneman, & Sicard, 1999). All studies thus far agree that the reliability and repeatability of heart period exceeds that of its cardiac counterparts.

Another physiological measure developed around the same time as cardiac measures incorporates blink rate and establishes a relationship where blink rate decreases as the demand imposed by the visual environment increases (Kramer, Sirevaag, & Braune, 1987). This relationship is not globally sensitive as blink rate has been found to be insensitive to communications demands in certain experiments (Casali & Wierwille, 1983). Blink rate has an additional different relationship established with mediational workload. As the workload associated with mediation, or central processing, increases, an individual’s blink rate will increase proportionally (Wierwille, Rahimi, & Casali, 1985).

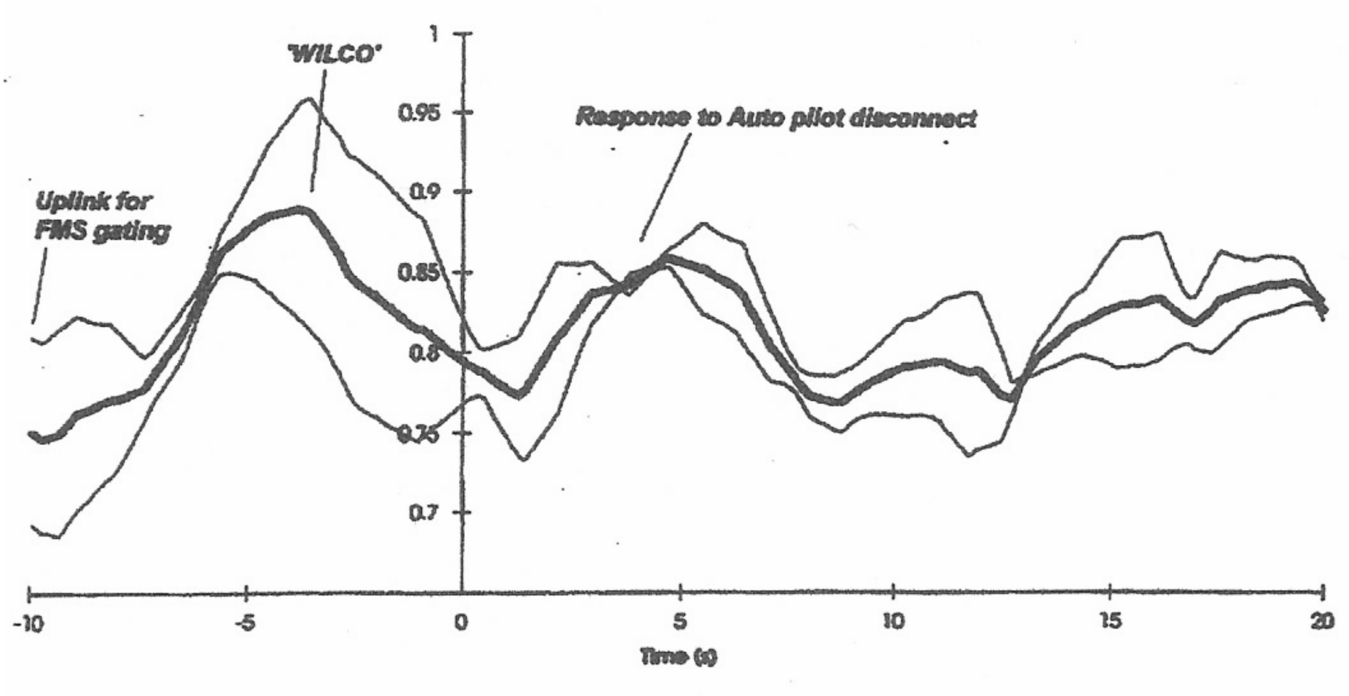


Figure 13. Event-related heart response (from Jorna, 1992).

In studies concerning mental workload imposed on pilots, a successful, well documented physiological measure that is especially relevant to simulated instrument flight is visual scanning consistency. Visual scanning, the direction of pupil gaze, can be used as a measurement of the workload necessary for information extraction in the visual environment, such as the flight information given by the standard six pack of instruments in a light aircraft while conducting IFR flights. Scanning, when coupled with fixations, where the gaze lingers on a certain visual area, also exhibits diagnostic properties, in that the fixations are usually the largest for the most information-rich instruments pointing towards the source of the workload. Furthermore, it has found that novices will fixate for much longer periods than will experts, indicating that workload for novices is much greater than that of the expert (Brown et al., 2002).

Another method that may prove particularly useful as a metric of mental workload is a composite measurement based on characteristics of the human voice. This appears especially attractive at first glance due to the intuitiveness of the underlying theory. The intuitiveness stems from a human's ability to derive information regarding the mental state of another individual based on what the individual is saying, or rather, how they are saying it. Anecdotally, this equates to our day-to-day experience where we can actually hear how the other person is feeling and derive an estimation of that person's mental state, especially if we know the person very well, such as a member of our family. Therefore, the question is, does a correlation exist between the characteristics of the human voice, workload, stress, psychological state, and physiological state.

The first time voice was investigated as a gauge to estimate operator state was in 1965. The Russians experimented with a measurement of the fundamental frequency of a voice to estimate psychological load on cosmonaut A.A. Leonov while he was conducting the first human extravehicular activity (a spacewalk) in space, outside the craft Voschod-2. The Russians continued the experimentation of voice as a measure of cosmonaut state while in space during the 1978 project “Speech 1”. During this project, the voice was broken down into three principal components. The voice pitch, formant frequencies, and intonation length, were studied during different phases of space flight using cosmonaut Sigmund Jaehn as the subject. Significantly higher mean values of voice pitch were reported during different types of physical and psychological stressful situations under the extreme conditions of space travel. These investigations provided some of the first evidence that voice pitch increases while undergoing heightened periods of stress (Johannes et al., 2000).

The results of the cosmonaut studies were later supported by studies performed in the United States. In 1975, researchers and astronauts involved in the Skylab project also investigated the use of voice output as a metric for astronaut psychological state. Workload was induced by increasing task difficulty, and the resulting voice communications were recorded on the ground and analyzed. The researchers found that the fundamental frequency of the astronaut’s voice did increase significantly (Older & Jenney, 1975).

Later studies have identified three vocal or speech characteristics, like the Russian study, which have consistently changed in relation to the amount of workload imposed upon an individual. The first characteristic has already been mentioned, the speaking fundamental frequency. This is one of the most recognized speech indices of workload and stress (Ruiz, Legros, & Guell, 1990; Scherer, 1981). Its increase in frequency in relation to an increase in workload is thought to reflect the resultant physical tension of the laryngeal muscle, the diaphragm, and the abdominal muscles (Johannes et al., 2000).

Speaking rate is another common measure of workload. The underlying theory being that an increase in speaking rate is related to a speeding up of the cognitive processes, and to a lesser degree, the motor processes, to meet an increase in workload demand (Brenner, Doherty, & Shipp, 1994). Lastly, an increase in vocal intensity is often seen in relation to an increase in workload and/or psychological stress. The physiological underpinnings are believed to be linked to the increased thoracic air pressure, again due to the tightening of the abdominal muscles and diaphragm, which occurs involuntarily under stress (Johannes et al., 2000).

Although the majority of studies found significant change in vocal components directly related to imposed workload, a few studies reported no observable changes in subjects' voices as these individuals underwent various physical or psychological stressors (Streeter et al., 1983). The most notable and consistently reported shortcomings of speech components deal with the variance of vocal characteristics both between different subjects and within the same subject. Studies have shown that while the fundamental frequency and intensity will vary according to workload, these variations are not always consistent (Hecker et al., 1968). Sudden changes, or jumps, in intensity and fundamental frequency in addition to rapid fluctuations from one syllable to the next have been observed in air to ground communication tapes during incidents which resulted in fatal consequences. Some have theorized that the seemingly random fluctuations may be the result of loss of accurate control over muscles and breathing during the life threatening, and ultimately fatal situations (Williams & Stevens, 1969). The theory has yet to be empirically substantiated for obvious reasons.

When studying vocal characteristics between subjects, researchers have noted a large variability (Levin, 1975). Due to this fact, results of one study do not generalize from one situation to another. Two main reasons seem to explain most of the variability. First, the fundamental frequency of one individual under workload may be 120 Hz, indicating the individual is in an active attention-laden state, while the very same frequency may be the resting fundamental frequency for another individual. Therefore, a baseline is needed to calibrate the measure for each subject, so he/she can act as their own control. Otherwise, the large intersubject variability may destroy the sensitivity of the experiment. This variability may be the cause of effects found, or rather, lack of effects found in the Streeter et al. (1983) investigation. Another shortcoming hindering the power of vocal characteristics as measures of workload, and their associated experiments, is the inability to generalize results to workload produced by tasks other than those used in the original investigation. The results do not generalize because there is a very specific relation between the vocal output and the workload, which produces the changes in a subject's voice (Ruiz, Legros, & Guell, 1990). In other words, different tasks produce different types of workload, and each type of workload is reflected differently in a subject's voice. At this point, the body of knowledge regarding this topic is not yet mature enough to create a taxonomy of voice changes, workload, and task type relationships. Therefore, each task or combination of tasks must be tested and the results calculated for each specific individual.

Fortunately, the body of knowledge continues to grow with each new study confirming the positive relationship between workload and voice changes. A fairly recent study conducted by Brenner, Doherty, and Shipp (1994) found, while using a simple tracking task, that the degree which vocal components change, while small, is statistically reliable. Vocal intensity, measured as the voice sound pressure level, increased 1 dB as a result of increased workload. The average fundamental frequency of an individual increased 2 Hz, and the average speaking rate increased by 4%. These small changes would not be noticeable in normal conversation; however, they are measurable.

Another interesting study, conducted in seven experimental parts, was performed recently by Johannes et al. (2000). Two experimental parts of particular interest incorporated the effects of psychological workload as compared to physiological workload for each subject. They found a significant increase in voice pitch related to the psychological loading, but interestingly, a standardized physical load produced no significant change in the subject's voice. This first part led to the hypothesis that while under physical workload, other physiological measures of a subject's mental workload would become confounded and rendered useless; however, since voice pitch is unaffected by physical work, it would still be a reliable estimator of mental workload. A field experiment was set up using Austrian soldiers as subjects. Voice pitch and heart rate were compared as measures of mental workload. The soldiers were informed of a dangerous new exercise they had never participated in before, a guerilla slide into a river. Voice pitch increased on the first slide, but by the third slide it had reduced significantly. Heart rate on the other hand increased significantly and remained at a heightened state for the duration of the exercise (Johannes et al., 2000). The only disappointing aspect of this study was that the report only contained a short verbal description of the field experiment. No data or statistical results were reported.

Subjective Measures. Subjective tools such as the Modified Cooper-Harper, the NASA TLX, and SWAT are some of the most effective and reliable measures employed in the field of mental workload. Generally subjective measures are considered to be globally sensitive. Therefore, while these measures possess excellent reliability and transferability from one application to another, they lack diagnostic power.

A total lack of diagnosticity is not entirely truthful concerning the multidimensional NASA TLX and SWAT. They can provide some degree of diagnostic information derived from their multiple subscales (Wierwille & Eggemeier, 1993). To further increase the diagnostic properties of these two scales, researchers have identified applications where the NASA TLX and SWAT are most relevant. SWAT seems to be most sensitive to identifying the cognitive mechanisms that affect mental workload, while the NASA TLX is more appropriate in an applied situation where the extra time needed to complete the survey is available. Also, the NASA TLX is generally considered more sensitive to lower levels of workload than SWAT (Moroney, Biers, & Eggemeier, 1995).

While there are many advantages to using a subjective measure of mental workload, and most researchers will usually include a subjective measure in the experimental framework, there are also disadvantages that must be kept in mind. First, as with any subjective measure, these metrics are easily influenced by context effects and the participant’s own bias, which can change from day to day. Second, administering a subjective test at the end of the experiment has been criticized that participants can not accurately recall that amount of workload they underwent at different points throughout the trial, although this point has been disputed. Moroney, Biers, and Eggemeier (1992) point out that the current evidence suggests that intervals between task and test of up to fifteen minutes may not be critical to the subjective scale's reliability.

Rationale for Chosen Mental Workload Experimental Methods

As suggested by many researchers, a battery of different mental workload tests will be used in this dissertation study. The reason so many tests are needed is to provide a higher measure of reliability, accuracy, and diagnosticity, than any one test can yield on its own. The subjective measure of choice will be the Modified Cooper-Harper. This test has proven itself a valid measure of workload many times over. It is simple, easy to administer, readily accepted by participants, and is especially sensitive to applied settings. A large constraint in this experiment is available time. The experimental trials are already long (3-4 hours) and detailed; therefore, the subjective tool used to measure workload must be administered in the least amount of time possible. This makes the MCH the obvious choice. Also, a large number of previously mentioned primary tasks measures will be collected automatically by the iGATE software.

In addition to those tests, a physiological measure will be included. Voice analysis has been chosen as the measure to be used. Based on previous findings, it seems that the human voice may hold characteristics which will prove to be very useful measures of mental workload. Especially promising is the phenomenon reported by Johannes et al. (2000) revealing that voice pitch, as measured by the fundamental frequency, is insensitive to increased physical workload, which confounds many presently used physiological workload measures. Voice analysis seems particularly appropriate for a simulation based flight experiment that incorporates a large of amount of ATC communications in the scenario design. Voice analysis could prove to be a powerful, transparent measurement tool of mental workload for flight experimentation.

HEADSET COMFORT

It is often said, “The best HPD is the one that workers will wear.” This simple phrase describes a variable with enormous influence on the attenuation characteristics of HPDs, comfort. Results of field research show that discomfort is the number one complaint shop workers have against HPDs (Royster & Holder, 1981). This discomfort often manifests itself in the physical actions of repeatedly doffing and donning a HPD, which has dire effects on the time-weighted noise protection. A HPD with an NRR of 25 dB not worn for 15 minutes out of an eight-hour workday will cause the NRR to be “time-corrected” to that of 20 dB. This results in the HPD not being worn for 3% of the total time, equating to a 20% reduction in the time-weighted hearing protection (Park & Casali, 1991). Therefore, comfort is critical to effective hearing protection in the work environment.

The literature in this area has shown exactly how difficult comfort research and data collection can be. The results of the majority of studies directly conflict with each other making it difficult to rate the comfort of different models of HPDs and headsets according to only a few common parameters.

Although comfort studies have been conducted on different types of HPDS (i.e. earplugs, earcaps, and earmuffs), the studies pertaining to earmuffs are of most interest to this study, as all GA headsets are based on the circumaural earmuff design. If the reader is interested in a more comprehensive review of all types of HPDs and comfort, he or she is directed to Casali et al (1987).

One of the first comfort studies incorporating earmuffs was conducted by Ivergard and Nicholl (1976). Subjects rated different earmuff models based on long-term and short-term wearing periods. A significant correlation was found between the short-term and long-term ratings for some dimensions of wearability and performance, which was interpreted as suggesting that short-term experience could allow users to make valid decisions on long-term experience. This was not the case for the subjective ratings of weight, also referred to as mass, and cushion softness. Weight was perceived as increasing as wearing time increased. In like manner, cushions were perceived as harder following long-term experience. The results pertaining to perceived pressure against the head were inconclusive.

Later studies also investigated pressure against the head, headband force, and clamping or compression force. The results again proved ambiguous (Flugarth & Wolfe, 1971). For example, one study specifically focused on the issues of compression force and pressure. Forty industrial workers evaluated 12 different earmuffs during a workday. They were then asked to choose the one they most preferred based on comfort and performance. The number one choice among workers had the highest contact area, the highest overall force (16.7 N), and the highest cushion pressure (5150 Pa) on the head. The number two choice among the workers also had one of the highest contact areas, but a significantly lower force (13.9 N) and pressure (3470 Pa) on the head.

The researchers concluded that pressure was the most important factor in earmuff comfort, while force and weight were minor considerations (Acton et al., 1976). Looking at the results, these conclusions seem strange as the number one and two choices differed so drastically in pressure. It seems that contact area is more likely the common comfort parameter between the two highly rated earmuffs.

Subsequent studies have also conflicted with the conclusions of Acton et al. (1976). A rating index based on the ratio of attenuation performance to the product of weight and clamping force was analyzed and determined to provide no relationship with user acceptance (Lhuede, 1980). Likewise, Savich (1981) found that clamping force and weight are not significant factors in induced comfort, but as the results of Acton et al. (1976) seem to point to, Savich concluded that cushion contact area was the most significant determinant of comfort. Other researchers concur with these results, finding almost no correlation between earmuff mass and comfort, and little relation between force or pressure and comfort. Furthermore, they conclude that high cushion compliance and contact area are the most important influences on comfort and user acceptance (Damongeot et al., 1981; Berger & Mitchell, 1989).

More detailed studies into headset comfort have utilized a bipolar rating scale to rate comfort across different variables. In 1990, Casali and Grenell examined the variables of headband compression / average pressure, cushion type, and wearing time / work activity. This time, however, the variable of headband pressure was broken down into three levels, high pressure (24.4 N), medium pressure (16.1 N), and low pressure (14.4 N). The other two variables, cushion type and wearing period, consisted of two levels, foam and liquid-filled, and 25 minutes and 75 minutes, respectively.

Analysis of headband pressure showed that only the level of high pressure resulted in significantly higher discomfort and lower acceptance. Medium and low pressure levels showed no significant relationship with comfort. These results could be a possible explanation for the discrepancies in past studies concerning pressure and earmuff comfort. The foam cushion type was consistently evaluated as more comfortable and acceptable than the liquid-filled cushions. However, in no case did the difference between liquid-filled and foam prove statistically significant. Analysis of the final variable, wearing time, showed that the high and medium headband pressures were significantly more uncomfortable after 75 minutes of wearing time than 25 minutes. This difference did not carry over to the low pressure level. However, when wearing time was analyzed for user acceptance, only the high pressure level became increasingly unacceptable over a longer wearing period.

A study reported one year later used the same bipolar rating scale. Although the objectives of the study were aimed more at a comparison of laboratory versus field-testing results and a second comparison of different types of HPDs, some of the results and conclusions have direct application in this dissertation study. First, the results further validated the sensitivity of the bipolar rating scale to the perceived comfort of an earmuff. Secondly, the results of the analysis of wearing time corresponded with those of the last study mentioned. Park and Casali (1991) found that perceived comfort of an earmuff degraded over a two hour wearing period. This discomfort was reported to be localized to the pinnae and the flesh surrounding the pinnae. They surmised that the discomfort was the cumulative result of headband pressure and heat buildup under the earmuff over the two hour wearing period.

Rationale for Chosen Headset Comfort Experimental Method

Although the results of some studies do tend to contradict each other, there seems to be certain variables, which consistently contribute to a user's perceived HPD comfort. Contact area and cushion compliance are generally agreed to be the most important influences on earmuff comfort. Pressure also seems to be influential at higher levels, especially over longer wearing periods.

The method which will be used to evaluate the aforementioned variables is the bipolar rating scale. This type of scale has already been validated in previous studies. Additionally, the scale has been shown to be sensitive to comfort parameters in studies incorporating variables of considerably different contexts. The use of the bipolar rating scale also has practical advantages to this study. It is easy to use and requires no training and only a minimal amount of instruction. Overall, the bipolar rating scale is a sensitive, effective, and time efficient subjective method. These characteristics will be very important to a study which already requires long subject participation time.

RESEARCH OBJECTIVES AND HYPOTHESES

From the preceding literature review, it is quite obvious that the potential theoretical advantages of active noise reduction have been established (e.g. increased low frequency attenuation, slightly improved speech intelligibility in certain situations, improved task performance in certain situations). However, with few exceptions (Gower & Casali, 1994; Urquhart, 2002; Lancaster, 2004) research attention has overlooked testing ANR in any applied environment. More testing to quantify the actual benefits ANR can provide is sorely needed, especially in high-risk, noise intensive environments such as that of general aviation. Light, single-engine, propeller driven aircraft have been empirically shown to produce some of the most potentially hearing perilous environments, with noise regularly peaking above 105 dB. When occupations such as flight instructors and FAA flight examiners, require the pilot to work in this environment 8 hours per day, 6 days per week, the highest quality, highest performance hearing protection headset must be a top priority; even then, some experts will likely argue the noise still is so overpowering that hearing loss is inevitable.

Another major problem exacerbated by the high intensity noise is that of unintelligible radio communications. The aviation system is completely reliant on the time critical exchange of information between an air traffic controller and a pilot, and also between a pilot and another pilot. Therefore, the comprehension of every radio message a pilot receives is of the utmost importance. ANR attenuating characteristics are most effective at the low frequencies of the audible spectrum. Fortunately, the spectrum of cockpit noise is also centered on the low frequencies. ANR should be more effective at attenuating cockpit noise then passive devices, which are most effective for midrange and high frequency noise. Furthermore, ANR attenuation of low frequency noise should prevent, or more likely reduce, the upward spread of masking, thereby increasing the speech-to-noise ratio and overall speech intelligibility. This could possibly be a tremendous advance in technology, aiding all pilots, from students to Boeing 777 airline captains.

Finally, a number of studies conducted by Whitaker and associates have demonstrated a relationship between task difficulty, speech intelligibility, performance, and workload. Their findings show that speech intensive complex tasks, or multi-tasking, are more susceptible to the adverse effects of decreased speech intelligibility. These adverse effects manifest themselves as a marked increase in mental workload, and a dramatic decrease in task performance. Thus, as speech intelligibility decreases, mental workload almost immediately begins to increase, followed shortly after by a sharp decrease in performance, so much so that mission success is often compromised.

This investigation sought to fill in the gaps to determine what, if any, benefits can be realized by using an ANR headset in the aviation environment. Specifically the research aims of this dissertation were:

1. To determine the differences between the attenuation of a passive hearing protection device (HPD) with an integrated aviation communications system, and that of several active noise reduction (ANR) headsets, also with integrated aviation communications systems. The attenuation characteristics of these

headsets were tested in the low-frequency dominant noise environment which typifies that of a propeller-driven aircraft cockpit.

2. To determine the impact that ANR headsets have on the speech intelligibility of aviation radio communications. Using the Speech Transmission Index the study examined the possibility that ANR could increase speech intelligibility through increased attenuation of low frequency noise and thus a release from the upward –spread of masking.

3. To determine the impact that ANR headsets have on a pilot’s communications workload. Using multiple, varied measures of workload, flight performance, and speech intelligibility, the project examined the possibility that ANR could reduce workload and increase pilot performance.

4. To determine and compare the protected exposure limits (PEL) of the ANR headsets utilizing measurements on an acoustic test fixture in the same propeller-driven aircraft cockpit noise as used in the experiment.

5. To collect and compare pilot participant opinions regarding the comfort, performance of the communications systems, and relative rankings of the passive and active attenuating headsets, following a three-hour cross-country flight simulation.

Along with these specific research objectives, several hypotheses were put forth to be empirically evaluated. These hypotheses were:

1. Significant differences in the attenuation performances of the four headsets were hypothesized, as measured by spectral attenuation via microphone in real ear (MIRE) and the protected exposure levels (PEL). The ANR headsets' performances exceeded that of the passive headset below 500 Hz and the performance of the passive headset exceeded that of the ANR headsets in the region of 1000 – 2000 Hz, where ANR devices typically exhibit amplification.

2. A significant increase in the pilots' speech intelligibility performance resulted while wearing the ANR headsets over the passive attenuating headset. This was most likely due to ANR's higher attenuation performance for low frequency, high intensity noise.

3. The increased speech intelligibility manifested as a concomitant decrease in measured mental workload and/or increase in flight task performance.

4. Using a measurement of the clamping force (per ANSI S3.19-1974), it was hypothesized that those headsets with lower clamping force (measured in Newtons) will be rated and ranked as more comfortable then those of higher clamping forces. These ratings and rankings were derived from the pilots' experiences with the headsets in long – duration, continuous flight. Based on this rationale, the ordering was expected to be:

    1. Bose (6.5 N)
    2. David Clark (9.7 N)
    3. Sennheiser (10.13 N)
    4. LightSPEED (11.28 N)

5. Subjective surveys and ratings showed pilots chose the ANR headsets as ranked higher than the passive headset in overall comfort, attenuation, and communications performance.

6. A subjective rating scale showed pilots felt less fatigued after finishing the flight scenario with the ANR headsets, as opposed to the passive headset.

7. Headsets which offered increased speech intelligibility, comfort, communication system quality, and protection levels (attenuation) were associated with higher levels of performance during the objective flight tasks.

METHODOLOGY

Experimental Design

The design of the investigation was a 4 x 4 x 3 completely within-subjects experiment with all levels of a communications workload variable nested within every cell of experimental design matrix (Figure 14). The study consisted of three independent variables and five categories of dependent variables, with some categories consisting of multiple dependent measures. The fully within-subjects design was chosen to make use of each participant as his own control. This methodology has met with success in past workload studies and headset comfort studies (Casali & Grenell, 1990; Casali & Wierwille, 1983; Wierwille & Connor, 1983; Casali & Wierwille, 1984; Wierwille, Rahimi, & Casali, 1985), increasing the sensitivity of the experiment due to relief from the high levels of between-subject variability likely to occur in a between-subjects, or in a mixed-measures design. Furthermore, a within-subject design is especially important to a study involving pilots in high workload scenarios. Pilots are encouraged in training to develop strategies and procedures to cope with the rigors of flight. These individualized, idiosyncratic strategies often differ widely from pilot to pilot, and can border on eccentricism.

One simple example is the immediate response to an urgent ATC command, such as (aircraft call sign), immediate left turn, heading 040, expedite! Some pilots, including the author, will immediately initiate the turn upon hearing the command, and then once stabilized in the turn, respond with a readback to ATC. However, other pilots will

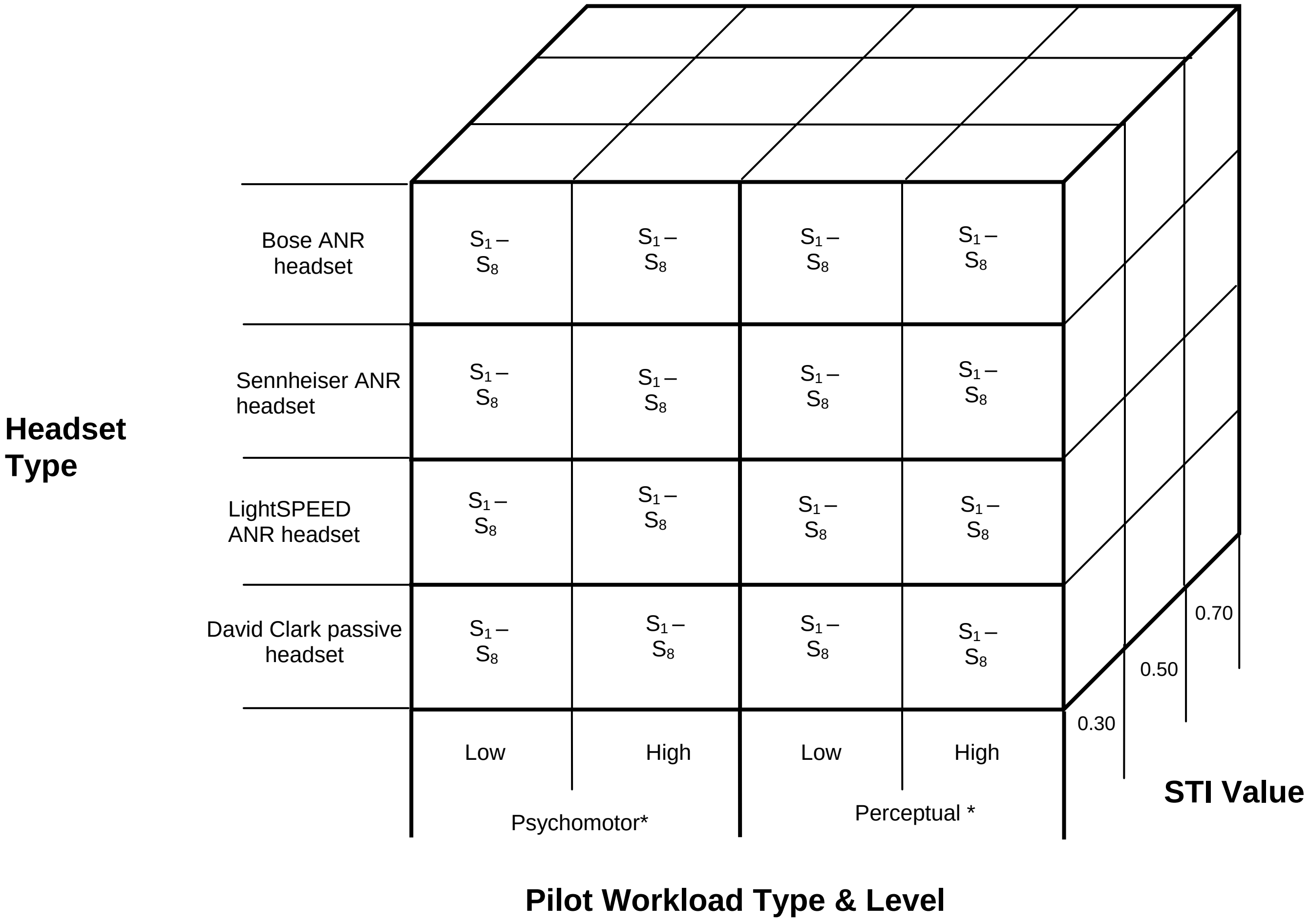


Figure 14. Experimental design block diagram. All cells of the matrix contain a nested communications workload variable with three levels representing low, medium, and high communications workload.

respond immediately with a readback, and upon completion of the readback initiate the left turn. This is a simple individual difference which could have dire effects in a between-subjects design.

Independent Variables

Headset. As previously stated, three independent variables were manipulated in this study. The first variable included four levels of different aviation communication headsets. A passive attenuating David Clark 13.4 headset, a Bose Aviation X ANR headset, a Sennheiser HMEC300 ANR headset, and a LightSPEED Thirty 3G ANR headset comprised the four levels of the headset variable, as depicted in Figure 16. These four headsets were chosen as prevalent headsets in aviation which would allow for comparison between passive attenuation and ANR, in addition to comparison between the individual headsets. Each of the four headsets were worn, one at a time, through one entire 3.5 hour instrument cross-country flight scenario. Therefore, four flight scenarios were needed in this experiment, one per headset. The volume at which ATC communications were played over each headset was controlled across headsets as explained in the later section entitled ‘Speech Intelligibility Under Headset Calibration’ (See Figure 15 for photo depictions of the four headsets).

Workload. The second independent variable manipulated in the study was that of workload type and level. Workload was divided into three types, psychomotor, perceptual, and communications workload. It was determined that these three types of workload were the primary types of workload experienced by pilots throughout flight.

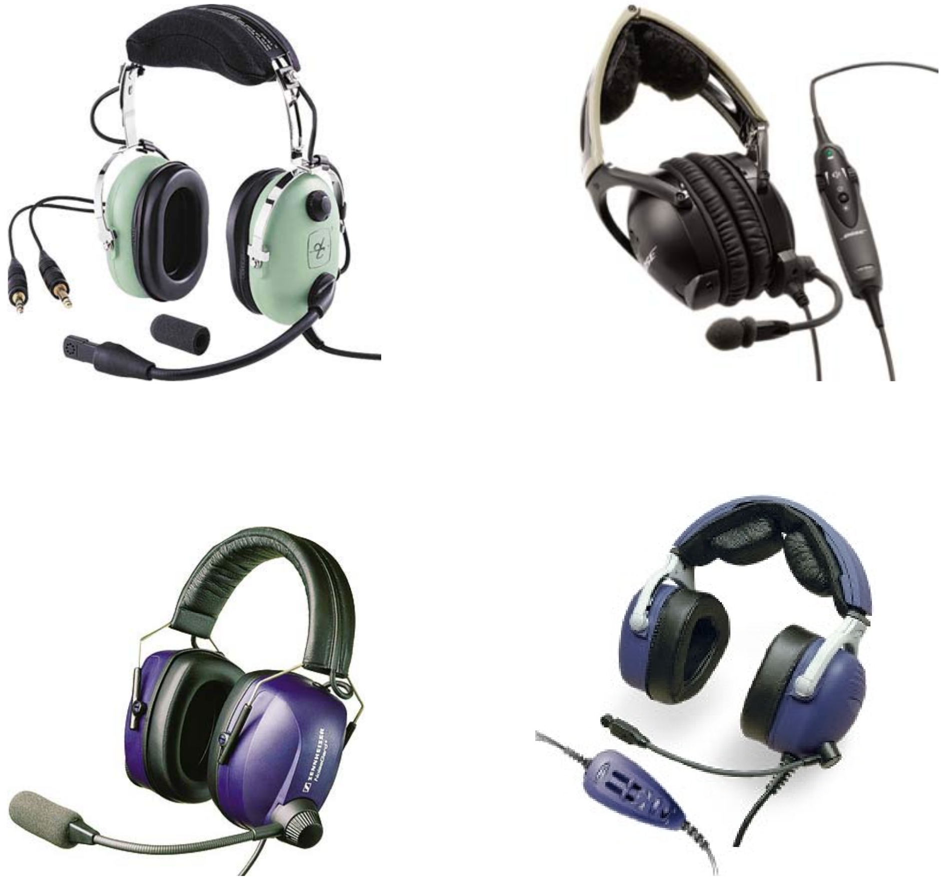


Figure 15. Four headsets used in this experiment, David Clark 13.4 (top left), Bose Aviation X (top right), Sennheiser HMEC300 (bottom left), LightSPEED Thirty 3G (bottom right).

Psychomotor workload focuses on motor skills, such as hand-eye coordination, and is experienced by the pilot as the flight control loop. Within the flight control loop, the pilot evaluates a given situation, determines the proper control input to make, and manipulates the control yoke; then the loop starts all over again. In the case of this experiment, psychomotor workload was manipulated by combining the level of difficulty of a flight task with the severity of the environmental conditions. In this manner, the manipulations of the flight yoke increased for the more difficult flight task, in combination with the more severe environmental conditions causing the aircraft to become less stable.

Perceptual workload focuses on awareness of what flight situations are transpiring. Perceptual workload was manipulated by increasing or decreasing the frequency of instrument and aircraft system failures. In the case of this experiment, the pilot had to recognize each specific hazard by reporting the hazard over the simulated aircraft radio. In this manner, the pilots' perceptual workload increased by having an increasing number of system failures to perceive and identify while continuing the flight simulation. Also, to prevent a pilot from completely ignoring his flight control responsibilities and allocating all attention to failure identification, light turbulence was added into the perceptual workload flight modules.

Communications workload was essential to many of the headset-related research questions. Therefore, all levels of communications workload were included in differing orders as a nested variable in all flight tasks regardless of psychomotor or perceptual workload. In other words, all levels of communications workload were presented in varying order, to counterbalance for order effects, to all participants across all experimental treatments, throughout the four flight sessions in which pilots were expected to participate. Due to the fact that all levels of communications workload were spread across every cell (treatment) in the experimental design, it was not possible to separate the levels of communications workload and measure their effects. Therefore, an analysis of communications workload will not be presented in this study.

Communications workload levels were manipulated according to Wickens & Hollands Information Theory (1999), which states that task difficulty can be determined by the amount of information processed per event. Therefore, by varying the amount of information contained in each ATC transmission the level of communications workload per ATC command can be controlled. In the ATC transmissions, low communications workload was achieved by including only one piece of information (e.g. Turn to heading 280). Moderate communications workload was achieved by including three pieces of information in the ATC command (e.g. Depart the hold on the 315 degree radial, descend and maintain 6000, increase speed to 110 knots). High Communications workload was achieved by including five pieces of information in the ATC command (e.g. Hold NE of the Rosewood VOR on the 030 degree radial, left turns, expect further clearance in 20 minutes).

The types of psychomotor and perceptual workload were further divided into two levels each, low workload and high workload. Workload was manipulated using flight task modules, such as vectoring, an ILS approach, a holding pattern entry, etc. Communications workload was nested as previously stated and was further divided into three levels, low workload, moderate, and high communications workload (See Table 1 for workload conditions of all levels). The difference between the subdivisions of psychomotor, perceptual, and communications workload was done based on the literature

Table 1.

Flight Tasks and Environmental Conditions for All Levels of Workload Variable.

Workload Type & Level	Flight Task Completed	Wind Direction & Speed	Cloud Cover	Turbulence Level	Nested Communications Workload Levels Included
High Psychomotor	ILS approach & Holding Pattern	10 knots perpendicular to final approach flight path or holding pattern entry flight path	Instrument Meteorological Conditions (flight through unbroken cloud layer)	Severe	High, Moderate, Low
Low Psychomotor	Vectoring	3 knots direct headwind to 30 degrees from flight path	Instrument Meteorological Conditions (flight through unbroken cloud layer)	Light	High, Moderate, Low
High Perceptual	One instrument and/or system failure every 5 seconds	No Wind	Instrument Meteorological Conditions (flight through unbroken cloud layer)	Light	High, Moderate, Low
Low Perceptual	One instrument and/or system failure every 50 seconds	No Wind	Instrument Meteorological Conditions (flight through unbroken cloud layer)	Light	High, Moderate, Low

(e.g. Casali & Wierwille, 1983), as well as for practical reasons. The practical reason was that the number of flight tasks needed per STI value would exceed the practical limit of a three and one-half hour flight simulation. If this were to occur, it was believed that subject attrition would endanger the completion of the study. The three levels of the communications workload variable were kept as their inclusion was of primary interest to the headset-related questions of the experiment, and all attempts to optimize its sensitivity were made.

Using this structure, the workload type, level, and all details were included in a flight task module. It was, in essence, self-contained. Therefore, the order of flight modules were changed between the four flight scenarios flown by each subject, so that the flight modules were counterbalanced in order to control for order effects which could contaminate data and skew results. Only differences in heading, altitude, and airspeed assignments were allowed between flight scenarios to accommodate different flight routes.

Speech Intelligibility. The third independent variable of the study was the communication Speech Transmission Index value. The Speech Transmission Index (STI) was chosen for the reasons detailed in the previous section entitled ‘Rationale for Chosen Speech Intelligibility Experimental Method’. Three STI values were administered at the levels determined by subject matter experts (a step-by-step procedure is given in the following section entitled ‘Equalization of Speech Intelligibility Under Headset’). These STI values were applied to the ATC commands and information queries given to the pilots as they flew through the various flight modules during the cross-country flight simulations. The STI values were achieved by varying the amount of additive and multiplicative white noise integrated with pre-recorded speech mimicking the sound of an aviation radio. A preset level of aircraft piston engine noise, representative of a Cessna 172 cockpit noise spectrum, was maintained throughout each flight scenario.

Dependent Measures

Primary Task Performance. There were six dependent measures which were collected as primary task metrics to measure the effects of the independent variables on performance during the flight modules. These six variables directly measured the pilot's flight performance and included magnetic heading, altitude, indicated airspeed, vertical speed, localizer track, and altimeter pressure setting. The iGATE flight simulator automatically collects this data approximately 32 times per second and stores it in a file on the hard drive of the experimenter station computer. These data were then reduced to one observation every 10 seconds for later statistical analysis.

It should be noted that of the six performance variables, altimeter pressure setting was considered to be a practically significant metric for instrument flight, but has not previously been utilized in this manner. It was considered practically significant because it required perceptual resources to comprehend the pressure setting level and make necessary changes, which could be affected by workload. The effects of a performance deviation in this measure are especially important to instrument flight as the deviation means a difference exists between the indicated aircraft altitude and the actual aircraft altitude, a potentially dangerous situation. Again though, this measure has not been previously utilized in this regard and is therefore considered to be an exploratory method.

Physiological Workload Measure. The experiment also incorporated a physiological measure of mental workload, voice analysis. This measure was not included as a dependent measure because the study required redundant measures of workload. On the contrary, for the purposes of workload measurement in this study, the Modified Cooper-Harper workload rating scale was sufficient by itself. Voice analysis was included as an exploratory method of workload analysis because this potential tool could be used as a concealed method of workload data collection. However, it has never been applied to an aviation or simulation environment, and is again considered to be an exploratory workload analysis technique.

Voice recordings were easily made by recording the readbacks of the ATC commands made by the pilots during the experimental sessions. Voice analysis has shown in a few studies to be an effective measure of workload without suffering from the confounds that prove problematic for other physiological measures of mental workload. Therefore testing and possibly validating this technique was of interest in this study. The measures of voice formant frequency and amplitude were derived from the readback recordings using Praat software (Boersma & Weenink, n.d.).

Praat software was developed by two professors, Paul Boersma and David Weenink, of the Department of Phonetic Sciences at the University of Amsterdam, Netherlands. It is a powerful, open source computer program that takes the digital signal from a computer port and converts it into numerical and graphical output. The same analysis of recorded sound files is also possible. This software easily derives the formant frequency and amplitude of recorded speech (Boersma & Weenink, n.d.).

Subjective Workload Measure. Subjective measures of workload were collected using the Modified Cooper-Harper Scale (Appendix G). This rating scale was administered immediately following each flight module in the flight scenarios. To allow the pilot to accurately complete this survey, the simulation was paused immediately following the completion of a flight module. This way, the pilots could devote their undivided attention to the Modified Cooper-Harper Scale. A short pause and simple measurement query during transitions between modules was believed to not disrupt the pilot's performance in any way, as has been demonstrated in previous flight simulation studies concerning situation awareness (Endsley, 1995).

Subjective Ratings and Rankings. Comfort data concerning the various headsets being tested were collected at the end of each flight session. Pilots were given a comfort rating scale (Appendix H) requiring them to rate aspects regarding comfort according to respective bipolar descriptors. This scale was developed by Casali and Grenell (1990), and had already been validated for earmuff comfort studies. It was modified to include dimensions which reflected the communication characteristics of each headset. The last three ratings measured topics of pilot fatigue, the realism of the simulation engine noise, and the realism of the actual simulation. At end of the entire experiment, (after experiencing all headsets, one per session) pilots were asked to rank order the four headsets according to overall communications performance, comfort, and noise reduction performance.

Protected Exposure Levels. The Protected Exposure Levels (PELs) under each headset were measured as specified by the procedures within ANSI 3.19-1974 for the use of an acoustical test fixture (conformed to ANSI S12.42-1995) in the noise (Cessna 172 engine noise) and measuring the octave band noise levels between 63Hz and 12.5 kHz, with and without the headset using a Larson-Davis 3200 series real-time spectrum analyzer and a 1 inch precision microphone (Larson-Davis model 2575, serial #1280).

Participants

Qualified participants were recruited by advertising the experiment at local centers of aviation (e.g., flight clubs, airports, flight schools). Of the ten qualified applicants (nine male and one female) who responded, eight were able to fit the experiment into their daily schedules. Thus, eight male, instrument-rated private or higher licensed pilots were used as research participants. It is believed that the study used all male pilots because the aviation industry (especially the pilot profession) is still heavily male-dominant, and therefore the odds of recruiting a female pilot are low unless the experiment were conducted near one of the large, commercial aviation airports.

Five of the pilots held private pilot's licenses with instrument ratings; two of these pilots also held multi-engine ratings. Two pilots were certified instrument flight instructors, and one pilot held an airline transport pilot's license. Experience in the aviation system is generally gleaned by the number of flight hours a pilot has accumulated. This tradition was continued in this investigation. The average experience of the eight pilots was 2392.5 total flight hours. Total flight hours were then broken down into specific types of flight hours relevant to this experiment (e.g., instrument hours). The pilots had an average of 313.3 instrument flight hours, which included both actual and simulated instrument flight. Further, they had an average of 1020.1 flight hours in any non-complex aircraft, and an average total of 621.8 flights hours in a Cessna 172 light aircraft (See Table 2 for break down of each participant's flight experience).

Pilots were paid $20.00 per hour for their participation. Each pilot was screened for an instrument-rating, current medical certificate, and their log book was scrutinized to ensure that they met Federal Aviation Regulations (FAR) Part 61.57 for instrument currency. This was done to ensure that the participants had the skills necessary to fly in high workload instrument meteorological conditions (IMC). If private pilots without instrument ratings were allowed to participate in the simulations, their lack of proper training and experience would undoubtedly artificially raise the workload ratings and flight performance errors, masking any true data of interest. Pilot participants also underwent a pure-tone audiogram to establish that their hearing met the minimum criteria set forth by FAR Part 67.205 and 67.305 (see Table 3 for the criteria).

Apparatus

All audiograms were conducted first using a Beltone Model 119 audiometer and an Industrial Acoustics Company, controlled acoustical environment portable audiometric booth. However, the main piece of equipment used for this experiment was Virginia Tech's Fly Elite iGATE-PC-ATD. This flight simulator (Figure 16) was specifically designed to be a realistic PC-based aircraft training device (ATD). The level of realism and fidelity achieved by the simulator is evidenced by the fact that the Federal Aviation Administration has certified that pilots can log instrument flight time accumulated through flights on this fixed-base simulator. The simulator apparatus was connected to a Dell Dimension 4300 (experimenter's station; Figure 17) in the adjacent

Table 2.

FAA certificates and ratings and break down of flight experience.

Participant	Highest Certificate Attained	Ratings Attained	Total Flight Time (hours)	Total Instrument Flight Time (hours)	Total Single Engine Non-Complex Flight Time (hours)	Total Cessna 172 Flight Time (hours)
1	Certified Flight Instructor - Instrument	Instrument, Multi-Engine	6360.0	730.0	4880.0	3190.0
2	Private Pilot	Instrument	215.0	40.0	215.0	70.0
3	Private Pilot	Instrument, Multi-Engine	337.0	56.0	262.0	188.0
4	Airline Transport Pilot	Instrument, Multi-Engine, BE-1900, B737	10860.0	1479.0	1600.0	800.0
5	Private Pilot	Instrument	275.0	54.0	275.0	275.0
6	Certified Flight Instructor - Instrument	Instrument	470.0	62.0	320.0	5.0
7	Private Pilot	Instrument, Multi-Engine	220.0	55.6	209.3	116.4
8	Private Pilot	Instrument	403.0	30.0	400.0	330.0

Table 3.

Minimum hearing thresholds as specified by FAR Part 67.205 and 67.305.

Frequency (Hz)	Hearing Threshold in Better Ear (dB)	Hearing Threshold in Poorer Ear (dB)
500	35	35
1000	30	50
2000	30	50
3000	40	60

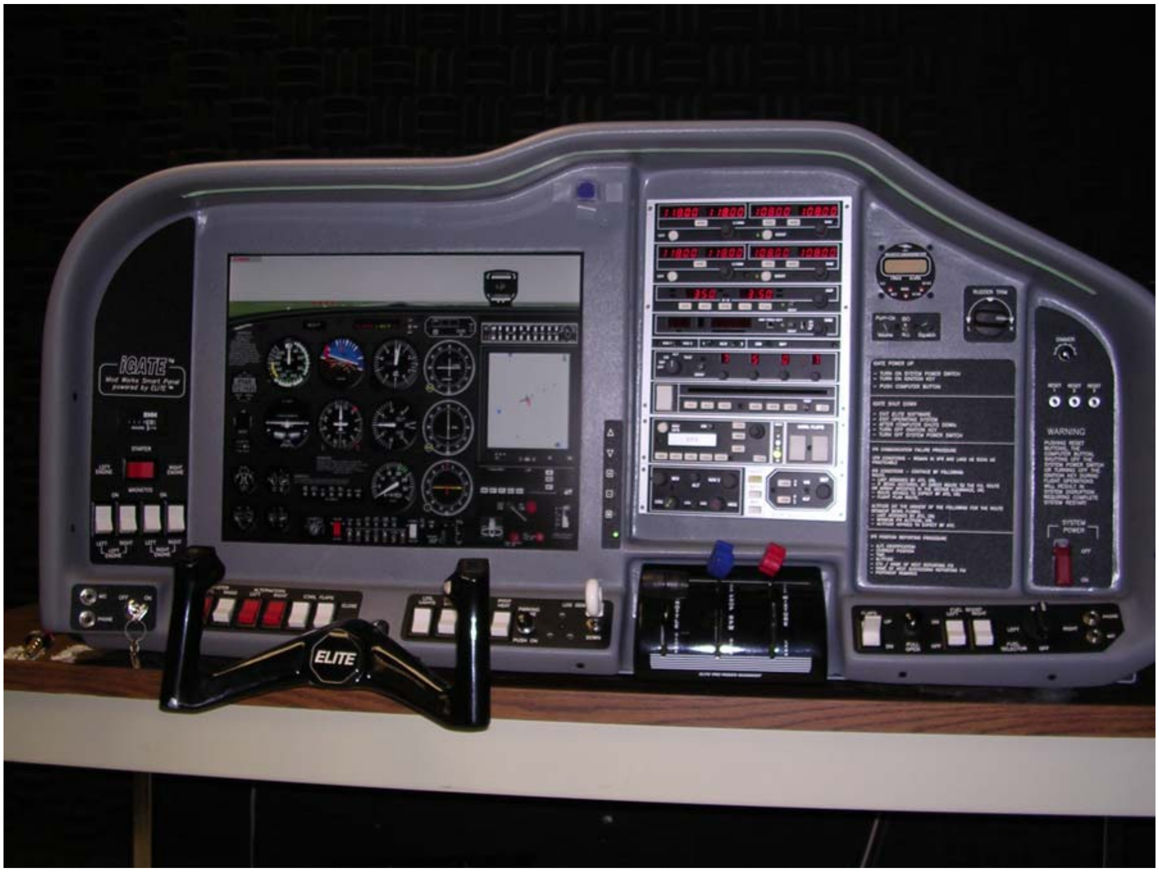


Figure 16. Virginia Tech's Fly Elite iGATE-PC-ATD flight simulator.

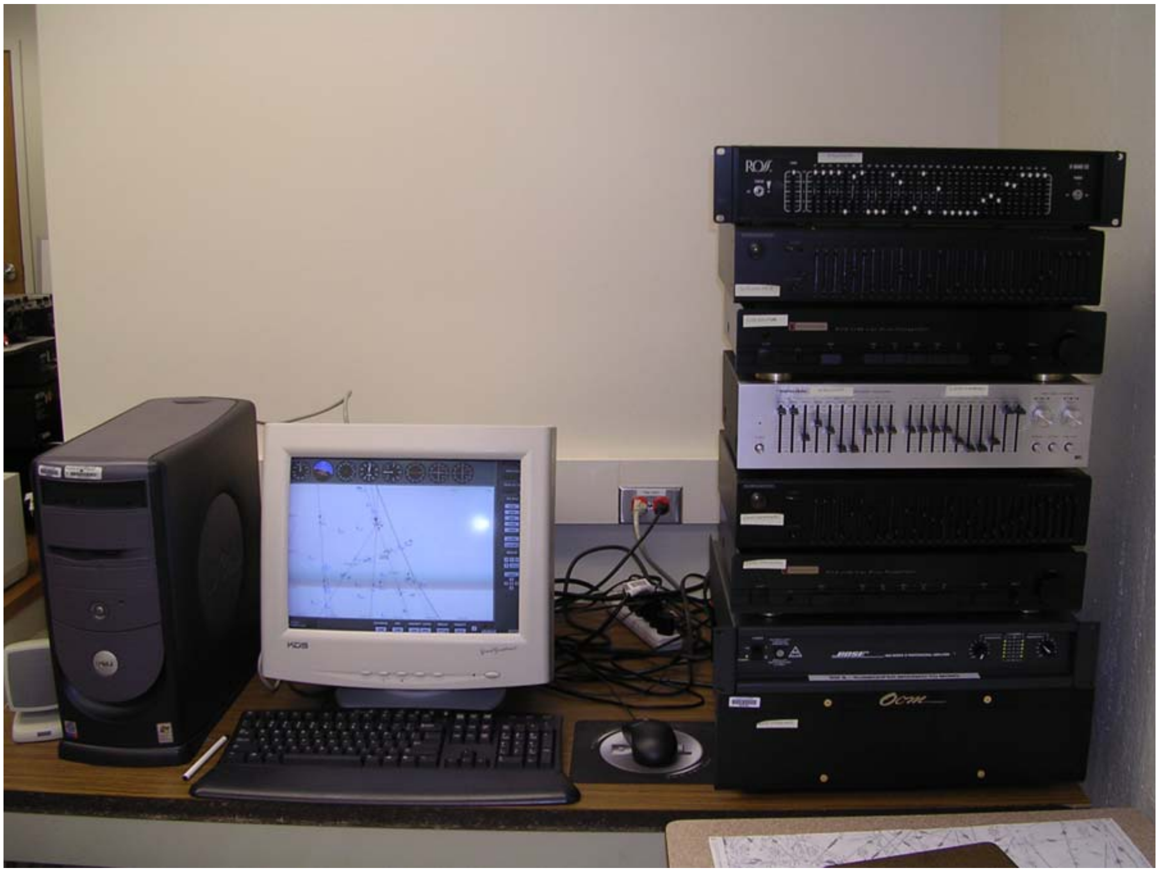


Figure 17. The experimenter's station for the iGATE flight simulator, and the stereo system controlling the speaker system in the simulation room.

room. The experimenter’s station has the capability to measure 65 different flight parameters which are automatically saved to the hard drive after each flight. Additionally, the experimenter’s station controls every aspect of a flight, such as the location of the aircraft, meteorological factors, or equipment failures. The flight simulation room was monitored using a Sony DSX-327 video camera and presented on a Sony Trinitron PVM-1341 monitor, located at the experimenter station.

Various Cessna 172 aircraft sound effects, such as engine, gear retraction and flap extension noises were produced by the experimenter station computer and played over two Bose 802 Series II professional loudspeakers and one Bose Panaray System 502B Acoustimass Bass subwoofer (Figure 18). The loudspeakers were driven by a Parasound P/LD – 1100 Line Drive preamplifier, an OCM 200 Series amplifier, and an Audio Control 1/3 octave band equalizer. The subwoofer was driven by another Parasound P/LD – 1100 Line Drive preamplifier, an Adcom Mosfet GFA 5500 amplifier, another Audio Control 1/3 octave band equalizer, and an additional Ross 1/3 octave band equalizer. An additional Realistic Model 31-2000A octave band equalizer was used to further boost the low frequencies of the subwoofer and the high frequencies of the loudspeakers (Figure 17). Calibration of these sound levels was accomplished using a RION NA-29E 1/3 – octave band analyzer to meet the 1/3–octave band levels recorded from a Cessna 172 in actual flight, as explained later (Figure 19).

Pre-recorded ATC commands were played over the four aviation communication headsets using a Dell Precision 380 computer connected directly to the headset. The pilot’s responses were recorded by connecting the aviation headset microphone plug to the iGATE’s integrated intercom and connecting the intercom to a Dell Precision 420

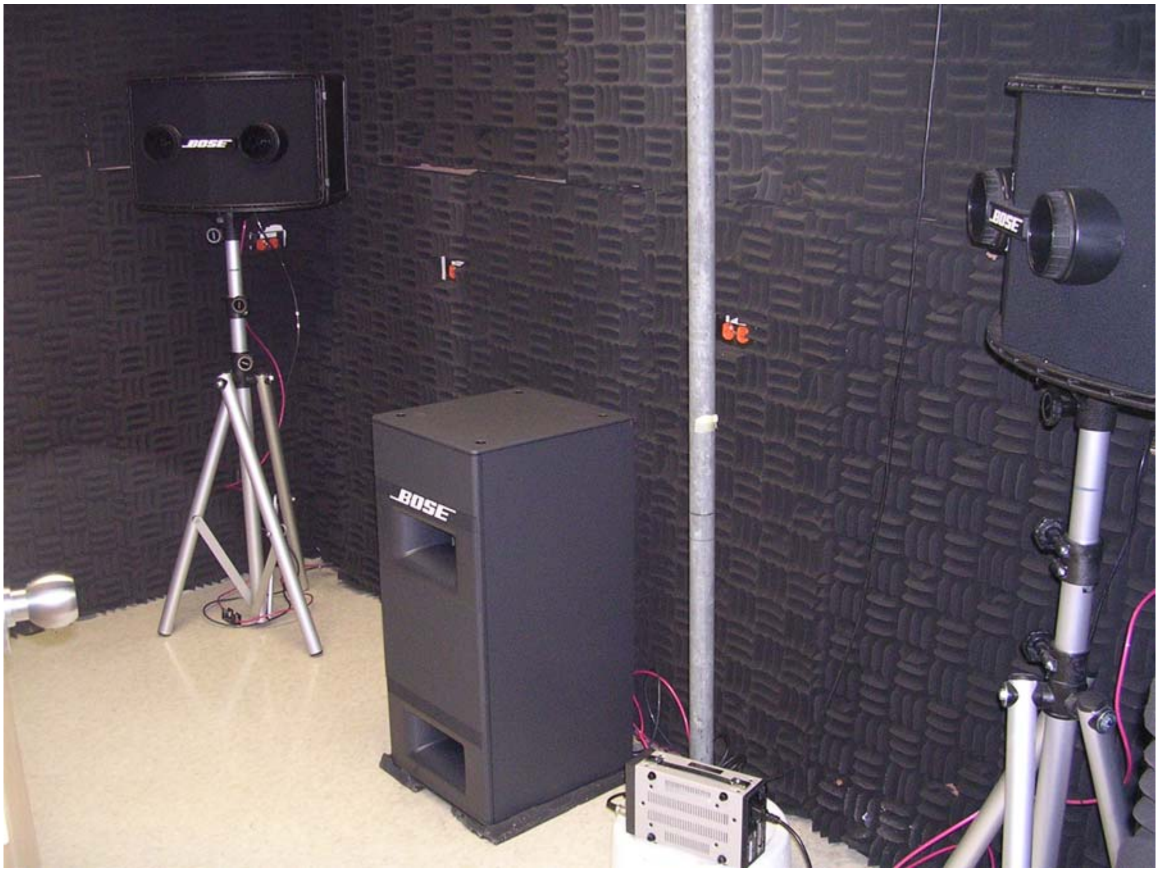


Figure 18. Speaker system which produced the Cessna engine noise inside the flight simulator room.

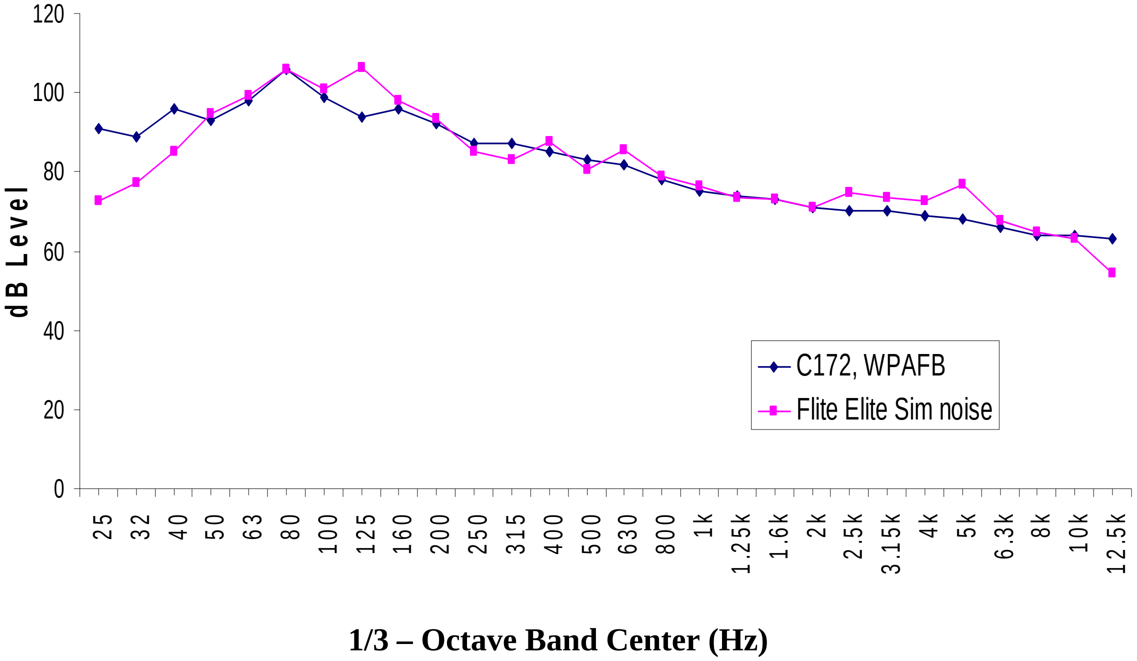


Figure 19. Noise spectra for a Cessna 172 and the Fly Elite simulator during cruise at 110 knots and throttles set at 2300 rpm.

running Adobe Audition. Additionally, the wav files recorded by Adobe Audition of the pilot's responses to the ATC commands were analyzed using the Praat software. The Praat software analyzed the speech contained in the wav file and produced the resulting formant frequency and amplitude.

Lastly, pilots were also allowed to bring any flight equipment they preferred to use, except a communications headset. This equipment included a knee or lap board, an electronic flight computer or slide scale flight computer, a plotter, a timer, or instrument approach plates, if they chose. If they chose not to, the default equipment of a Sporty's pilot kneeboard, a Sporty's electronic E6B flight computer, a Jeppesen instrument plotter, appropriate Jeppesen instrument approach plates, a checklist for the Cessna 172, and a standard digital kitchen timer were supplied. It was preferred (and suggested) that pilots bring the equipment they were already familiar with, as it may have falsely increased workload levels if they were to conduct flight operations while learning to use new types of equipment.

To conduct the protected exposure level without commincations input measurements, the aforementioned sound system was used in addition to an acoustical test fixture (Figure 20) which conformed to ANSI S12.42-1995. The Larson – Davis 3200 real-time spectrum analyzer was used to collect the attenuation data. Protected exposure levels with communications input were collected using a Goldline DSP-30 digital signal analyzer and Bose "swim plug" in-ear microphones.

The headband clamping force of each headset was measured using a Nova Tech Type F241CF00H0 clamping force measurement device, which conforms to ANSI S3.19-1974 (Figure 21).

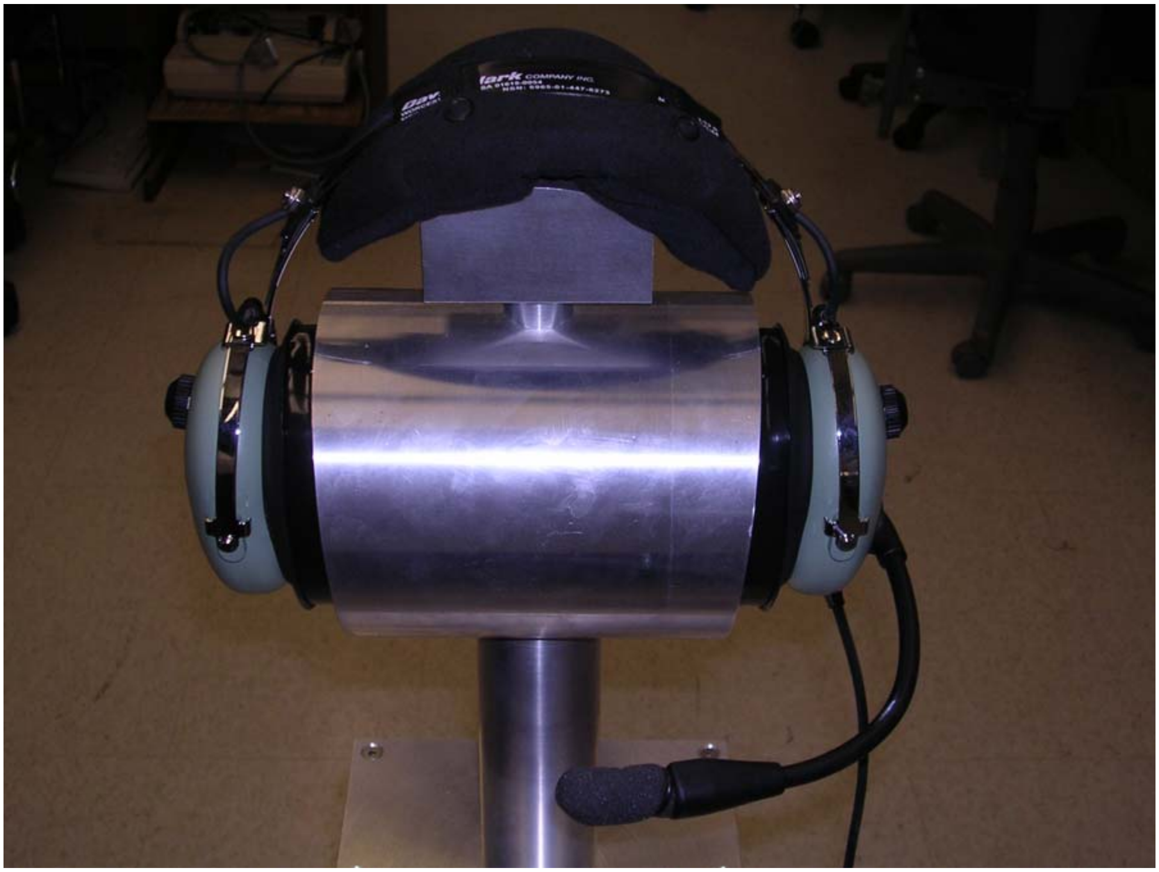


Figure 20. Acoustical test fixture used to measure protected exposure levels without communications (ANSI S12.42-1995).

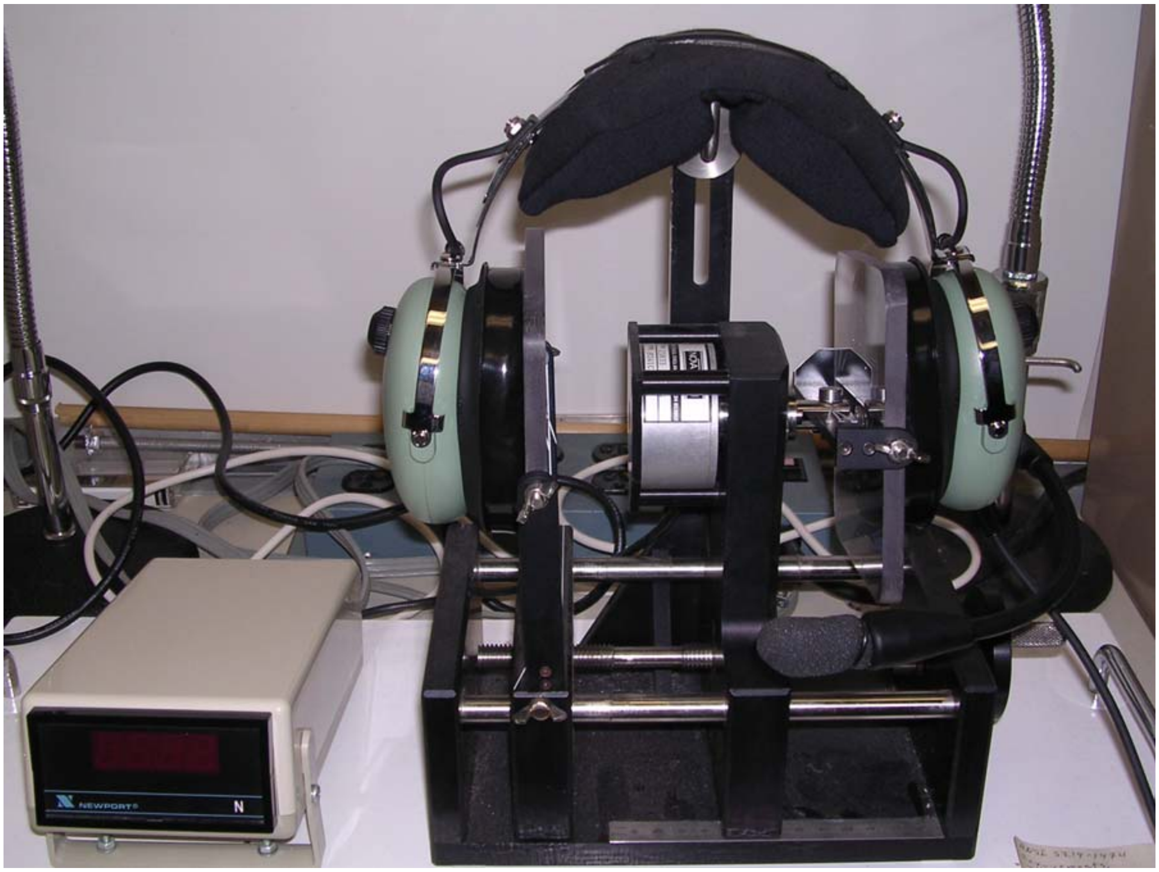


Figure 21. Headband clamping force measurement device (ANSI S3.19-1974).

Procedure

Aircraft Engine Noise Calibration. The Cessna 172 engine noise played by the experimenter station computer over the loudspeakers/subwoofer system was calibrated to that experienced in an actual Cessna 172. Consequently, a 1/3-octave band spectral analysis was obtained of the engine noise as experienced in the cockpit of a Cessna 172 during cruise flight (see Figure 19). The octave band and 1/3-octave band equalizers were adjusted until the spectrum in the simulator room was as close to the actual spectrum as the equipment allowed; the comparison is depicted in Figure 17. The simulator engine noise spectrum, which was constant in level, produced a 30 second L_{eq} of 95 dBA, which is representative of a Cessna 172 at cruise. Using these data, the spectrum and noise level were checked before every screening and experimental session to ensure these were in line with the original calibration noise. If deviations were detected, the appropriate modifications were made to the equalizer settings.

Equalization of Speech Intelligibility Under Headset. While some previous studies have allowed subjects to set a “most comfortable” volume setting for hearing protection/communications devices, it was determined that having subjects manipulate the headsets’ volume controls would not be the best design for this investigation. Instead, the four headsets’ output volumes were equalized and the volume controls were taped over to prevent manipulation by the subjects. This procedure was chosen because the primary interest in the study was speech intelligibility, and it was thought that a potential sacrifice in realism by equalizing volume before the experiment would be more beneficial toward achieving a high level of experimental control. By equalizing the headsets at a very high level of speech intelligibility, differences between the headsets, and their effects on flight performance and workload, might be teased out as a controlled degradation of speech intelligibility was introduced into the ATC commands. This procedure was also more efficient to conduct. If subjects were allowed to set the volume, then the nearly four-hour flight simulation sessions would have dragged out even longer with the need to collect the at-ear sound amplitude and STI values for speech intelligibility determination.

In response to the need to equate the speech intelligibility across headsets, the following procedure was devised. The Bose headset was chosen as the reference headset to which each the other headsets would be equated. This choice was made based on the attenuation data provided by Bose Corporation which they collected using a microphone-in-real-ear (MIRE) procedure. (These data were not released by Bose for inclusion in this dissertation. If the reader is interested in obtaining these MIRE data, they should contact Bose directly.) The MIRE data show that the at-ear A-weighted attenuation of the Cessna 172 engine noise for the Bose headset was the middle of the range of the four headsets used in this study.

Using the Bose headset, a sample ATC command was played at a Speech Transmission Index (STI) value with no degradation (STI = 1.0) and was listened to by two subject matter experts (SMEs) in the presence of Cessna 172 engine noise at 95 dBA. The headset volume was set to what the SMEs agreed upon was a “normal” intensity level for listening to ATC communications. The at-ear dBA level and STI value of the ATC command was documented (82.7 dBA; STI = 0.80). This and all subsequent measurements were conducted by repeating the MIRE measure, underneath the Bose headset, three times and averaging the measurements. Then, the same speech was played at varying speech intelligibility values until the SMEs, while wearing the Bose headset, had determined the speech intelligibility values which would represent the worst speech intelligibility experienced in aviation radio communications, the midrange value, and the best available speech intelligibility based the pilots' experiences with in-flight radio communications. The STI values chosen were 0.30, 0.50, and 0.70, respectively. These STI values were also documented using a MIRE procedure, as above. The other three headsets were then equalized in under-earphone output level by measuring the at-ear dBA level and the STI value using the ATC command (STI = 1.0) in the engine noise. The earphone volume levels were manipulated until the STI value for each headset was equal with that of the Bose headset. The at-ear dBA level and STI value were documented again through a MIRE procedure. The Sennheiser and LightSPEED headsets were equalized at approximately the same earphone sound level and STI values as the Bose headset, 83.3 dBA and 81.9 dBA (STI = 0.80), respectively. However, the David Clark headset required a higher earphone sound level to reach the same speech intelligibility level (93.1 dBA, STI = 0.80). At these settings, the headsets were considered equalized. It should also be noted that each headset was calibrated using the MIRE method before each experimental session to ensure that the equalization had been preserved.

Following the equalization of the headsets, an SME, independent of the project, was brought in to review the decisions made by the previous two experts. This SME listened to the sound levels and speech intelligibility levels for each headset, and perceived these settings as equivalent. The third SME also agreed that the three speech intelligibility levels chosen did represent realistically poor, midrange, and best available aviation radio communications levels. This was the final validation of all decisions made for the equalization of the four headsets used in the investigation.

Screening Session. Pilots participating in the study were expected to attend 5 sessions. The first session was used for screening and familiarization with the flight characteristics of the iGATE simulator. During the screening session, each pilot was required to produce their FAA-issued pilot's license and medical certificate.

They were also required to produce their logbook so that it could be verified they have made the six instrument approaches, or received a new pilot certificate in the past six months, which was required for instrument currency (Federal Aviation Regulations, 2004). Following verification of currency, participants underwent the pure-tone audiogram, described earlier. This was done to ensure that a pilot's hearing level met the minimum hearing requirements as stated in the FARs. Pilots whose flight experience or hearing levels did not meet the FAR requirements would have been paid for their time and dismissed as an experiment participant. Fortunately, all pilots met the participation requirements. Next, all pilots were required to read and sign an informed consent form (Appendix A) documenting the purpose of the experiment, and the conditions of their participation. Following this, pilots engaged in a familiarization session with the flight simulator. They were given 14 minutes of "free flight" where they were allowed to perform any flight maneuvers they wished to gain familiarity with the flight characteristics of the simulator. This was done in a no turbulence, no weather, and a high visibility flight environment to reduce workload to the point which is associated with only stick and rudder maneuvering. Following the "free flight" segment, pilots were given six ATC commands with which they were required to readback and comply. Their responses were recorded for later use as the voice baseline in the analysis of voiced speech for workload measurement. At completion of the familiarization flight, the screening session ended.

Flight Scenario Sessions. Pilots were then scheduled for the four sessions of long duration experimental flight scenarios. The 3.5 hours each flight simulation lasted was considered long duration based on many factors the pilot encountered in the flight simulations. First, for the entire 3.5 hours across four simulations the pilots flew within instrument meteorological conditions (IMC) without any chance of visual reference to the ground. This type of flying is far more taxing and skill-intensive than normal visual flight, as evidenced by the fact that the FAA requires special training, a separate rating, and recency of experience to legally fly within the instrument environment. However, these flight simulations took this intensive environment one step further, by adding in multiple flight tasks (e.g. ILS approaches to minimums, holding patterns) which were accomplished within adverse weather conditions (e.g. severe turbulence, crosswinds). Lastly, it is important to note that research has shown an increase in fatigue of operator when exposed to high intensity noise environments (Kjellberg et al., 1996). It is believed that these factors created a fatiguing factor which was above normal and lead to the consideration that 3.5 hours was of long duration for a cross-country flight. At the beginning of each flight scenario, pilots were given an instruction sheet (Appendix B) with a few basic instructions and details regarding the performance requirements expected of the pilot during the flight. All performance requirements were based on the FAA's practical test standards for the certification of new instrument-rated pilots. Other instructions informed the pilots that they must follow all ATC commands, the pre-planned flight route, and must respond to all ATC commands and queries will a full and proper read-back, from which voiced speech analyses and speech intelligibility measures were obtained. For purposes of measuring workload through voice analysis, the simplified read-backs of “roger,” “wilco,” and “unable” were disallowed. Pilots were also given instructions (Appendix C) about the use of the Modified Cooper-Harper Scale (Appendix D).

A different headset was given to the pilots with each flight session. The order in which the pilots were exposed to the headsets was determined by a counterbalancing routine according to a full Latin-Square design. They were then given a flight log, which has already been prepared for that particular flight scenario, a Jeppesen instrument sectional map with the flight route already highlighted, and any necessary instrument approach plates. Pilots were given 20 minutes to study the cross-country flight in detail and ask any questions they may have concerning the pre-planned route, the Cessna 172 checklist, or the instructions.

The ordering of the flight modules within a cross-country flight simulation was primarily dependent upon the available flight paths that presently exist in the National Airspace System (NAS). For example, when the flight module incorporating a holding pattern came up in the flight module sequence, the aircraft must be in the vicinity of a VOR radio navigation aid. However, extensive counterbalancing, utilizing a full Latin-Square design, was still employed to insure that the order in which the pilots flew the scenarios, the scenarios themselves, or the order of the flight modules within the scenarios did not influence the analysis of the data collected. Full Latin-Square counterbalancing was employed to establish the order in which each subject flew each scenario, which headset was worn while flying each scenario, the order in which the STI values were presented within each scenario, and also the order in which each of the four workload flight modules per STI value were presented. Due to the previously mentioned physical and geographical constraints of the NAS, not all flight tasks fit the prescribed order determined by the counterbalancing. However, these deviations only occurred twice throughout the four scenarios. Careful planning enabled realistic flight plans to incorporate all twelve necessary flight modules in orders such that no order was repeated throughout the four flight scenarios. Therefore, it was determined not to be a concern for influence in the data (i.e., skewing performance data). Following each module, the scenario was briefly paused to query the pilot's opinion regarding the overall flight module workload using a Modified Cooper-Harper workload rating scale (Appendix D).

Upon completion of a flight scenario, the pilots were given a short rating scale to fill out regarding the comfort and communication system of the headset worn throughout the flight (Appendix E). Pilots were then paid for the flight session and were scheduled for the next flight session. At the end of the fourth and final experimental session, pilots were asked to rank each of the headsets in the order of their preference according to the criteria: overall communications performance, comfort, and noise reduction performance.

Protected Exposure Levels. The data for the protected exposure level (PEL) with communications input for each headset were gathered during the 'equalization of speech intelligibility under headset' portion of this study and the procedures are explained in detail within the section entitled Equalization of Speech Intelligibility Under Headset. The PEL with communications data, which represented the under-earphone, at-ear dBA levels required for each headset to achieve a speech intelligibility level of STI = 0.80, were collected prior to each experimental session.

The PELs without communications were collected following the completion of the entire experimental flight trial portion of the study. At this time, human participants were no longer needed and an acoustical test fixture was placed in the flight simulation room. The PELs without communications of the four headsets were measured in accordance with the ANSI S12.42-1995 standard. Each headset was placed on the acoustical test fixture in the 95 dBA aircraft engine noise. The PEL measurements incorporated two headsets of each model for all four models of headsets tested in this study. Therefore, each of the two headsets per model was tested twice for the protected exposure level, for a total of four protected exposure level data measurements per headset model. The noise levels under each headset were collected at a 1/3-octave band level resolution, and were analyzed at octave band intervals beginning with 63 Hz.

Headband Force Measurements. The first data collected in this investigation, before the screening and flight sessions had begun, were measurements of the headband force exerted by each headset. The headband force measurements were taken in accordance with the procedures laid out by ANSI S3.19–1974 using an INSPEC Earmuff Headband Force Measurement Rig. As with the protected exposure levels, each of the two headsets per model for all four headsets incorporated in the flight simulations were tested twice. Therefore, each model of headset had four data points for the headband force analysis.

Flight Performance Data Reduction. The iGATE flight simulator automatically collects data on 64 different variables at a rate of 32 times per second. The 3.5 hour flight simulations each yielded approximately 403,200 rows of data. The amount of sensitivity in these performance matrices (32 data points per second) was unnecessary for this experiment, and the analysis software was unable to handle the size of these matrices. Therefore, it was determined that one data point every 10 seconds would yield the sensitivity desired, to see effects while reducing the data enough to be manageable. It was further believed that if a more sensitive resolution was chosen (i.e., less than 10 seconds) that undesirable fluctuations, or noise, would appear in the data, unnecessarily raising the variability in the data analyses.

A MATLAB routine was written which read the performance matrix, then averaged every 320 rows of data together and output the result to an Excel file. Since this experiment was only interested in six flight performance measures, the other 58 variables the simulator had collected were deleted. At this point, each Excel file showed the individual pilot's performance on a given flight scenario. However, the pilot's deviation from the ATC commanded performance needed to be calculated so that the data could be compared across the four different flight scenarios. If this was not done, it would impossible to compare, for example, an altitude of 4500 feet with an altitude of 6200 feet. However, it would be possible to compare, for example, a flight performance deviation of 500 feet and a deviation of 1200 feet. Therefore, the ATC commanded performance (i.e., assigned heading, altitude, airspeed, etc.) was subtracted from the pilot's performance and the absolute value was taken. Once this procedure was complete, the flight performance data was ready for statistical analysis.

Data Analysis Overview

Primary task performance data and *voice analysis data* were each grouped and analyzed using a 4 x 3 x 3 multivariate analysis of variance (MANOVA). If the MANOVA resulted in significance differences, the individual measures were separated and analyzed using a 4 x 3 x 3 univariate analysis of variance (ANOVA). Post hoc comparisons were conducted on significant ANOVA results using a Tukey HSD test.

The *speech intelligibility data* was tested using a 4 x 3 x 3 ANOVA procedure. Where significant differences were found, further post-hoc analysis was carried out using the Tukey HSD procedure.

Workload data collected through the Modified Cooper-Harper scale were analyzed using a nonparametric Friedman's *F*-Test because this particular test can analyze three or more categorical variables in a data set which incorporates an ordinal data type. If the *F*-test resulted in significance, further pairwise contrasts were carried out using a Bonferroni correction to protect against type one error inflation.

All *bipolar scale ratings* (comfort, communications performance, and miscellaneous ratings) were each converted into numerical scores ranging from 1 (on the far left) to 7 (on the far right) and analyzed separately using the ANOVA procedure as was carried out by Park & Casali (1991). Further post-hoc analysis was carried out using the Tukey HSD procedure.

The final *rankings* were analyzed using the Fisher's Exact Test. The major reason this test was chosen over a Chi-Square was that the Chi-Square is only appropriate if all cells of the experimental design contain a value of 5 or greater (Ott & Longnecker, 2001). When this assumption does not hold, as in this case, the Fisher's Exact Test is more appropriate. Therefore, the Fisher's Exact test was used to analyze for categorical differences of the pilots' assigned rankings of each level of the headset variable. Where significant differences were found, further pairwise contrasts were performed using additional Fisher's Exact Test procedures.

The headset *protected exposure levels* were collected at 1/3-octave band levels from 63 Hz to 12.5 kHz. The data collected were separated and analyzed by octave band frequency. A separate one-way ANOVA was used to analyze each frequency with headset as the sole variable. Where significant differences across headsets were found, the data were further analyzed using the Tukey HSD procedure.

The *headset headband forces* were collected and analyzed using a one-way ANOVA. Where significant differences across headsets were found, the data were further analyzed using the Tukey HSD procedure.

RETESTING THE LIGHTSPEED TREATMENT

Following the completion of data collection, it was discovered that the black tape placed over the LightSPEED headset logos had actually blocked an air passage which was hidden, very subtlely, in a decorative groove surrounding the logo. Blocking such a passage was thought to have changed and probably degraded the performance of the headset's ANR attenuation. Therefore, it was determined that retesting was the most appropriate course of action. However, before embarking on retesting, an analysis of the four flight simulation scenarios was conducted to determine whether order or learning effects existed in the data, such that the pilots' performance in the dependent measures improved between their first flight simulation session and subsequent flight sessions. If order effects did exist, the entire experiment (all four headsets) was to be run again with new participants. On the other hand, if order effects did not exist, it was deemed acceptable (based on statistical determination) to rerun only the LightSPEED treatment with the original eight participants.

The dependent measures associated with the three major independent variables, speech intelligibility, workload, and flight performance, were analyzed using the same statistical procedures outlined in the previous section. However, the comparison of interest was the pilots' first flight session compared with the second, third and fourth flight sessions.

Speech intelligibility was analyzed using a one-way ANOVA. Flight session order was found be non-significant, $F(3, 379) = 1.88$, $p = 0.133$ (Table 4). Next, the Modified Cooper–Harper scale ratings were tested to see if there were any differences in perceived workload based on the order of the flight sessions. The data were analyzed using a Friedman's *F*–test. It was also found that flight session order was non-significant regarding the workload measure, *F*(3, 380) = 0.98, *p* = 0.400 (Table 5). Lastly, the six flight performance measures were tested for order effects. As stated in the Data Analysis Overview section, these measures were first analyzed using a repeated measures MANOVA. If this test resulted in significance, then the ANOVA procedure would be applied to the individual flight measures. The MANOVA also showed that flight session order was not significant, *F*(3, 1991) = 2.15, *p* = 0.092 (Table 6).

Based on the above analyses, it was determined that retesting *only* the LightSPEED treatment was acceptable. The previously collected data related to the LightSPEED treatment (i.e., with the port taped over) was deleted from the data set. Once retesting was complete, the new data collected for the LightSPEED headset was added to the data set previously collected. This data set was then analyzed in accordance with the procedures set forth in the Data Analysis Overview section.

Table 4

ANOVA summary table for speech intelligibility order effects.

Source	df	MS	*F*	*p*
Session Order (SO)	3	0.215	1.88	0.133
Subject (S/SO)	379	0.114		

Table 5

ANOVA summary table for Modified Cooper-Harper workload rating order effects.

Source	df	MS	*F*	*p*
Session Order (SO)	3	4.156	0.98	0.400
Subject (S/SO)	380	4.222		

Table 6

Wilk's Lambda MANOVA summary table for flight performance order effects.

Source	Wilk's Lambda Value	Num df	Den df	*F*	*p*
Session Order (SO)	0.997	3	1991	2.15	0.092

RESULTS AND DISCUSSION

Individually, the measures concerning workload, speech intelligibility, headset comfort, communications performance, headset protected exposure levels, and the final rankings were simpler analyses, each affected by fewer variables, comparatively, then the performance results. Consequently, comprehension of the results and subsequent interpretations of these less complex analyses should be a simpler task. On the other hand, the primary task performance measures were potentially influenced by all variables involved. Therefore, it was determined that the experimental results of the less complex analyses should be presented before the results of the primary task measures were explored. In this way, an understanding of the results of each of the less complicated measures will hopefully facilitate an understanding of the more complex primary task performance results and interpretations, wherein most, if not all, variables contribute some influence.

Speech Intelligibility

ATC commands and Pilot Readbacks. The speech intelligibility metric represents the number of times a pilot required air traffic control commands to be transmitted to respond with a perfect command readback. This procedure is akin to speech intelligibility tests that incorporate a percent-words-correct paradigm. However, in the method incorporated here the number of words correctly repeated was not recorded, but instead, the number of times a pilot must hear a stimulus (ATC command) to repeat the stimulus with 100% accuracy was captured. Therefore, it was theorized that as speech intelligibility decreases, whether due to STI value or headset characteristics, the number of times a pilot required the ATC command to be spoken would increase. Also, the research hypotheses predicted that as workload increased, more mental resources would be engaged leaving less resources to accurately comprehend radio communications or to stay within the FAA Practical Test Standards for flight performance. Therefore, either task performance would decrease or comprehension of the ATC commands would decrease, depending upon the pilot's priority schema.

Analysis to confirm the data did not violate any assumptions was necessary. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted utilizing a scatterplot of the predicted versus residual values from the speech intelligibility data and showed the tell-tale horizontal band of values. This indicated the data did not violate the homogeneity of variance assumption. Lastly, Mauchly's test of sphericity was conducted which confirmed that the data did not violate the sphericity assumption, $p = 0.37$. The collected data was determined to fit all assumptions of a parametric ANOVA. Therefore, a 4 x 4 x 3 repeated measures ANOVA was used in the statistical analysis. The results given in Table 7 show a significant main effect for headset, $F(3,21) = 11.62$, $p < 0.0001$, STI value, $F(2,14) = 29.39$, $p < 0.0001$, and workload, $F(3,21) = 4.20$, $p = 0.006$. Furthermore, a significant interaction was found between headset and STI value, $F(6,42) = 6.91$, $p < 0.0001$. Post-hoc comparisons of all pairwise combinations were carried out using the Tukey HSD procedure.

Table 7

ANOVA summary table for the number of ATC commands given for a correct pilot readback (speech intelligibility measure).

Source	df	MS	*F*	*p*
<u>Between</u>				
Subject (S)	7	0.204		
<u>Within</u>				
Headset (H)	3	2.452	11.62	<0.0001
H x S	21	0.211		
STI	2	5.555	29.39	<0.0001
STI x S	14	0.189		
Workload (W)	3	0.203	4.20	0.006
W x S	21	0.048		
H x STI	6	1.368	6.91	<0.0001
H x STI x S	42	0.198		
H x W	9	0.033	0.46	0.902
H x W x S	63	0.071		
STI x W	6	0.060	1.00	0.427
STI x W x S	42	0.060		
H x STI x W	18	0.021	0.49	0.961
H x STI x W x S	125	0.042		

These post-hoc comparisons showed that for the main effect of headset, pilots wearing the Bose headset required a significantly fewer number of times an ATC command must be transmitted than with the LightSPEED. The Sennheiser headset did not differ significantly from the Bose headset or the LightSPEED headset. The results also showed the number of times an ATC command must be transmitted for correct readback was significantly higher when a pilot wore the passive David Clark headset than when the pilot wore any one of the three ANR headsets (Figure 22). These results support the research hypothesis that speech intelligibility did increase while wearing an ANR headset, and is reflected in the fact that pilots required fewer ATC command repeats for a correct readback while wearing any one of the three ANR headsets. When the pilot wore the passive headset, speech intelligibility decreased causing the pilot to require an ATC command to be repeated a significantly higher number of times.

This relationship has significant implications for safety in the aviation system. It has been reported that 78% of emergency medical service aviation accidents are directly caused by communications difficulties (Connell & Reynard, 1993). Additionally, the NTSB Aviation Accidents Reports have attributed some of the worst aviation accidents in history to degraded radio transmissions leading to pilot – ATC miscommunications, such as the Tenerife catastrophe described in the beginning of this dissertation (Aviation-Safety.net, 1996). These results suggest that the use of ANR technology in aviation communications headsets will significantly increase the speech intelligibility of ATC commands.

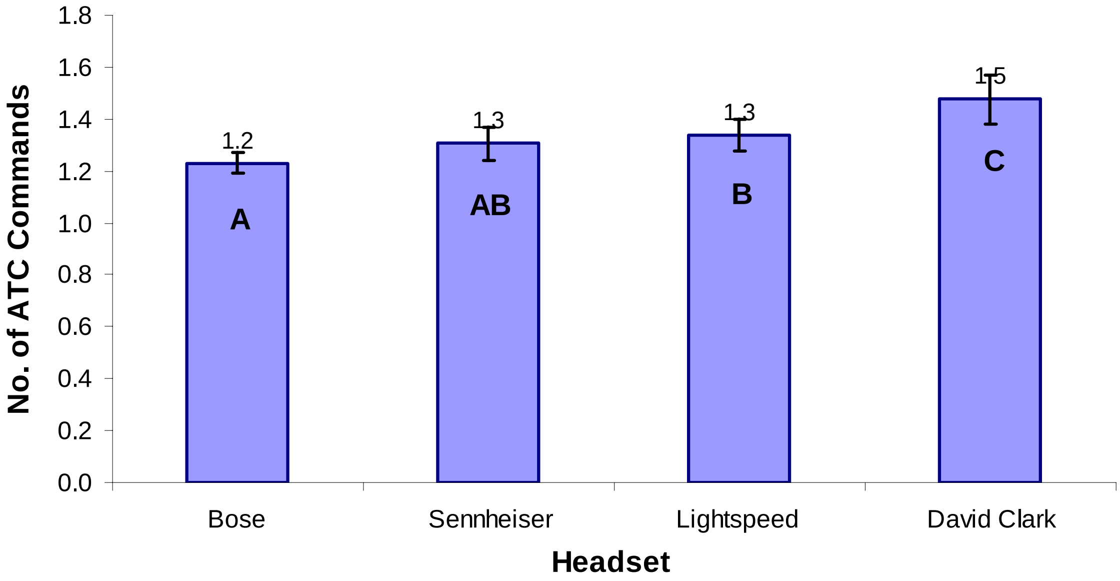


Figure 22. The effect of aviation headset on the mean number of times an air traffic control command had to be transmitted to the pilot for a 100% correct readback. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Now of course, ANR technology will not solve all miscommunications problems, but the increase in speech intelligibility afforded by ANR headsets over their passive counterparts could allow the pilot to understand enough of a degraded transmission to mean the difference between safe comprehension of instructions and the execution of an erroneous, dangerous flight maneuver.

Another benefit to safety in the aviation system is the potential for ANR headsets to cut down on the amount of time an ATC controller must spend relaying or repeating instructions to a pilot, thereby decreasing the controller's workload, which would be especially important to ATC directing the busiest controlled airspaces (Class B airspace).

Further analysis of the main effect of Speech Transmission Index (STI) speech intelligibility values resulted in a significantly higher number of times an ATC command must be spoken for a correct pilot readback at the lowest STI value (0.30), than the 0.50 or the 0.70 STI value (Figure 23). The analysis also found that there was no significant difference in speech intelligibility between the 0.50 and 0.70 STI values. It seems that the difference in speech intelligibility between 0.50 and 0.70 is not enough of an increase to cause pilots to misunderstand radio commands, or miss a command altogether. This finding coincides with the results of past studies where task performance did not begin to suffer until speech intelligibility dropped below 50% (Whitaker, Peters, & Garinther, 1989; Whitaker, 1991; Whitaker & Peters, 1993). This 50% speech intelligibility threshold has been equated by Steneeken (2004) to an STI value of 0.50, exactly the threshold that this study establishes before the pilots' communication task performance begins to deteriorate.

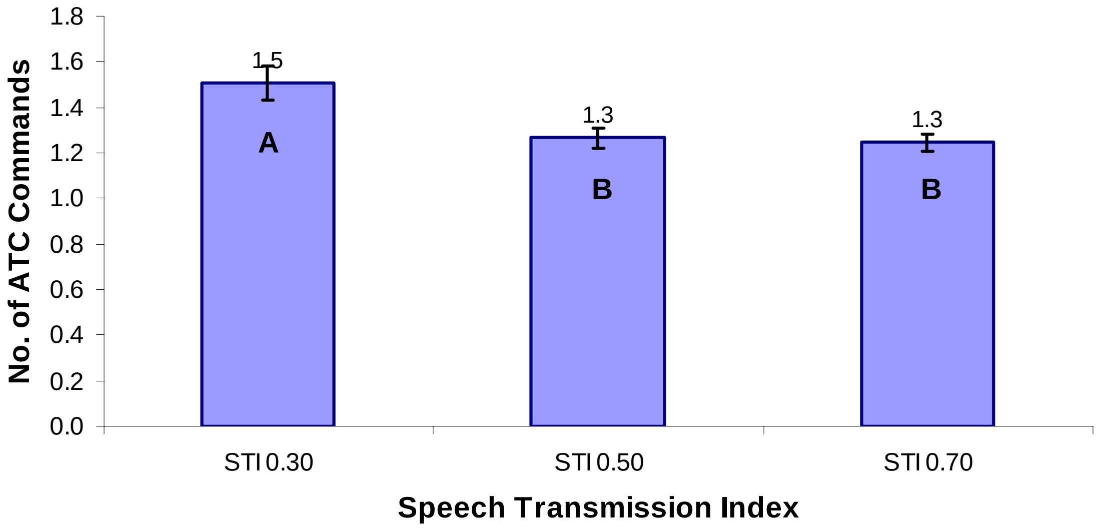


Figure 23. Effect of various STI values on the mean number of times an air traffic control command must be transmitted to the pilot for a 100% correct readback. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

So, the communications task performance does not decrease sharply between the higher values, such as 0.50 and 0.70, but once speech intelligibility begins to drop below 0.50 pilots require significantly more repeated instructions to correctly readback the ATC command.

Analysis of the final main effect of *workload* showed that the only significant difference between workload conditions for speech intelligibility occurred between the high psychomotor and high perceptual workload condition, where the number of ATC commands was significantly higher for high psychomotor workload than high perceptual workload. The low psychomotor workload and low perceptual workload proved not to be significantly different from both the high psychomotor workload condition and the high perceptual workload condition (Figure 24). It was expected that post-hoc comparisons would show that the number of times an ATC command needed to be transmitted would increase as workload moved from the low level to the high level. However, it is obvious this is not the case.

A potential explanation for the variance seen between these results and those predicted by the research hypotheses could lie in the realism of the flight tasks designed into the flight simulation. Every variable available in the flight environment (e.g. weather, turbulence, and flight maneuvers commanded by ATC) was used to create flight modules which were realistic, yet designed to incorporate enough control to retain a uniform type and level of workload throughout the course of the flight module. Unfortunately, according to these results, this may not have been accomplished. As will be shown in the analysis of the Modified Cooper Harper workload ratings, the overall workload was consistent with the intentions of the experimental design.

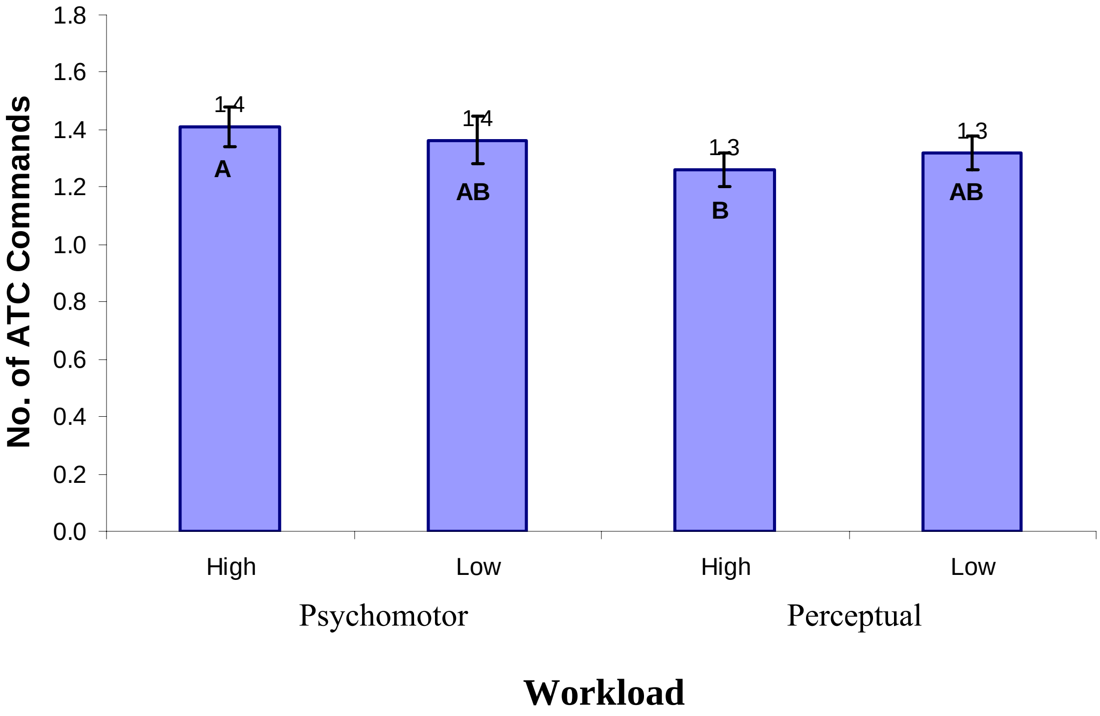


Figure 24. Effect of workload condition on the mean number of times an air traffic control command must be transmitted to the pilot for a 100% correct readback. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

However, the speech intelligibility measure was not an overall measure. Flights tasks by nature have periods of higher workload and lower workload throughout their execution. For example, entering a holding pattern is of higher workload, than circling within the holding pattern. Therefore, ATC commands to change heading, altitude, and airspeed were employed to even out the temporary decreases in workload. However, it is likely that these attempts were unsuccessful. Consequently, the speech intelligibility data which were collected six times per flight module (approximately every two minutes) would reflect the increases and decreases in workload within each flight module. These effects may have then been averaged out through the statistical analysis to the point that the results show little to no difference between the workload conditions and levels as is displayed by Figure 24.

The analysis also resulted in a significant headset by STI value interaction for the speech intelligibility measure (Figure 25). The interaction effect was further analyzed using the method present in Keppel (1971). Levels of one variable were analyzed using an *F*-test while holding the other variable constant. This was done until all levels of both variables had been analyzed. Where significant *F*-tests resulted, further simple contrasts where made to isolate the source of the statistical significance. Figure 27 shows the end results of the interaction effect analysis.

This interaction supports two interpretations addressed in this subsection of the Results section. First, the significant difference in speech intelligibility performance between the STI = 0.30 value and the upper two levels of STI= 0.50 and STI = 0.70 for the David Clark headset supports the conclusion drawn by examining the main effect of STI value for this measure.

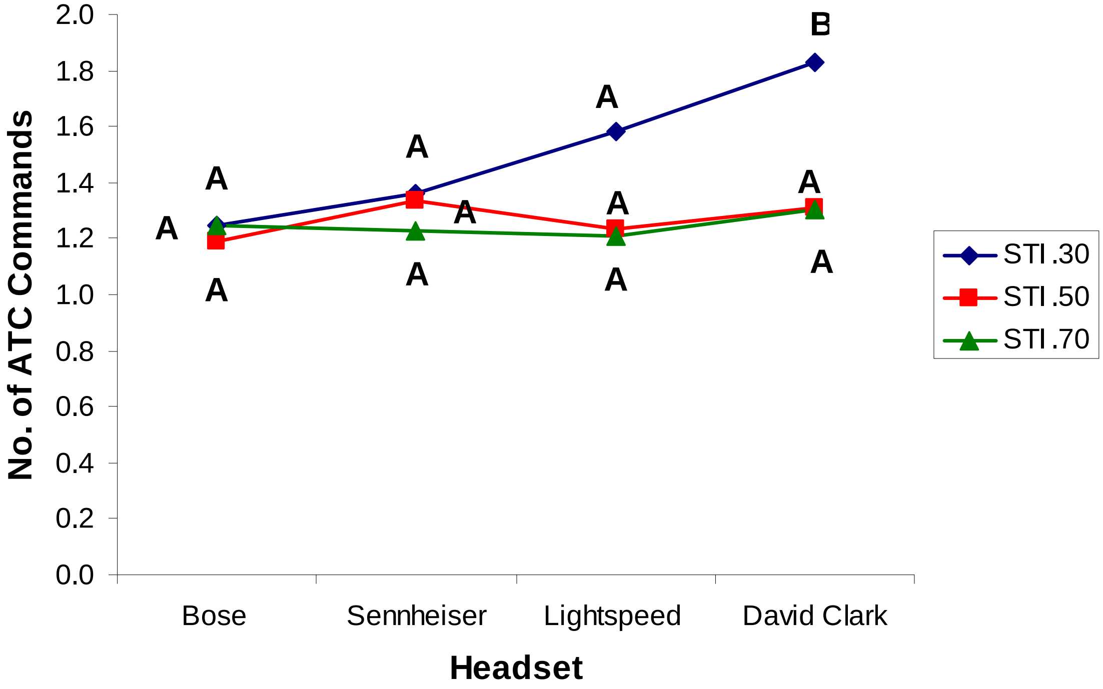


Figure 25. Interaction effect of the speech intelligibility measure (headset * STI value). Different letters represent significant differences at the $p < 0.05$ level.

As previously stated, there is a performance threshold at approximately STI =0.50, below which speech intelligibility performance deteriorates rapidly. However, the interaction clearly shows that this performance threshold does not exist for the ANR-based headsets (Bose, Sennheiser, LightSPEED), supporting one of the original hypotheses of this investigation. ANR has increased the speech intelligibility (which has also been supported by previously stated main effects results in this section) of the ATC transmission overcoming the theorized speech intelligibility performance threshold, as evidenced by the non-significant differences found between the three STI values for each of the three ANR headsets.

Additionally, there is a significant practical implication found in this interaction effect. The passive David Clark headset shows a significant decrease in speech intelligibility performance during the worst speech intelligibility level. To put this into practical terms, the passive David Clark headset allows a pilot's speech intelligibility performance to deteriorate when it is needed most, during situations when radio communications are questionable and time-critical ATC commands could more easily be misinterpreted or completely missed. On the other hand, the ANR-based headsets have increased the speech intelligibility, raising the pilot's performance at the lowest speech intelligibility level. In practice, the use of ANR headsets could potentially mean the difference between comprehending a critical ATC transmission and executing an erroneous, potentially dangerous flight maneuver.

Workload

Modified Cooper-Harper Ratings. The Modified Cooper-Harper rating scale is divided into 10 discrete categories, which are further grouped into the following four larger categories.

- 1 – 3 = an acceptable level of workload
- 4 – 6 = high workload, workload should be reduced
- 7 – 9 = Major design deficiencies, system redesign is strongly recommended
- 10 = Major design deficiencies, system redesign is mandatory

A Friedman's *F*-test was chosen to analyze the workload ratings given by the pilots after each flight module through the Modified Cooper-Harper rating scale. The Friedman's *F*-test was deemed most appropriate because it was designed to compare more than two categorical variables with an ordinal data type. The Friedman's *F*–test is the nonparametric equivalent to a repeated-measures ANOVA. The results are given in the Friedman's *F*-test summary table of Table 8. These results show significant main effects for headset $F(3, 21) = 12.25$, $p < 0.0001$, STI value $F(2, 14) = 9.09$, $p < 0.0001$, and workload $F(3, 21) = 83.66$, $p < 0.0001$. The analysis further indicated that all interactions resulted to be non-significant at the $p < 0.05$ level. Pairwise contrasts were carried out utilizing a Bonferroni correction to control for Type I error inflation. These contrasts showed that for the main effect of headset, there was no significant difference in workload between the three ANR headsets (Bose, Sennheiser, and LightSPEED). However, workload was rated as significantly higher when a pilot wore the passive David Clark headset than when the pilot wore any one of the three ANR headsets (Figure 26).

Table 8

Friedman's F-Test summary table for Modified Cooper-Harper workload ratings.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	0.543		
Within				
Headset (H)	3	2413.250	12.25	<0.0001
H x S	21	197.319		
STI	2	1959.404	9.09	<0.0001
STI x S	14	215.556		
Workload (W)	3	13908.391	83.66	<0.0001
W x S	21	166.249		
H x STI	6	122.001	0.87	0.855
H x STI x S	42	140.231		
H x W	9	64.421	0.82	0.547
H x W x S	63	78.563		
STI x W	6	105.768	1.78	0.098
STI x W x S	42	59.420		
H x STI x W	18	42.622	0.55	0.956
H x STI x W x S	125	77.495		

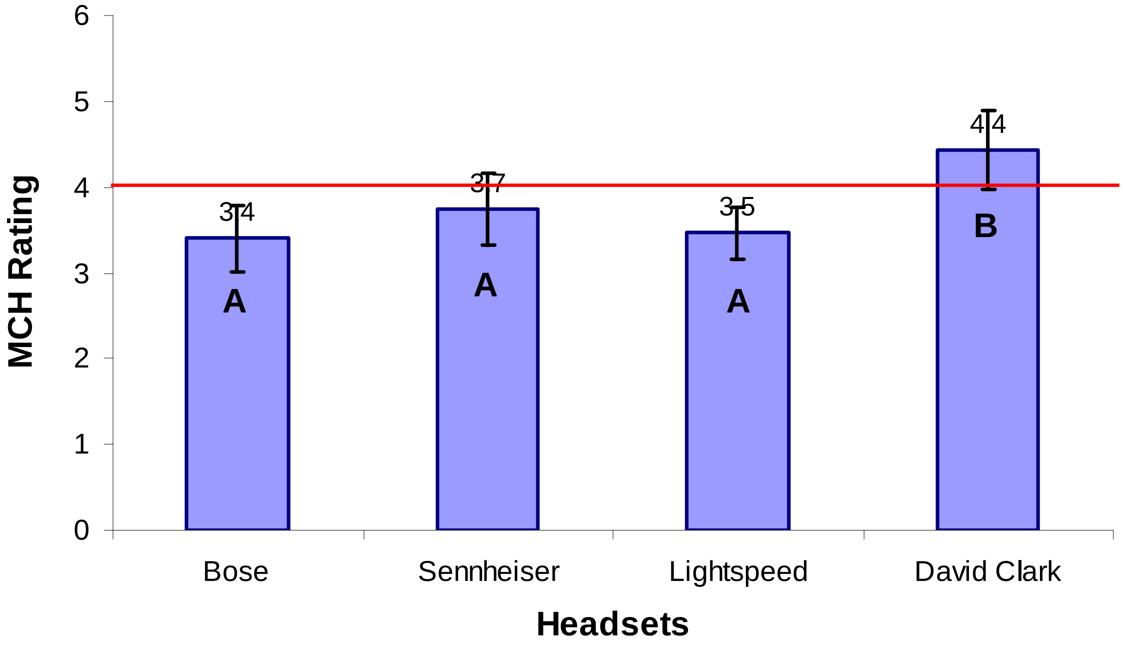


Figure 26. Pilot workload ratings per headset as collected using the Modified Cooper-Harper (MCH) Rating Scale (1 = an acceptable level of workload, 10 = major design deficiencies, system redesign is mandatory). Different letters signify a significant difference. Horizontal line at four indicates boundary between acceptable workload and high workload. Vertical range bars represent 95% confidence intervals about the means.

This result further supported the research hypotheses demonstrating that the increased speech intelligibility afforded by the ANR headsets, as shown by the results of the speech intelligibility measure, decreases the amount of mental resources necessary to effectively respond to ATC commands while flying the aircraft. This decrease in required mental resources manifests itself in the pilot's perception of decreased workload, as the main effects of headset show.

Practically speaking, the difference shown by the workload analysis between the ANR headsets and the passive headset is important. The ordinal scale of the Modified Cooper-Harper rating is broken up into different categories with one through three given as ratings of acceptable workload, whereas anything with a four or higher indicates the workload is too high and a redesign is suggested. As Figure 26 shows, the rated workload levels of pilots wearing the ANR headsets are all within the range of an acceptable level of workload. However, the rated workload levels of pilots while wearing the passive headset undergoing the same flight conditions had increased enough to push the workload levels into the high workload range, suggesting a redesign of the system. Further implications of these results are similar to those of the headset main effect for the speech intelligibility measure. The ANR headsets reduce the mental workload of the pilot, freeing mental resources which could be reallocated to other tasks, such as navigation or flight maneuvers. This greater amount of available resources increases the safety buffer between adequate performance and errors due to task overload. Therefore, wearing an ANR headset potentially creates a safer pilot operating in the aviation system.

Further analysis of the main effect of STI value resulted in a significantly higher workload rating for the lowest STI value (0.30) than the middle (0.50), or high (0.70) STI values. The analysis also found that there was no significant difference in workload ratings between the 0.50 and 0.70 STI values (Figure 27). As with the main effect of headset, these results concur with those of the STI value main effect of the speech intelligibility measure showing that workload does not significantly increase as STI decreases from 0.70 to 0.50. However, as STI decreases from 0.50 to 0.30, it becomes more difficult to understand the ATC commands, which necessitates that additional mental resources be allocated to attend to, and correctly comprehend, the specific commands. The result of low speech intelligibility is a significant increase in mental workload, as the analysis shows.

Finally, analysis of the main effect for workload showed that the highest workload rating was placed on the high psychomotor workload condition. The high perceptual workload rating was significantly lower than the rating of the high psychomotor workload condition, but a significantly higher rating than those of the low psychomotor and low perceptual workload ratings. Lastly, there was no significant difference between the low psychomotor and low perceptual workload ratings (Figure 28). Although the results of the speech intelligibility measure for the main effect of workload show that workload was probably not uniform throughout the flight module, the MCH ratings of the different workload types and levels are a validation of the experimental design in terms of overall workload perceived for each flight module.

The high perceptual workload condition was originally designed to be at an equal workload level as the high psychomotor workload.

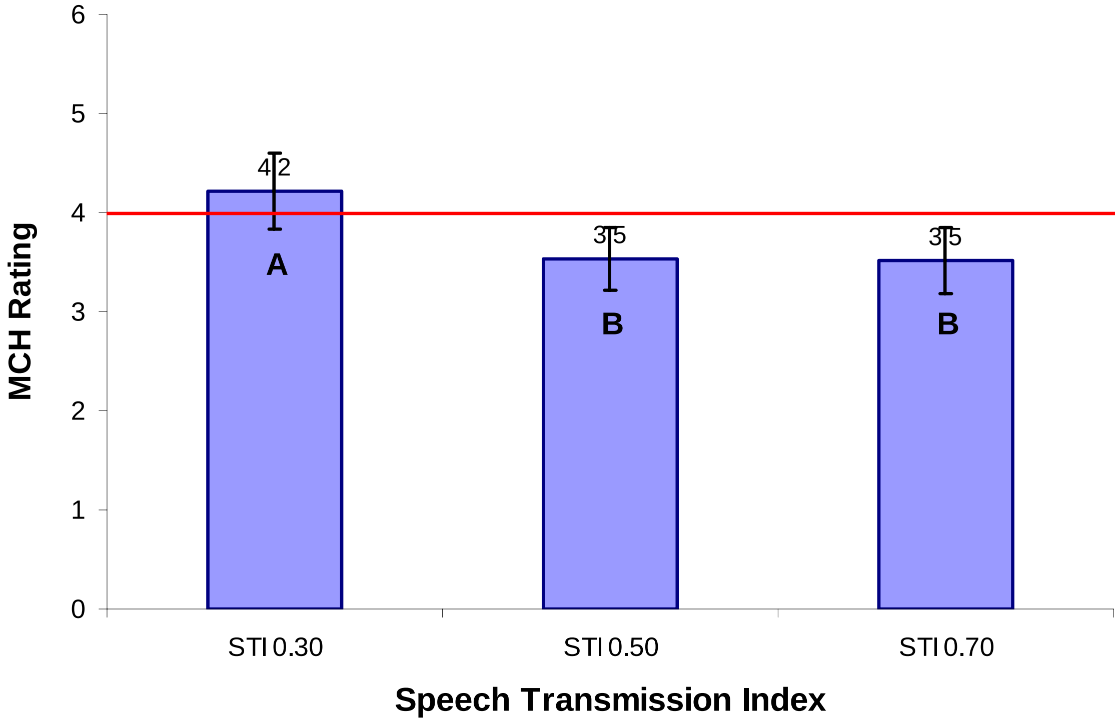


Figure 27. Pilot workload ratings per STI value as collected by the Modified Cooper-Harper (MCH) rating scale (1 = an acceptable level of workload, 10 = major design deficiencies, system redesign is mandatory).Different letters signify a significant difference at the *p* < 0.05 level. Horizontal line at four indicates boundary between acceptable workload and high workload. Vertical range bars represent 95% confidence intervals about the means.

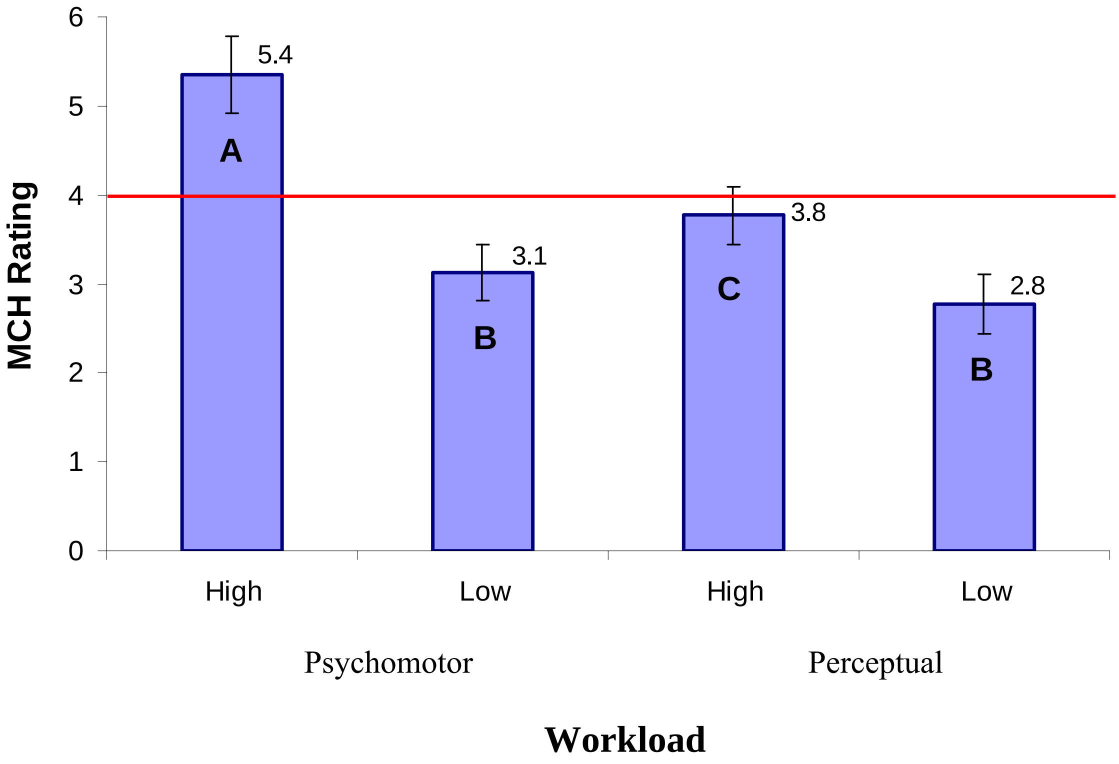


Figure 28. Pilot workload ratings per speech intelligibility value as collected by the Modified Cooper-Harper (MCH) Rating Scale (1 = an acceptable level of workload, 10 = major design deficiencies, system redesign is mandatory). Different letters signify a significant difference at the $p < 0.05$ level. Horizontal line at four indicates boundary between acceptable workload and high workload. Vertical range bars represent 95% confidence intervals about the means.

However, due to limitations in the simulator's software, the only instrument failures that could be used during the simulation were very obvious failures such an attitude indicator failure. Therefore, the workload for the perceptual high workload condition was significantly lower than the high psychomotor workload condition. The MCH ratings did show that both high psychomotor and high perceptual workload were significantly greater than their low workload counterparts. Therefore, it could be considered that the overall effects of the experimental and simulation designs were successful in their intended functions.

Voice Analysis. As stated in the methods section, data were collected for two measures of workload through the recording of the pilot's ATC readbacks. The average amplitude and formant frequency of each readback was derived from the recorded readbacks, and the pilot's own baseline amplitude and frequency were subtracted to yield amplitude and frequency deviations. These two measures were then analyzed using a 4 x 4 x 3 repeated measures MANOVA. The MANOVA indicated significance for only one of the main effects, headset, F (6,181) = 15.87, $p < 0.0001$. As Table 9 shows, the main effects for STI values and workload, along with all interactions, were found to be non-significant. Further analyses were conducted on the headset main effect by separating the two speech analysis measures and conducting individual 4x4x3 repeated measure ANOVAs.

Voice Amplitude ANOVA. Analysis to confirm the data do not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation.

Table 9

Wilk's Lambda MANOVA summary table for speech analysis.

Source	Wilk's Lambda Value	Num df	Den df	*F*	*p*
Headset (H)	0.938	6	181	15.87	<0.0001
STI	1.000	4	163	0.05	0.955
Workload (W)	0.997	6	181	0.70	0.554
H x STI	0.995	18	210	0.63	0.705
H x W	0.993	9	225	0.58	0.813
STI x W	0.995	12	210	0.60	0.729
H x STI x W	0.985	36	243	0.59	0.907

Another visual analysis was conducted on a scatterplot of the predicted versus residual values from the amplitude data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption. Lastly, Mauchly's test of sphericity was conducted which confirmed that the data did not violate the sphericity assumption, $p = 0.053$.

A 4 x 4 x 3 repeated measures ANOVA was conducted on the pilot voice amplitude data. This result showed a significant main effect for headset, $F(3, 21) = 4.79$, $p = 0.003$ (Table 10). Main effects for STI value and workload could not be considered, nor could any interactions be considered as they did not prove to be significant in the previously presented voice analysis MANOVA procedure. Further analysis of the main effect of headset was carried out using a Tukey HSD method to make all pairwise comparisons. Post hoc comparisons of the main effect of headset showed that speech amplitude deviations from baseline were significantly lower when the Bose headset was worn, than the David Clark headset. There were not any significant differences found between the Bose, Sennheiser, and LightSPEED headsets. Significant differences were found between the LightSPEED headset and both the Sennheiser and the David Clark headset, such that the speech amplitude deviations were significantly lower when the LightSPEED headset was worn than when the other two headsets were worn. Pairwise comparisons did not find any significant differences between the Sennheiser headset and the David Clark headset (Figure 29).

The voice analysis method employed in this investigation was intended to be a measure of workload that could be used simply by recording the pilot's speech.

Table 10

ANOVA summary table for voice amplitude deviations from baseline.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	2.275		
Within				
Headset (H)	3	2.098	4.79	0.003
H x S	21	0.438		
STI	2	0.169	0.68	0.507
STI x S	14	0.249		
Workload (W)	3	0.011	0.23	0.874
W x S	21	0.047		
H x STI	6	0.027	0.29	0.941
H x STI x S	42	0.093		
H x W	9	0.036	0.64	0.763
H x W x S	63	0.056		
STI x W	6	0.004	0.10	0.996
STI x W x S	42	0.038		
H x STI x W	18	0.092	0.14	1.000
H x STI x W x S	125	0.657		

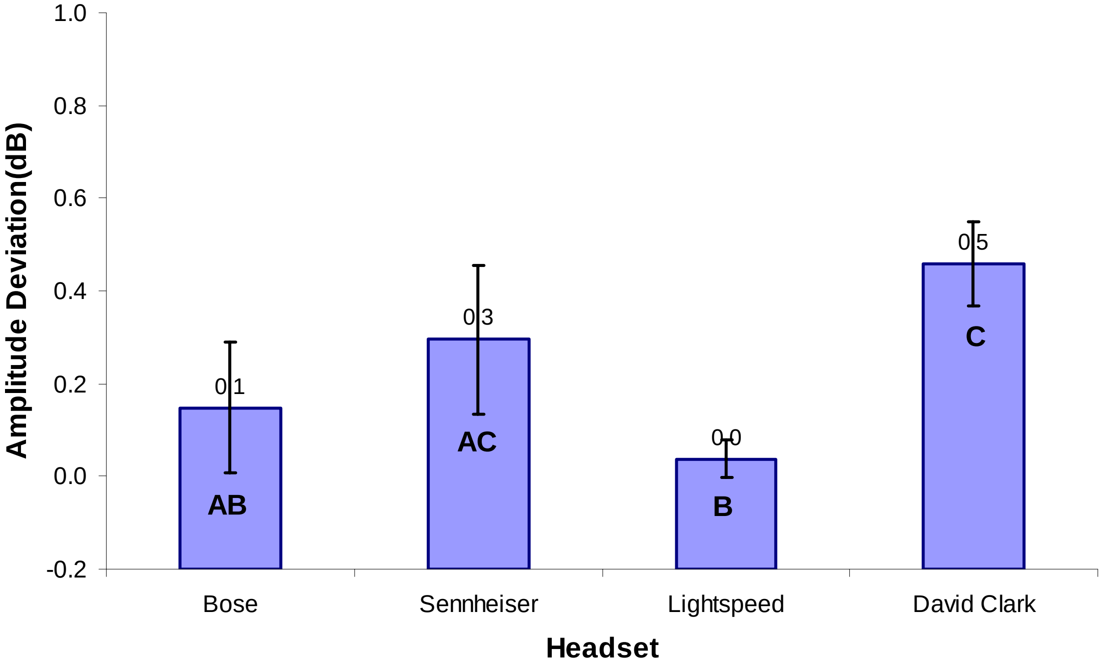


Figure 29. Difference between the recorded pilot voice amplitude during flight from the recorded non-workload baseline, for each headset. Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

As such, it was tested in the flight simulation environment without any controls or manipulations in the environment implemented with hopes that success in this study would demonstrate the viability of this method to be readily implemented in aviation systems without change to those systems.

Unfortunately, the results, namely the non-significant relationship between voice amplitude and workload, do not show this measure to be deployable. However, with further research, voiced speech analysis may still be an effective measure of workload to specific aspects of the air transportation system. It seems that one of the main confounding factors affecting the speech analysis is that of the piston engine noise. The great majority of general aviation aircraft in operation today employ reciprocating engines, which was one of the main focuses of this ANR investigation. However, the intense engine noise was picked up by the pilot's microphone and recorded along with their speech. It is highly likely that the noise was of sufficient influence to mask the subtle differences in speech which reportedly indicate differences in workload.

One area of air transportation which may be suitable for the use of this measure is the commercial airlines. The difference in engine noise between a Cessna 172 and a Boeing or Airbus turbojet is immense. Inside a Cessna 172, the pilot is situated immediately behind a piston engine inside a fuselage with little sound isolation. On the other hand, a pilot of a modern airliner is situated in the nose, a good distance from the engines, which are under the wings. Also, the fuselage is sound isolated to reduce noise as much as possible for the comfort of the passengers. As a result, the voiced speech analysis measure may be unsuitable for general aviation, but may be applicable to commercial air carriers. Of course, this requires further investigation.

The results did show that a significant difference existed between the four headsets. It cannot be said that this indicates which headset helped alleviate workload and which headset contributed to workload, as the workload variable was not significant. Therefore an explanation for this relationship must be sought elsewhere. The only source of uncontrolled variability that might affect the speech amplitude between headsets is each headset's microphone. Recording of the baseline speech amplitude was done while the pilot was wearing an Aviation Communication (AvCom) headset so as not to predispose pilots to any one of the four headsets to be tested. It is possible that the deviation in the speech measures actually reflect how similar or dissimilar (to the AvCom microphone) that the respective headset microphones transduce the speech into electrical signals, rather than speech deviations due to workload. It would be necessary to rule out this possibility and the potential engine noise confound before the viability of the voice analysis measure is decided.

Voice Formant Frequency ANOVA. Analysis to confirm the data do not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted on a scatterplot of the predicted versus residual values from the formant frequency data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption. Lastly, Mauchly's test of sphericity was conducted which confirmed that the data did not violate the sphericity assumption, $p = 0.06$.

A 4 x 4 x 3 repeated measures ANOVA was conducted on the pilot speech frequency data. This analysis also showed the significant main effect for headset, *F* (3,21)= 16.42, $p < 0.0001$. Main effects for STI value and workload were not significant at the $p < 0.05$ level, nor were any interactions in this analysis (Table 11). Further analysis of the main effect for the variable headset was carried out using a Tukey HSD method to make all pairwise comparisons. Post hoc comparisons of the four headsets showed that speech frequency deviations from each pilot's baseline were significantly lowest when the Bose headset was worn as compared to the other three headsets. In addition, the frequency deviations for the David Clark headset were significantly lower than the Sennheiser headset, but not significantly different from the LightSPEED headset. Although the Sennheiser speech deviations were significantly higher than both the Bose and David Clark headsets, they were not found to be different from the LightSPEED headset (Figure 30).

The results of the voice formant frequency measure show the same pattern as that of the speech amplitude measure. Again, it is believed that a combination of the light aircraft engine noise and the microphone transduction characteristics of each headset was the major influence driving these results. Light aircraft engine noise and microphone characteristics could easily affect the frequency and amplitude of speech. Consequently, these confounding variables must be ruled out in subsequent experiments before a decision regarding the analysis of speech as a measure of workload can be made.

Table 11

ANOVA summary table for voiced speech formant frequency deviations from baseline.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	0.008		
Within				
Headset (H)	3	0.082	16.42	<0.0001
H x S	21	0.005		
STI	2	0.001	0.04	0.960
STI x S	14	0.002		
Workload (W)	3	0.0007	0.70	0.551
W x S	21	0.001		
H x STI	6	0.0001	0.66	0.685
H x STI x S	42	0.0002		
H x W	9	0.0006	0.61	0.790
H x W x S	63	0.001		
STI x W	6	0.0003	0.62	0.714
STI x W x S	42	0.0005		
H x STI x W	18	0.001	0.62	0.888
H x STI x W x S	125	0.002		

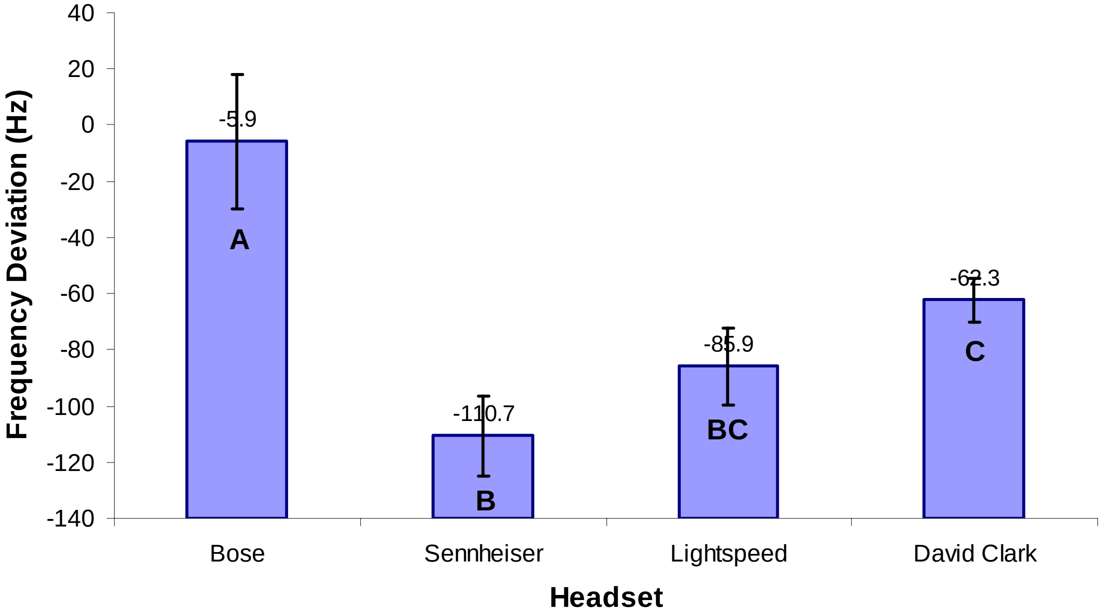


Figure 30. Difference in formant frequency of the recorded pilot voiced speech from the recorded non-workload baseline for each headset. Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Headset Comfort and Communication Performance Ratings

Headband Force Measurements. Analysis to confirm the data do not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted on a scatterplot of the predicted versus residual values from the headband force data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption.

The amount of force, as measured in Newtons (N), applied to the head by a headset was determined according to the ANSI S3.19–1974 standard. These measurements were then analyzed using a one-way ANOVA for the four levels of the headset variable. A significant difference was found between the four headsets, $F(3, 12) = 38.57$, $p < 0.0001$ (Table 12). Post-hoc comparisons utilizing the Tukey HSD method revealed the headband force of the Bose headset was significantly lower than the other three headsets. The headband force of the Sennheiser was not different from the LightSPEED headset or the David Clark headset. However, the headband force of the David Clark headset was significantly lower than that of the LightSPEED headset (Figure 31).

Many researchers in the area of hearing protection device (HPD) comfort have proposed headband clamping force as a major influential factor, especially with circumaural HPDs (Acton et al., 1976). Therefore, the headband force, which presses the ear cup against the head, would be a likely sign of the comfort a pilot could expect from

Table 12

ANOVA summary table for headband force analysis.

Source	SS	df	MS	*F*	*p*
Headset (H)	49.587	3	16.529	38.57	<0.0001
Within Groups (S/H)	5.143	12	0.429		
Total	54.729	15			

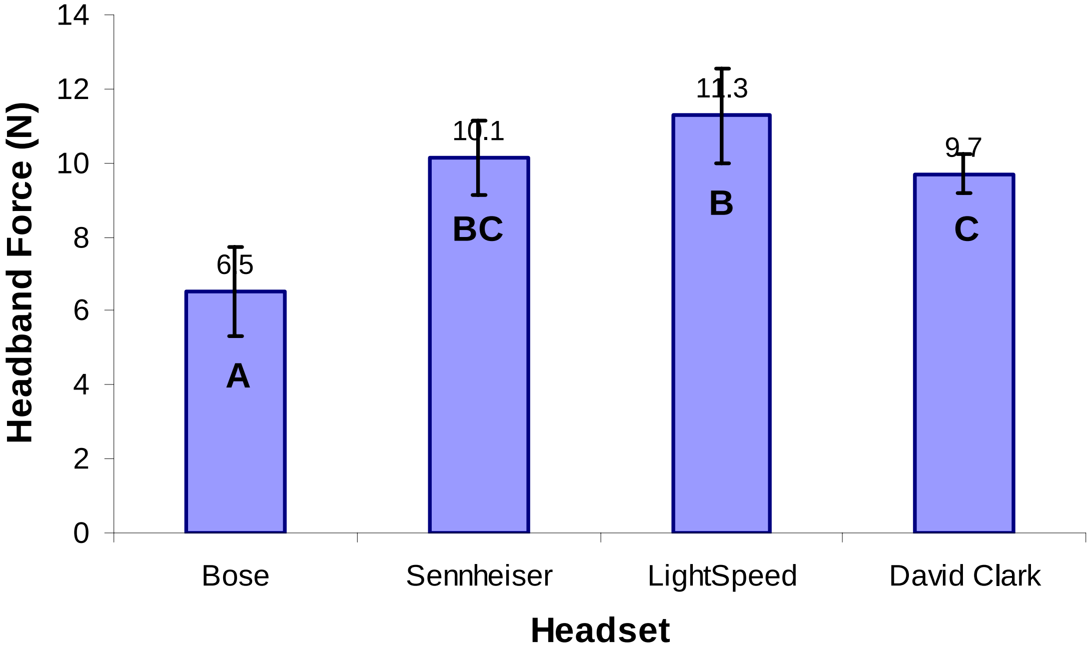


Figure 31. Headband force in Newtons (N), collected according to ANSI S3.19–1974. Different letters signify a significant difference at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

each headset, especially when compared against each other. Unfortunately, the results of the headset comfort ratings and the final rankings the participants made of the headsets based on comfort and communications performance do not support the results of the headband force analysis. This may actually come as no surprise as the literature concerning objective measures for HPD comfort could be considered ambiguous at best (Lhuede, 1980; Savich, 1981).

Headset Comfort Ratings. As discussed previously, the comfort rating scales had been validated in previous studies, as had the method of statistical analysis (Park & Casali, 1991). Therefore, this study utilized the same method of analysis. Each scale, identified in this section by its bipolar descriptor, was treated as a Likert-type scale, and the pilot responses were transformed into numeric values, with 1 situated on the left side of the scale and 7 situated on the right side of the scale. The numeric data derived from each of the bipolar rating scales were then analyzed separately using a repeated measures ANOVA for the four levels of the headset variable. Before the ANOVA analysis began analysis to confirm the all data do not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted on the scatterplots of the predicted versus residual values from the comfort rating data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption.

A Mauchly's test of sphericity showed that the data for the *Painless/Painful* rating data did conform to the sphericity assumption, *p* = 0.07. The bipolar rating scale of

Painless/Painful showed a significant difference across the four headsets, *F* (3, 21) = 3.66, *p* = 0.024 (Table 13). Post hoc analysis utilized the Tukey HSD method to make all comparisons. The results showed that the Bose headset was considered significantly less painful than the David Clark headset, but no different from the Sennheiser or LightSPEED headsets. Furthermore, the Sennheiser and LightSPEED headsets did not differ significantly from each other or the David Clark headset (Figure 32).

Further analyses of variance resulted in the remaining comfort ratings demonstrating no significant differences across the four levels of the headset variable.

- Uncomfortable/Comfortable; *F*(3, 21) = 1.50, *p* = 0.236, *ns* (Figure 33).
- No Uncomfortable Pressure/Uncomfortable Pressure; *F*(3, 21) = 0.68, *p* = 0.572, *ns* (Figure 34).
- Intolerable/Tolerable; *F*(3, 21) = 2.18, *p* = 0.113, *ns* (Figure 35).
- Tight/Loose; *F*(3, 21) = 0.63, *p* = 0.601, *ns* (Figure 36).
- Not Bothersome/Bothersome; *F*(3, 21) = 2.08, *p* = 0.126, *ns* (Figure 37).
- Heavy/Light; *F*(3, 21) = 1.01, *p* = 0.403, *ns* (Figure 38).
- Cumbersome/Not Cumbersome; *F*(3, 21) = 2.72, *p* = 0.064, *ns* (Figure 39).
- Soft/Hard; *F*(3, 21) = 0.41, *p* = 0.744, *ns* (Figure 40).
- Feeling of Complete Isolation/No Feeling of Complete Isolation; *F*(3, 21) = 1.94, *p* = 0.146, *ns* (Figure 41).
- Ear Open/Ear Blocked; *F*(3, 21) = 0.54, *p* = 0.658, *ns* (Figure 42).
- Ear Empty/Ear Full; *F*(3, 21) = 0.39, *p* = 0.762, *ns* (Figure 43).

Table 13

ANOVA summary table for the bipolar comfort rating scale, painless/painful.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	1.911		
Within				
Headset (H)	3	4.868	3.66	0.024
H x S	21	1.330		

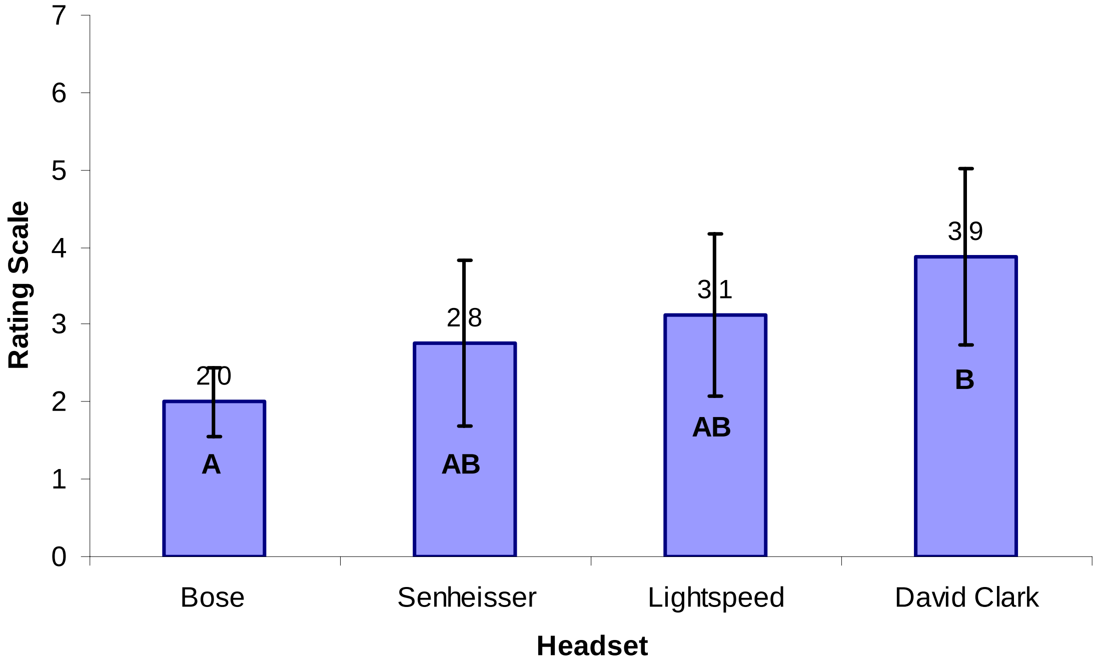


Figure 32. Pilot ratings of pain levels under each headset (0 = Painless, 7 =Painful). ANOVA analysis found no significant differences between headsets at the p < 0.05 level. Vertical range bars represent 95% confidence intervals about the means.

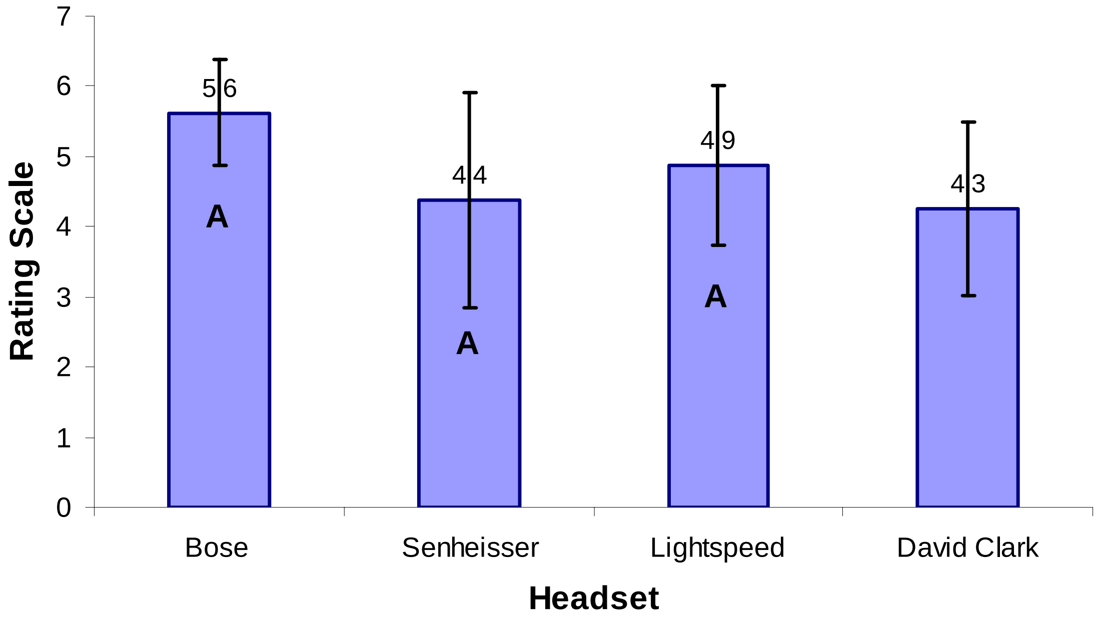


Figure 33. Pilot ratings of comfort levels experienced under each headset (0 = Uncomfortable, 7 =Comfortable). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

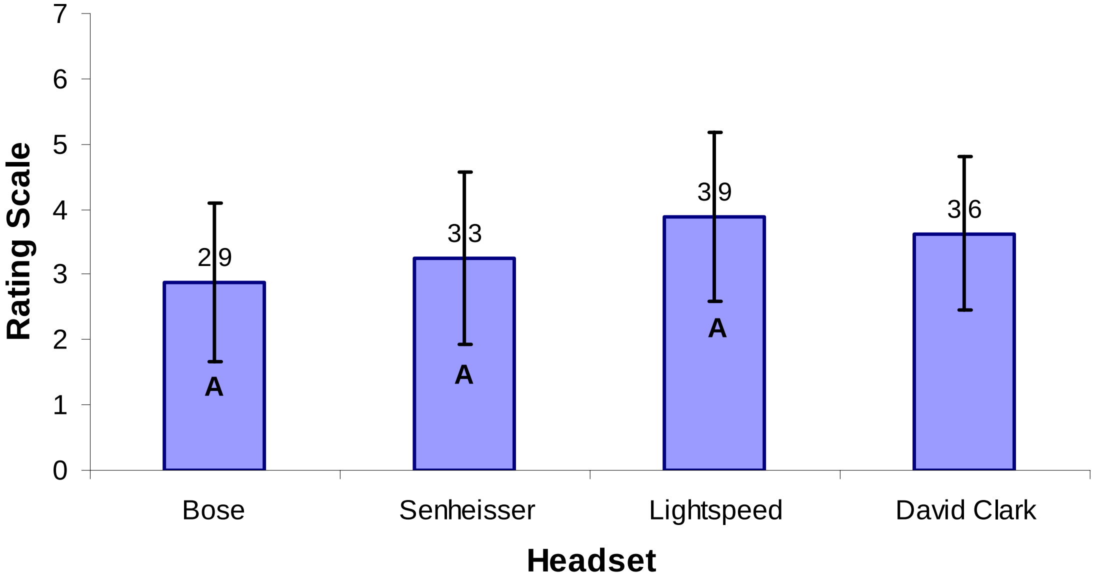


Figure 34. Pilot ratings of pressure levels experienced under each headset (0 = No Uncomfortable Pressure, 7 = Comfortable Pressure). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

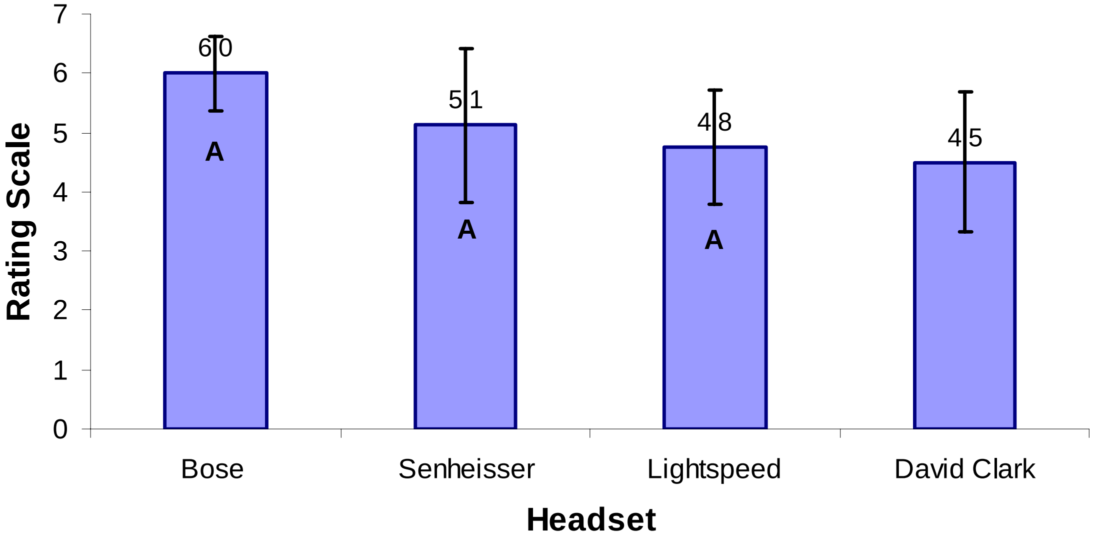


Figure 35. Pilot ratings of tolerance levels for each headset (0 = Intolerable, 7 = Tolerable). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

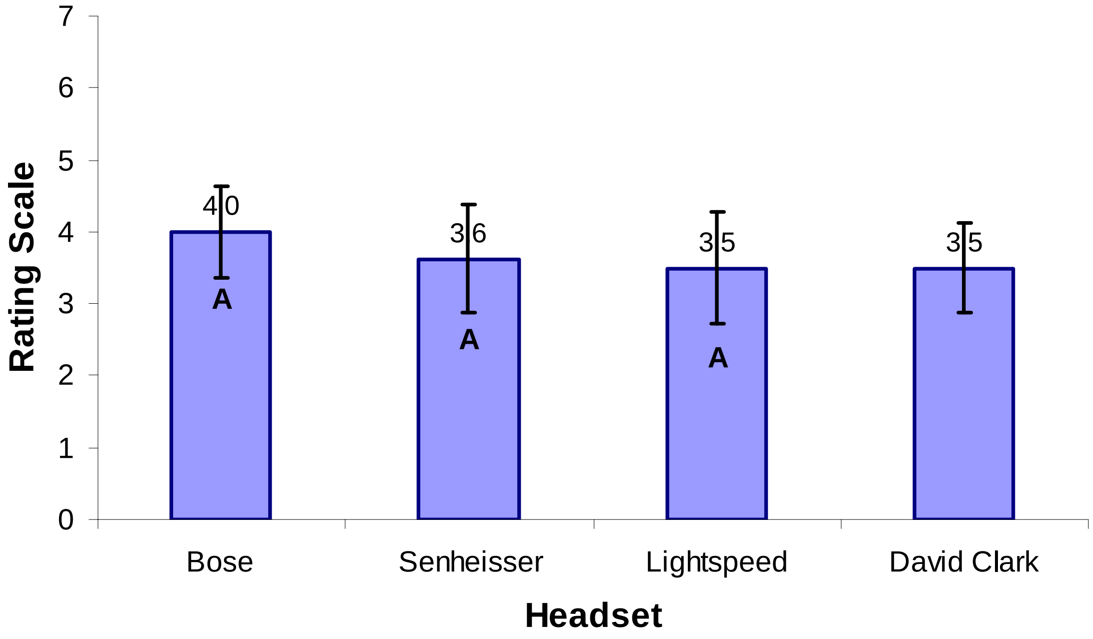


Figure 36. Pilot ratings of tightness experienced under each headset (0 =Tight, 7 = Loose). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

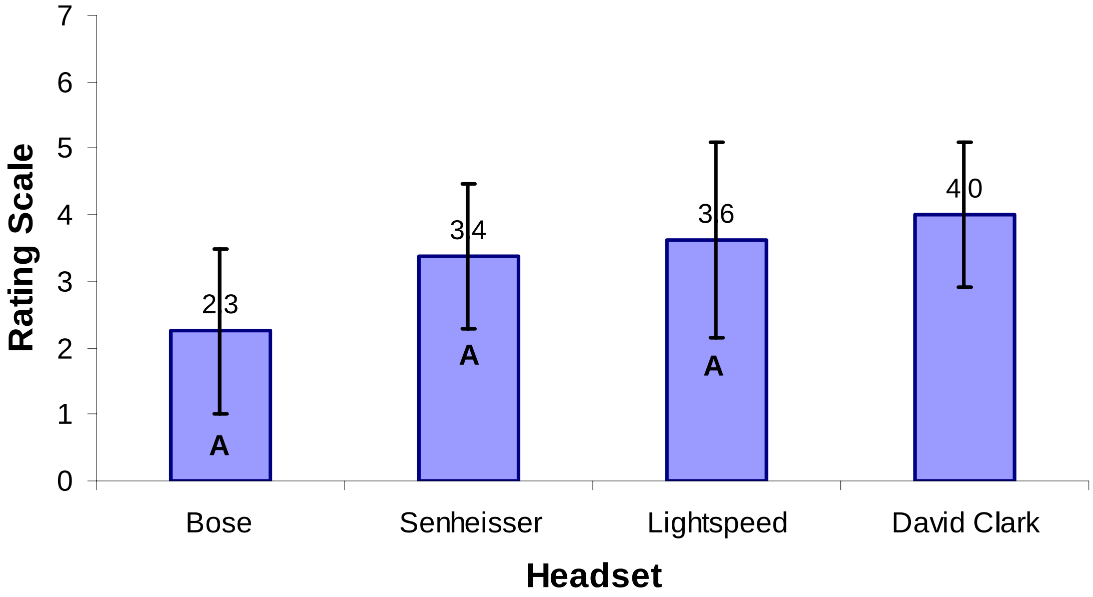


Figure 37. Pilot ratings of how bothersome each headset was (0 = Not Bothersome, 7 = Bothersome). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

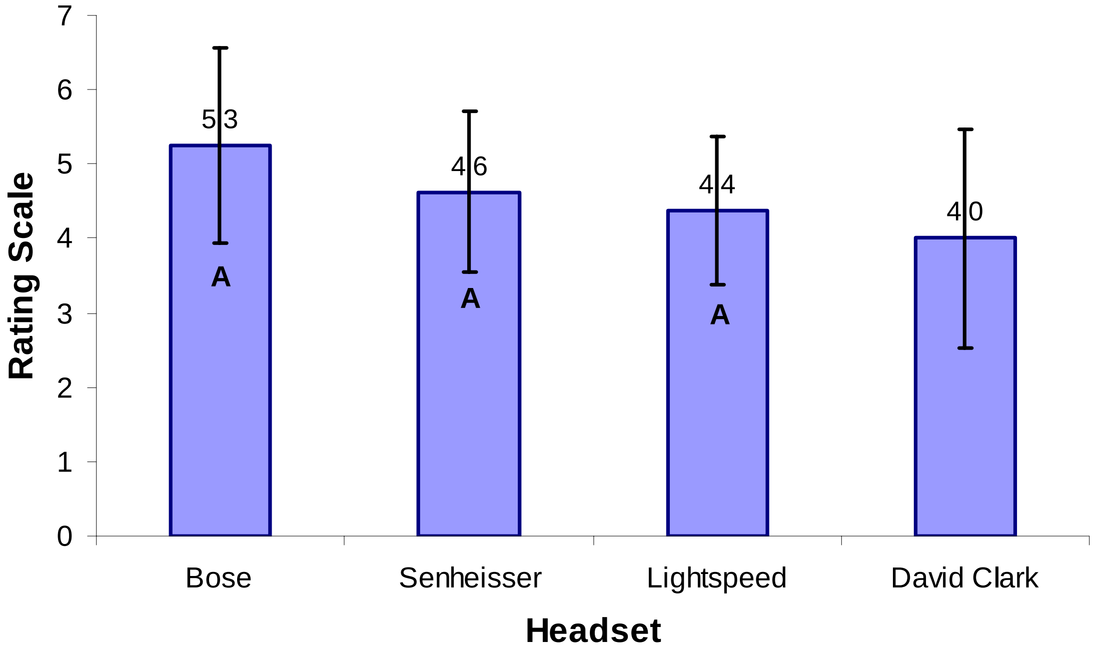


Figure 38. Pilot ratings of weight of each headset (0 = Heavy, 7 = Light). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

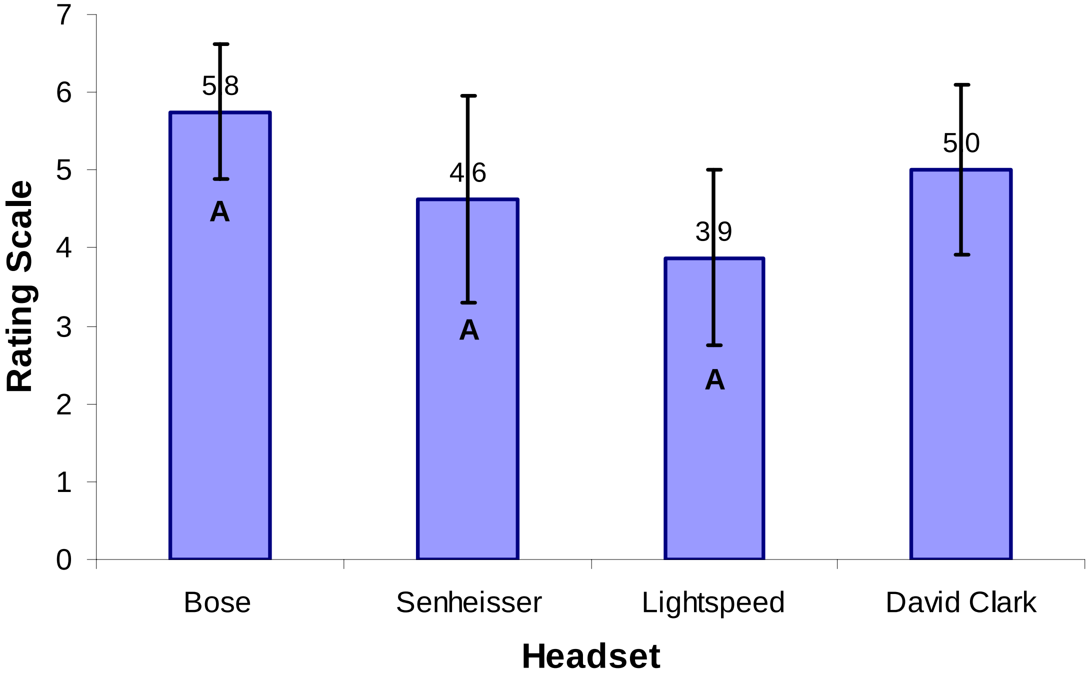


Figure 39. Pilot ratings of how cumbersome each headset was (0 = Cumbersome, 7 = Not Cumbersome). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

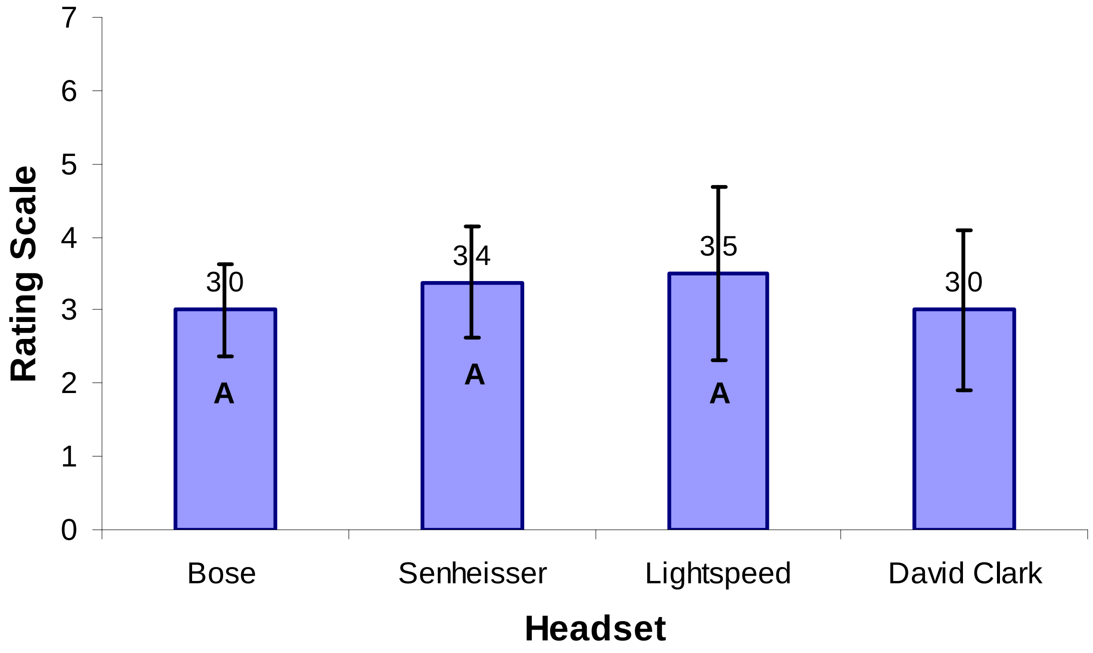


Figure 40. Pilot ratings of the softness of each headset (0 = Soft, 7 = Hard). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

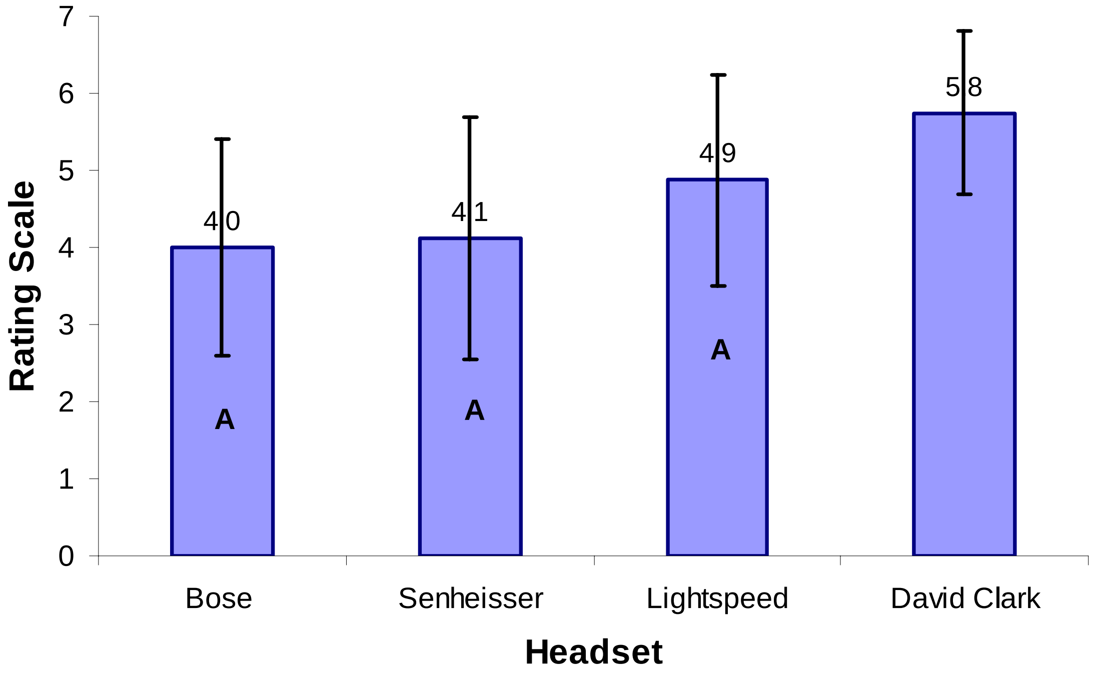


Figure 41. Pilot ratings of the feeling of isolation experienced under each headset (0 = Feeling of Complete Isolation, 7 = No Feeling of Complete Isolation). ANOVA analysis found no significant differences between headsets at the *p* < 0.05 level. Vertical range bars represent 95% confidence intervals about the means.

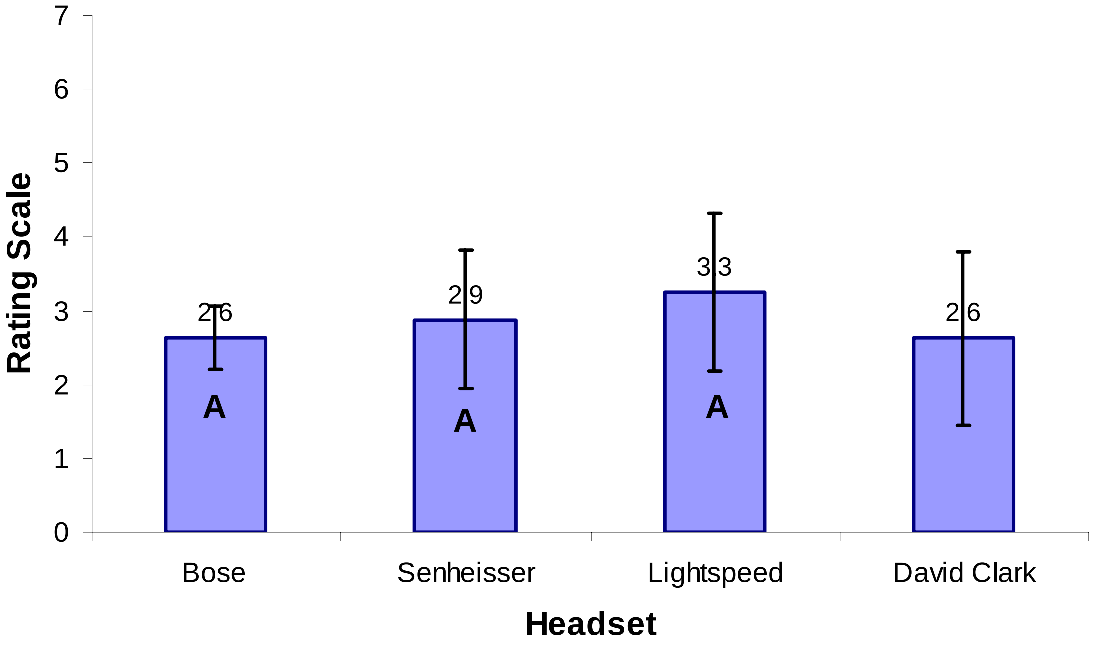


Figure 42. Pilot ratings of the feeling that their ears were open or blocked under each headset (0 = Ear Open, 7 = Ear Blocked). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

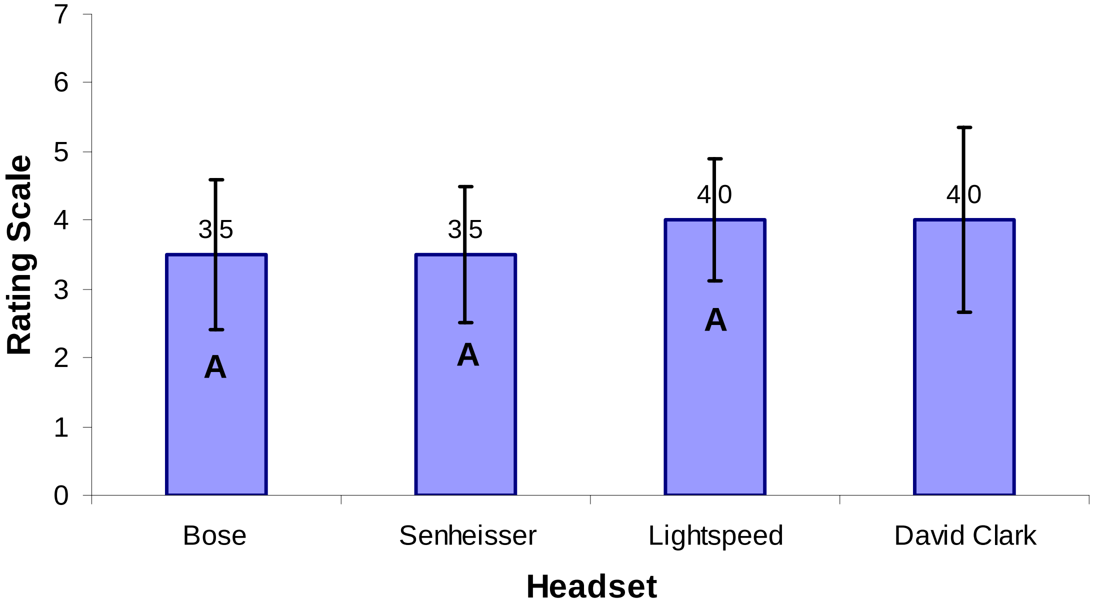


Figure 43. Pilot ratings of the feeling that their ears were empty or full under each headset (0 = Ear Empty, 7 = Ear Full). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

The results of the comfort ratings were somewhat surprising. The analyses showed no differences between any of the headsets, except the *Painless/Painful* rating, and even this rating showed a difference only between the Bose and David Clark headsets. This does not align with the expectations about design variables which were expected to influence comfort, i.e., headband force.

The headband force analysis, previously presented, and the subjective rankings and comments regarding the headsets concerning comfort and performance, to be presented in a subsequent section, both point to the fact that comfort influences the pilot and their perceptions of the quality of the headset. However, the well-established ratings scales (which are the subject of this section) say something completely opposite to the other two measures; that is, the majority of the rating scales show pilot participants do not perceive any difference between the four headsets, in terms of comfort (with one exception between Bose and David Clark in the *Painless/Painful* rating).

These conflicting results may be a function of a sample size, which was too small for the rating scales to function effectively. In other words, with only eight pilots reporting their perceptions through the ratings, the ratings were not sensitive enough to find the comfort differences that were shown through the pilots' subjective rankings and comments about the headsets.

This would be the simplest hypothesis to test. The experiment could basically be rerun in a geographical area where a larger sample size is feasible. To simplify the experiment, simply running a larger sample size through a cross-country flight without the need to collect speech intelligibility, workload, and flight performance data would suffice to investigate whether sample size is the cause for the disparity between the comfort rating scale results and the pilots' rankings and comments on the comfort of the four headsets.

It is also possible that pilots view comfort as a dichotomy of tolerable/intolerable and the four headsets were each acceptable in that they did not cause intolerable pain, annoyance, or infringe upon their flight performance (at least based on the pilots' perceptions). Therefore, the pilots may have concluded that each headset was fit for use in the cockpit. There is some anecdotal support for this speculation. The David Clark headset, which is generally thought to be the most widely used headset in general aviation, is referred to as the "David Clamp" by many pilots for its perceived high clamping force on the head (even though this claim is made by pilots, the headband force measurements did not support it). Even after the pilots' complaints concerning the David Clark headset, these same pilots continue to use it on a daily basis, year after year. This norm in the flight training culture may ingrain in pilots that comfort is to be viewed as a tolerable/intolerable dichotomy, and not as the many degrees of comfort offered by the comfort rating scales.

One thing can be said for sure. The area of comfort regarding aviation communication headsets will require further research to gain an understanding concerning how pilots conceptualize comfort and what aspects of comfort they regard as highest in importance.

Communications Performance Ratings. In addition to the comfort ratings, several ratings were designed to measure the pilot's perception of performance of the integrated communication system (earphone, microphone) in each headset. These ratings were analyzed in the same manner as the previous comfort ratings. As previously stated, each rating was analyzed separately using a repeated measures ANOVA to test the four levels of the headset variable. If significant differences were found, further post-hoc comparisons were carried out using the Tukey HSD procedure. Before the ANOVA analysis began, analysis to confirm the all data do not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted on the scatterplots of the predicted versus residual values from the comfort rating data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption.

A Mauchly's test of sphericity showed that the data for the *Low Fidelity Communications/High Fidelity Communications* rating data did conform to the sphericity assumption, $p = 0.33$. The first communications performance rating was characterized by the bipolar descriptor, *Low Fidelity Communications/High Fidelity Communications*. The ANOVA analysis resulted in significant differences among the four headsets, $F(3, 21) = 3.03$, $p = 0.046$ (Table 14). Post hoc comparisons showed that pilots perceived fidelity of communications while wearing the Bose headset to be significantly higher than communications while wearing the David Clark headset. There were no significant differences between the Bose, Sennheiser, and LightSPEED headsets. There were also no significant differences found between the Sennheiser, LightSPEED, and David Clark headsets (Figure 44).

Table 14

ANOVA summary table for the rating Low Fidelity Communications / High Fidelity Communications.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	1.357		
Within				
Headset (H)	3	4.509	3.03	0.046
H x S	21	1.488		

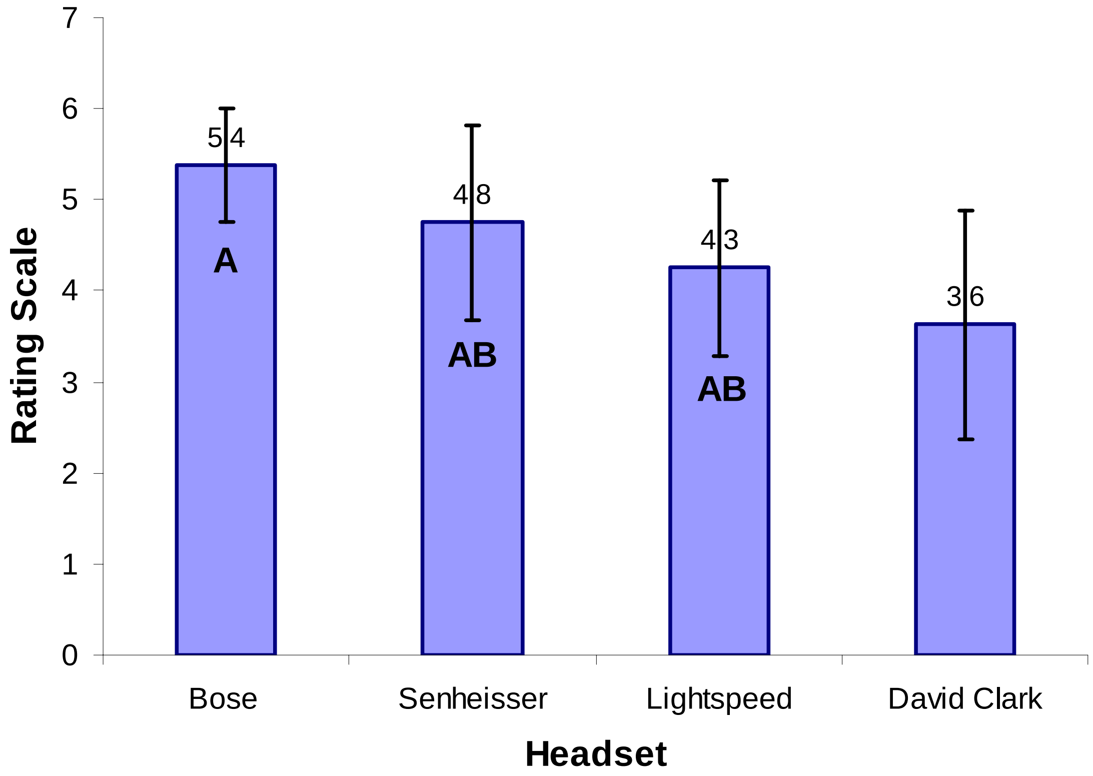


Figure 44. Pilot ratings of the communications fidelity levels for each headset (0 = Low Fidelity Communications, 7 = High Fidelity Communications). Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

A Mauchly's test of sphericity showed that the data for the *Extraneous Noise/No Extraneous Noise* rating data did conform to the sphericity assumption, *p* = 0.13. The analysis of the rating scale *Extraneous Noise/No Extraneous Noise* showed no significant difference among the headsets, *F*(3, 21) = 2.43, *p* = 0.094 (Figure 45).

A Mauchly's test of sphericity showed that the data for the *Sound Distortion/No Sound Distortion* rating data did conform to the sphericity assumption, *p* = 0.06. The analysis of the rating for the scale of *Sound Distortion/No Sound Distortion* showed a significant difference among the headsets, *F*(3, 21) = 3.51, *p* = 0.028 (Table 15). Post-hoc analysis showed that pilots rated the Bose headset as producing significantly less sound distortion during communications than the David Clark headset. There was no significant difference in sound distortion between the Bose, Sennheiser, and LightSPEED headsets. Additionally, the results did not show any significant differences between the Sennheiser, LightSPEED, and David Clark headsets (Figure 46).

A Mauchly's test of sphericity showed that the data for the *Interferes with Communications/No Interference with Communications* rating data did conform to the sphericity assumption, *p* = 0.18. The analysis of the rating scale *Interferes with Communications/No Interference with Communications* showed a significant difference among the headsets, *F*(3, 21) = 4.31, *p* = 0.013 (Table 16). Post-hoc comparisons showed that pilots rated the Bose headset as producing significantly less interference during communications than the David Clark headset. There were no significant differences found between the Bose, Sennheiser, and LightSPEED headsets. Additionally, the results did not find any significant differences between the Sennheisser, LightSPEED, and David Clark headsets (Figure 47).

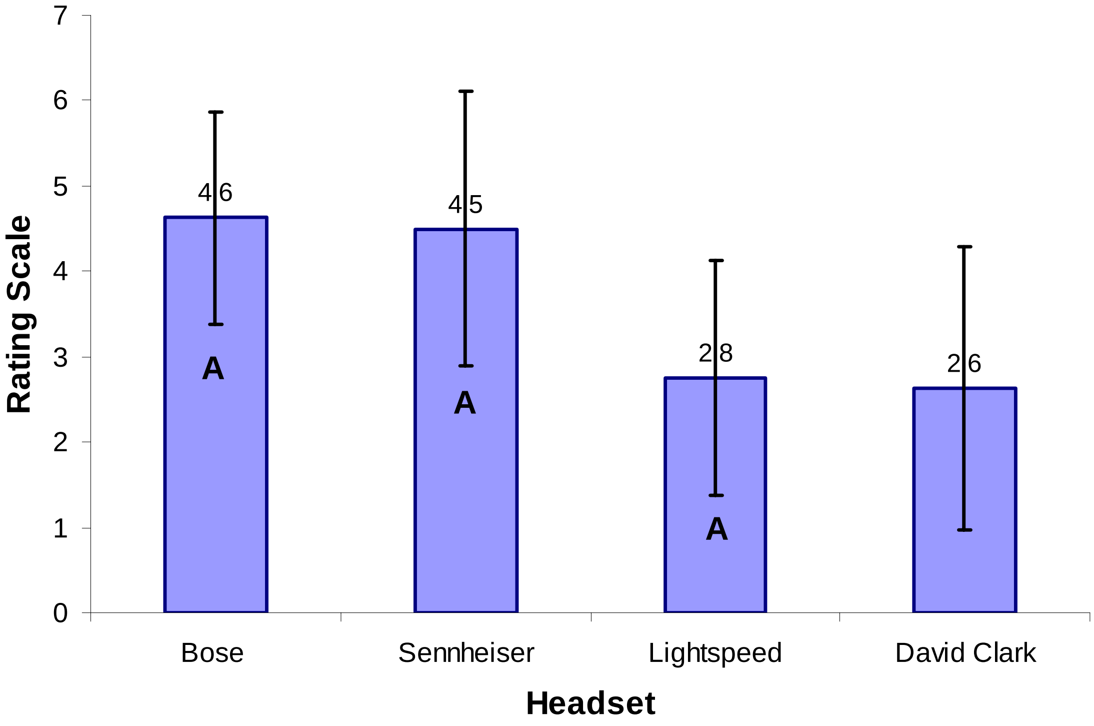


Figure 45. Pilot ratings of the extraneous noise levels for each headset (0 = Extraneous Noise, 7 = No Extraneous Noise). ANOVA analysis found no significant differences between headsets at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 15

ANOVA summary table for the rating Sound Distortion/No Sound Distortion.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	1.429		
Within				
Headset (H)	3	5.556	3.51	0.028
H x S	21	1.583		

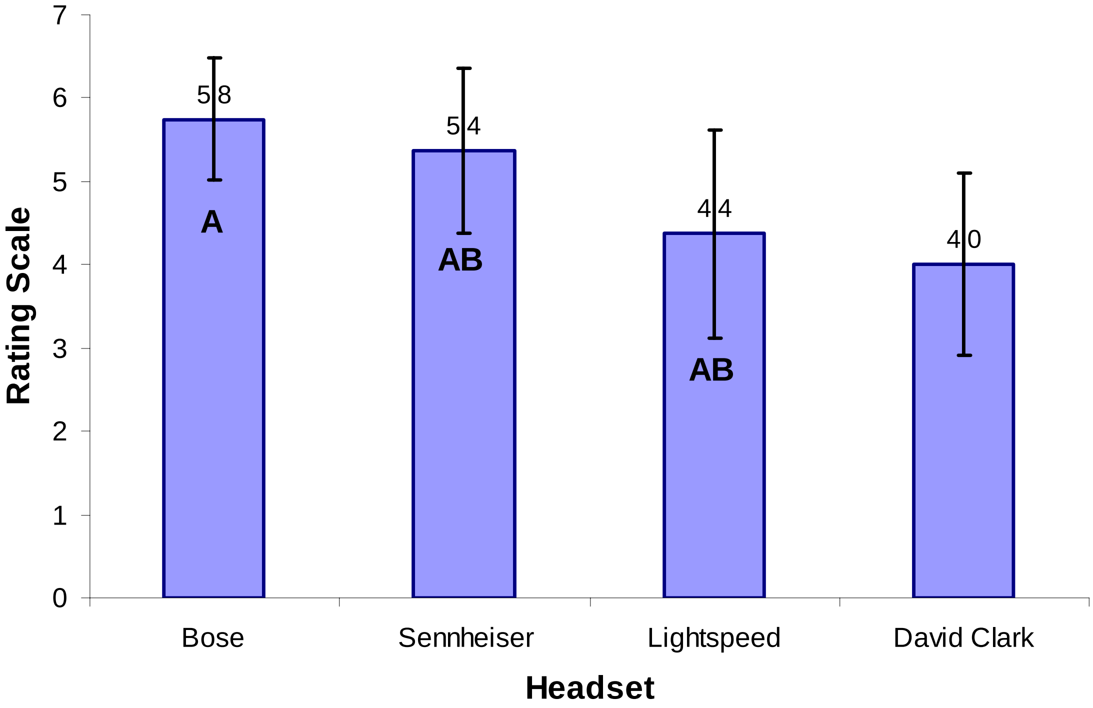


Figure 46. Pilot ratings of the sound distortion levels for each headset (0 = Sound Distortion, 7 = No Sound Distortion). Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 16

ANOVA summary table for the rating scale: Interferes with Communications/No Interference with Communications.

Source	df	MS	*F*	*p*
<u>Between</u>				
Subject (S)	7	0.924		
<u>Within</u>				
Headset (H)	3	8.603	4.31	0.013
H x S	21	1.996		

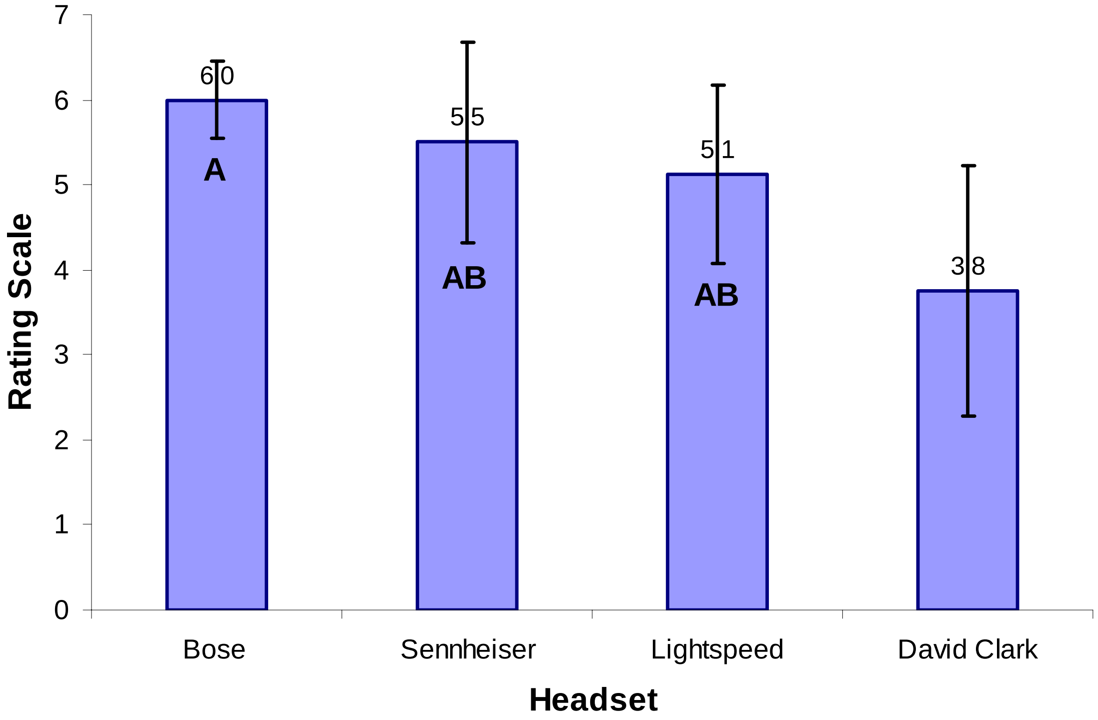


Figure 47. Pilot ratings of communications interference experienced under each headset (0 = High Communications Interference, 7 = No Communications Interference). Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

A Mauchly’s test of sphericity showed that the data for the *Interferes with Communications/No Interference with Communications* rating data did conform to the sphericity assumption, $p = 0.18$. The analysis of the scale concerning *Low Overall Communications Quality/High Overall Communications Quality* showed a significant difference among the headsets, $F(3, 21) = 6.60$, $p = 0.002$ (Table 17). Further post-hoc comparisons showed that pilots rated overall communications quality of the Bose headset as significantly higher than that of the David Clark headset. Post-hoc analysis also found no significant differences between the Bose, Sennheiser, and LightSPEED headsets, in terms of overall communications quality. Moreover, no significant differences were found between the Sennheiser, LightSPEED, and David Clark headsets (Figure 48).

A Mauchly’s test of sphericity showed that the data for the *Background Hum Present/No Background Hum Present* rating data did conform to the sphericity assumption, $p = 0.19$. The analysis of the scale concerning *Background Hum Present/No Background Hum Present* showed a significant difference among the headsets, $F(3, 21) = 5.76$, $p = 0.003$ (Table 18). Further post-hoc analysis showed that pilots rated the Bose, Sennheiser, and David Clark headsets as having significantly less background hum present than the LightSPEED headset. Also, there were no significant differences in background hum found between the Bose, Sennheiser, and the David Clark headset as rated by the pilots (Figure 49).

Table 17

ANOVA summary table for the rating scale: Low Overall Communications Quality/High Overall Communications Quality.

Source	df	MS	*F*	*p*
<u>Between</u>				
Subject (S)	7	2.143		
<u>Within</u>				
Headset (H)	3	24.829	6.60	0.002
H x S	21	3.762		

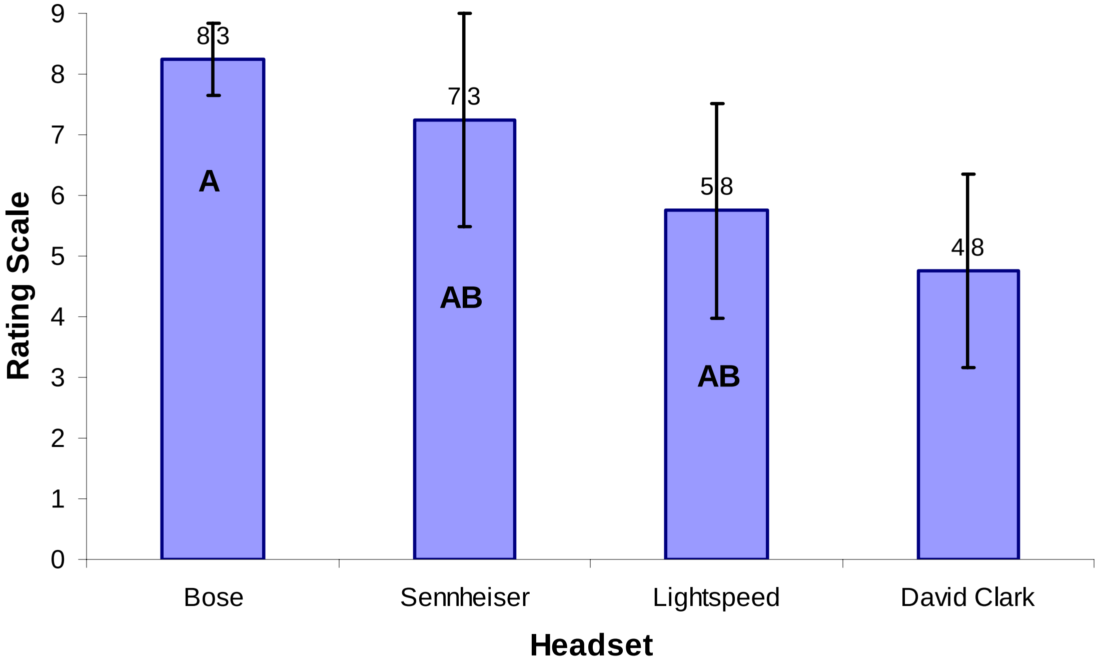


Figure 48. Pilot ratings of the overall communications quality experienced with each headset (0 = Low Overall Communications Quality, 9 = High Overall Communications Quality). Different letters signify significant differences at the p < 0.05 level. Vertical range bars represent 95% confidence intervals about the means.

Table 18

ANOVA summary table for the rating scale: Background Hum Present/No Background Hum Present.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	2.268		
Within				
Headset (H)	3	16.836	5.76	0.003
H x S	21	2.923		

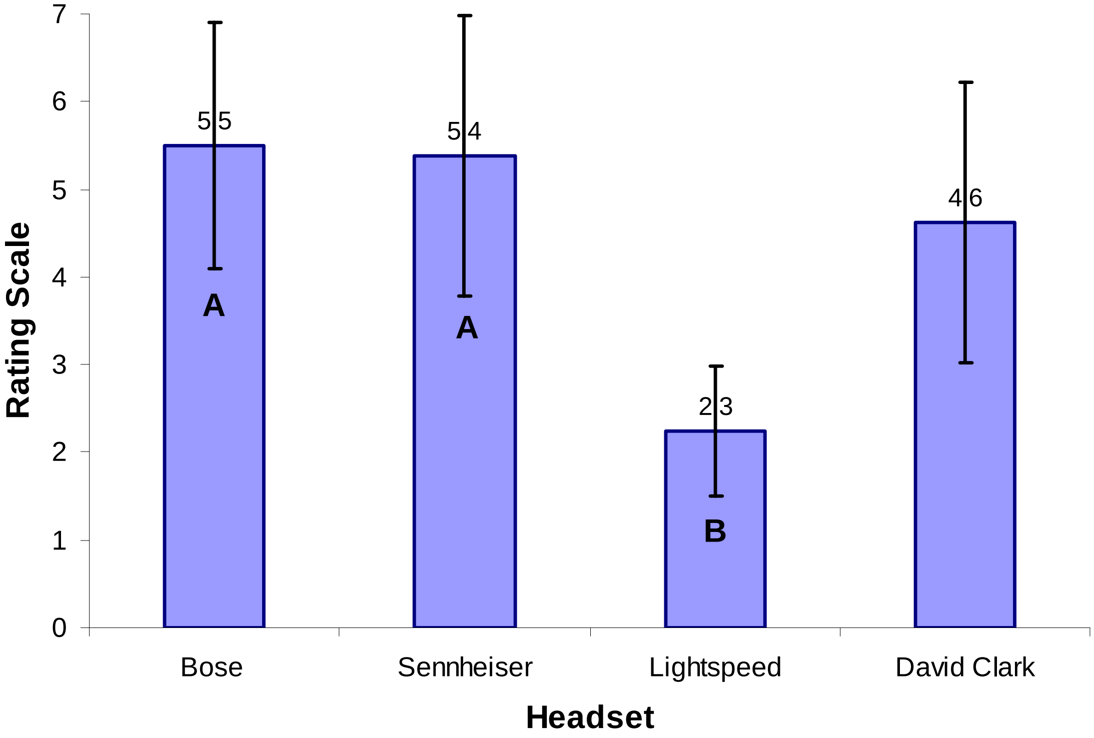


Figure 49. Pilot ratings of the background hum experienced under each headset (0 = Background Hum Present, 7 = No Background Hum Present). Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

A discussion of the *background hum* rating scale will be given first. This rating scale presents an important, but quite different relationship from the other five rating scales. Therefore, following the discussion of the *background hum* rating scale the remaining five rating scales can be presented together.

The *background hum* rating scale was designed to gauge the amount of electronic noise, or hum, the pilots perceived while wearing each headset. The Bose, Sennheiser, and David Clark headsets were rated numerically highest, meaning they have little background hum present. On the other hand, the LightSPEED headset is a completely different case for background hum. As soon as the ANR electronics are turned on, a very noticeable electronic hum is heard without any stimulation from a communications signal. This electronic hum appears to intensify as the headset is moved closer in proximity to other electronics. It could be the case that the ANR and communications electronics within the LightSPEED headset are not sufficiently isolated. However, without a detailed analysis of the electronics, which is beyond the scope of this investigation, it is impossible to know for sure. One thing is for sure, though, with the amount of avionics in a light aircraft, the growing use of new electronics such as GPS, and the general aviation push towards the “glass cockpit”, the electrical hum is an issue with the LightSPEED design which will only be exacerbated by the increasing amount of electronics in the general aviation cockpit.

As further proof that the electronic background hum is easily noticeable, even in the 95 dBA engine noise, more than one pilot remarked, after wearing the LightSPEED headset, that the constant background hum “quickly became very irritating throughout the 3.5 hour flight simulation.”

The other five ratings which assessed different aspects of the headsets' communications systems show that the Bose headset was rated consistently highest, or in the highest grouping, for all ratings which reached statistical significance. The Sennheiser and LightSPEED headsets followed closely behind the Bose headset in their mean ratings different (from each other) through the majority of the communication performance rating scales. However, the Sennheiser and LightSPEED headsets were also rarely found to be significantly difference from the passive David Clark headset. These relationships are the obvious trend through four of the remaining five communication performance rating scales, with the major exception being the *extraneous noise* rating scale which showed no significant differences between the headsets.

According to the results of the *communication fidelity* rating scale, communications with ATC while wearing the Bose headset were rated as significantly higher fidelity than communications while wearing the passive David Clark headset. The mean ratings for the Sennheiser and LightSPEED headsets were not significantly different from the Bose headset or the David Clark headset, indicating that communications fidelity for the Sennheiser and LightSPEED headset are within the midrange between the Bose headset, which is at the upper rating (highest communications fidelity) and the David Clark, which is at the lower end of the ratings (lowest communications fidelity).

In the *sound distortion* rating, the Bose headset was rated as distorting the ATC communications significantly less than the David Clark headset. The mean ratings for the Sennheiser and LightSPEED headsets were once again not significantly different from the Bose headset or the David Clark headset. These results follow the exact same relationship as the *communications fidelity* rating scale, indicating that the distortion of the communications for the Sennheiser and LightSPEED headset is in the midrange between the Bose headset at the upper rating (least amount of sound distortion in communications) and the David Clark at the lower end of the ratings (most sound distortion in communications).

The relationship seen in the last two ratings was repeated in the *interference with communications* rating. Again, the Bose headset was rated as interfering significantly less with communications than the David Clark headset. The Sennheiser and LightSPEED fell within the midrange of the between the Bose headset and the David Clark headset, with the David Clark rated as interfering most with communications.

The *communications quality* rating was one which assesses the overall communications quality of each headset. Here a similar trend to the previously discussed ratings, other than the background hum rating, was expressed. The mean rating for communications quality was highest for the Bose headset, but not significantly different from the Sennheiser headset. Additionally, the mean ratings of the Sennheiser, LightSPEED, and David Clark headsets were found not to be significantly different, although the means showed that pilots rated the headsets in the order given above, from highest to lowest in communications quality.

The implications of these ratings are very important. The *communications performance ratings*, with the exception of the *background hum* rating, show that pilots perceive the quality of the ATC communications to be significantly higher with the Bose ANR headset than with the passive David Clark. The quality of the communications system had a direct impact on the communications exchange between the pilot and air traffic controller, and/or other aircraft. If the communications system is poor, the speech intelligibility will decrease and could potentially create a situation which negates the positive effects of the ANR. Therefore the potential exists for a poor communications system to decrease speech intelligibility below the performance threshold of STI = 0.50 (50% speech intelligibility) seen in the speech intelligibility analysis, thereby decreasing the pilot's speech intelligibility performance, requiring an increase in workload for the pilot to comprehend the ATC commands, and also increasing the air traffic controller's workload by repeating the commands and verifying that the pilot understands the flight maneuver he or she must accomplish.

An additional implication of the *communications performance ratings* concerns the lack of difference between the Sennheiser ANR, the LightSPEED ANR, and the passive David Clark headset. This lack of difference has a major implication for the design of headsets, especially ANR headsets. It is evident that adding ANR capabilities will not automatically create a superior headset over their passive counterparts. Attention must be paid not only to the quality of the ANR system, but also to the quality of the integrated communication system. This may be a difficult engineering task as the ANR and communications are generally integrated into one system; for instance, it is crucial that the ANR not cancel the critical portions of the speech bandwidth, and any "noise" it produces should not mask speech frequencies. However difficult an engineering task, it is one of utmost importance, and is attainable; this declaration is evidenced by the perceived superior quality of the Bose headset, at least as rated by the pilot participants in this study.

Fatigue Level Ratings. The analysis of the scale concerning the *Well-Rested/Entirely Exhausted* ratings showed no significant differences among the headsets (Figure 52), *F*(3, 21) = 0.23, *p* = 0.872 (Table 19; Figure 50). There was an unfortunate flaw discovered in the design of this rating scale. This scale was designed to be presented to the pilot with the other rating scales following the completion of each flight scenario. However, the nature of the variable to be tested, fatigue, is such that factors outside the simulation influence a person's fatigue level measured. A better design would have been to include a pre-test, post-test paradigm to gain a measure of the fatigue induced by the cross-country flight simulation. Induced fatigue was the measurement objective of this fatigue rating scale, but the execution was faulty. Therefore, the results of this scale are considered to be confounded, making a credible interpretation impossible.

Simulation and Engine Noise Realism. The scales for engine noise realism and simulator realism were not analyzed according to any of the independent variables, as it was deemed not be relevant to these variables or the hypotheses of this study. However, descriptive statistics are relevant as an indication of the pilots' perception of the realism of the flight environment as presented by the iGATE simulator, and the realism of the engine noise presented over the subwoofer and loudspeaker system (Table 20).

The realism of the flight simulation and the engine noise were rated fairly highly, approximately 5 out of 7 possible. The main criticism concerning the entire simulation was that the engine noise was not the same as an actual Cessna 172. Pilots remarked that the engine noise had the familiar "raspiness" and "droning" of a Cessna 172, but it was slightly too low in the frequency spectrum.

Table 19

ANOVA summary table for the rating scale: Well-Rested/Entirely Exhausted.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	4.125		
Within				
Headset (H)	3	0.174	0.23	0.872
H x S	21	0.756		

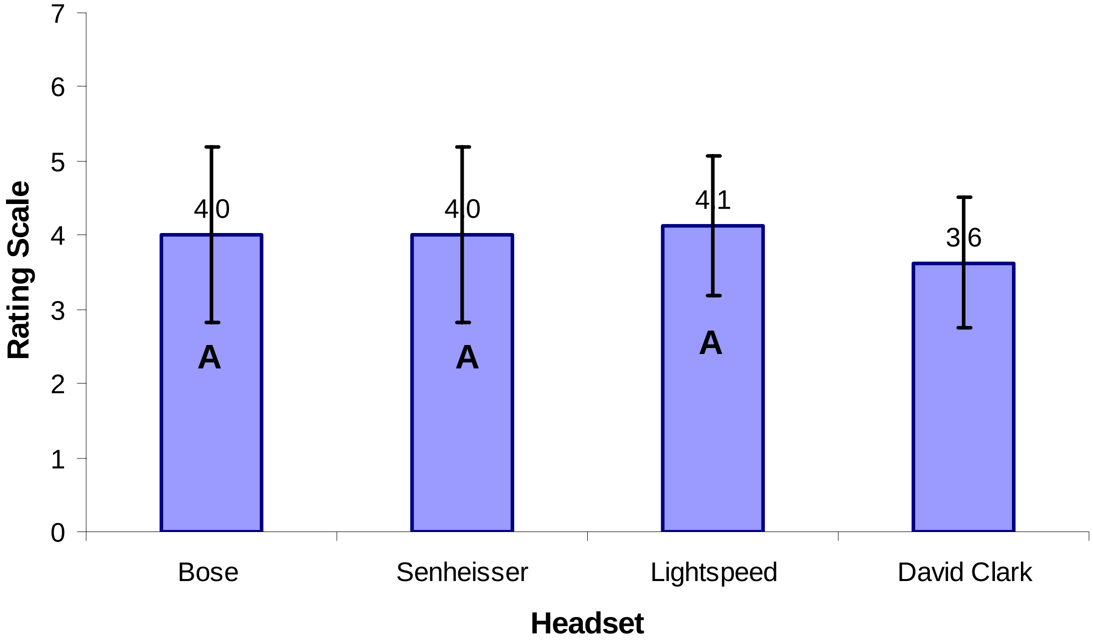


Figure 50. Pilot ratings of the fatigue levels following each 3.5 hour flight session with each headset (0 = Extremely Exhausted, 7 = Well–Rested). Different letters signify significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 20

Pilot's perceptions of the realism of the simulator and of the engine noise (0 = Completely Unrealistic, 7 = Highly Realistic).

Variable	Mean	Standard Deviation	Minimum	Maximum
Simulation Realism	5.16	1.76	3.0	7.0
Engine Noise Realism	4.85	2.06	1.0	7.0

However, the criticism that was received almost unanimously from pilot participants was that the noise level was distinctly louder than that of an actual Cessna 172, but this was not borne out by actual measurements in the aircraft. Perhaps the noise level should be reduced in future experiments, but first this issue requires careful review.

Pilots also rated the realism of the entire flight simulation as high (just above 5 out of 7). Many of the excursions from realism were done intentionally for the purposes of experimental control. For example, to impose high workload during the ILS approach, turbulence was intentionally set higher than a pilot in a Cessna 172 would safely attempt. For this experiment, these deviations from realism were necessary to garner important results. In spite of the inaccuracies perceived by the pilots, their major criticism was focused on the simulator. Again, almost unanimously the pilots reported that the pitch of the aircraft simulation was too unstable, and the response to pitch control inputs was far more sensitive than in an actual Cessna 172. The participants did comment that the combination of the flight route, the flight tasks, weather, turbulence, the interactive ATC commands, and all other aspects of the four flight simulations made for a realistic simulator experience. This experience was compared by the pilots to their previous experiences with simulators, which ranged from a fixed-base instrument simulator with no field of view, to a full-motion Boeing 737 simulator for a major airline.

Headset Protected Exposure Levels without Communications Input

As the past literature has shown, the ANR component of hearing protector attenuation is best at the lower frequencies, while passive HPD performance is best in the mid to high frequencies of sounds (Gower & Casali, 1994).

Therefore, the PEL measurements taken were analyzed by octave band frequencies to differentiate between headset PELs .

PELs were measured using an acoustical test fixture while situated in the 95 dBA engine noise. No ATC communications were transmitted over the headset while these PELs were measured. Figure 51 shows a plot of the PELs for all four headsets used in the flight sessions across all octave band frequencies between 63 Hz and 12.5 kHz.

Before any ANOVA analyses began, analysis to confirm that all data do not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted on the scatterplots of the predicted versus residual values from the comfort rating data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption.

PELs at 63 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets and resulted in no significant differences found between the PEL measurements at 63 Hz, $F(3, 12) = 2.30$, $p = 0.129$ (Table 21; Figure 52).

PELs at 125 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets resulting in significant differences found between the PEL measurements for the four headsets at 125 Hz, $F(3, 12) = 16.85$, $p = 0.0001$ (Table 22). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise levels under the three ANR headsets were all significantly the lower than the passive David Clark headset, with no differences found between the three ANR headsets (Figure 53).

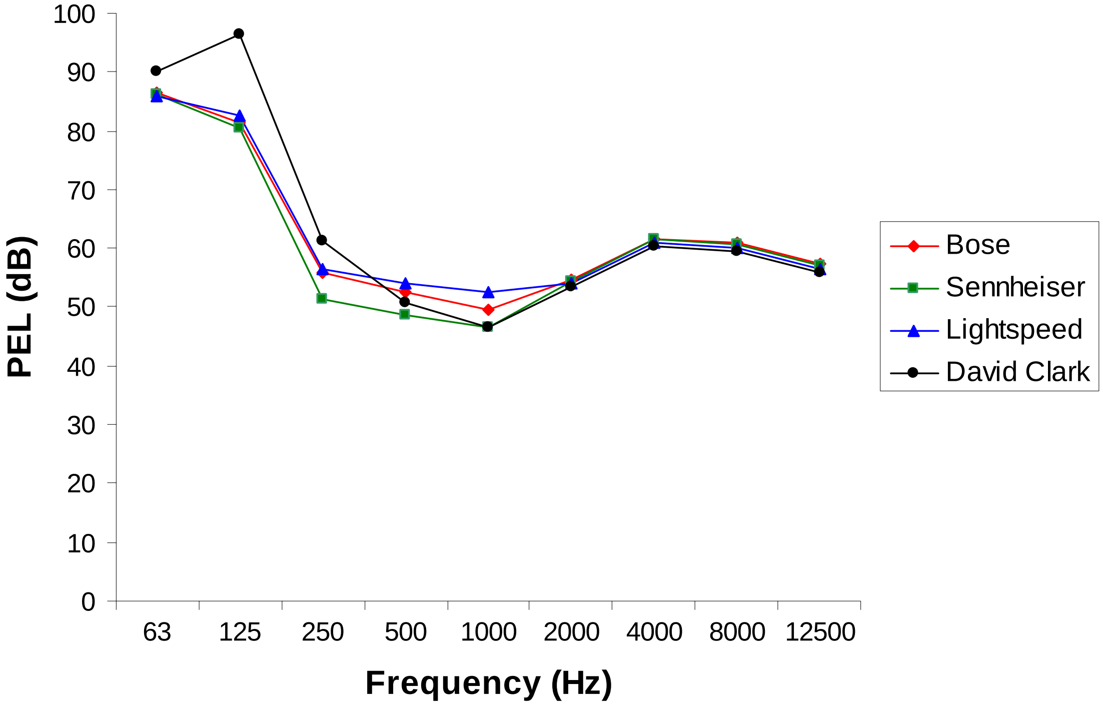


Figure 51. Headset protected exposure levels (PELs) in the aircraft engine noise at octave band intervals (63 Hz – 12500 Hz).

Table 21

ANOVA summary table for the headsets' protected exposure levels at 63 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	14.716	2.30	0.129
Within Group (S/H)	12	6.387		

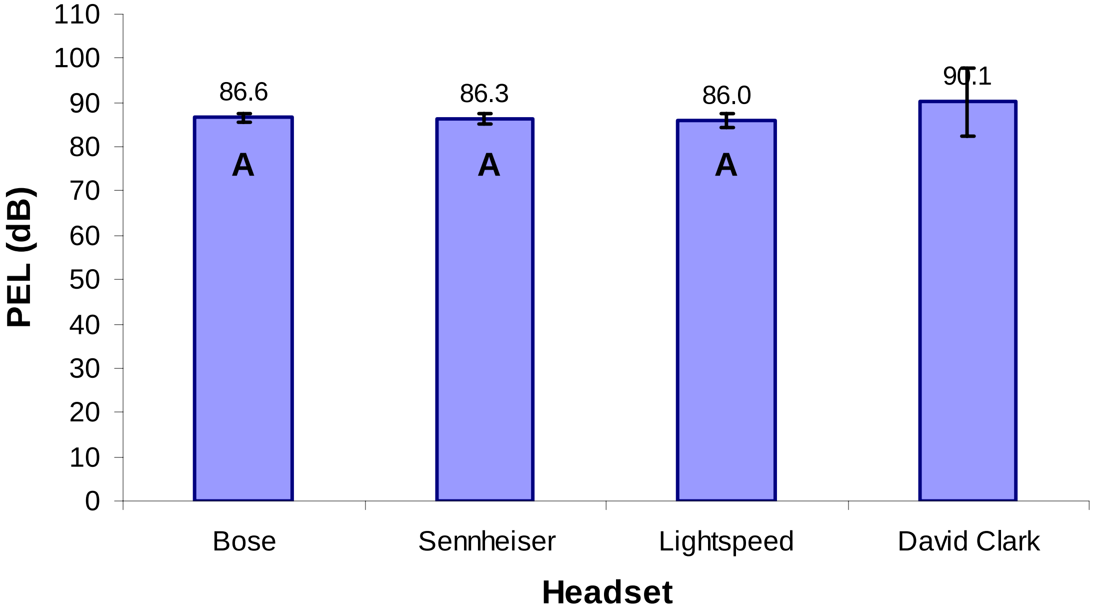


Figure 52. Headset protected exposure levels (PELs) in the aircraft engine noise at 63 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 22

ANOVA summary table for the headsets' protected exposure levels at 125 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	226.662	16.85	0.0001
Within Group (S/H)	12	13.454		

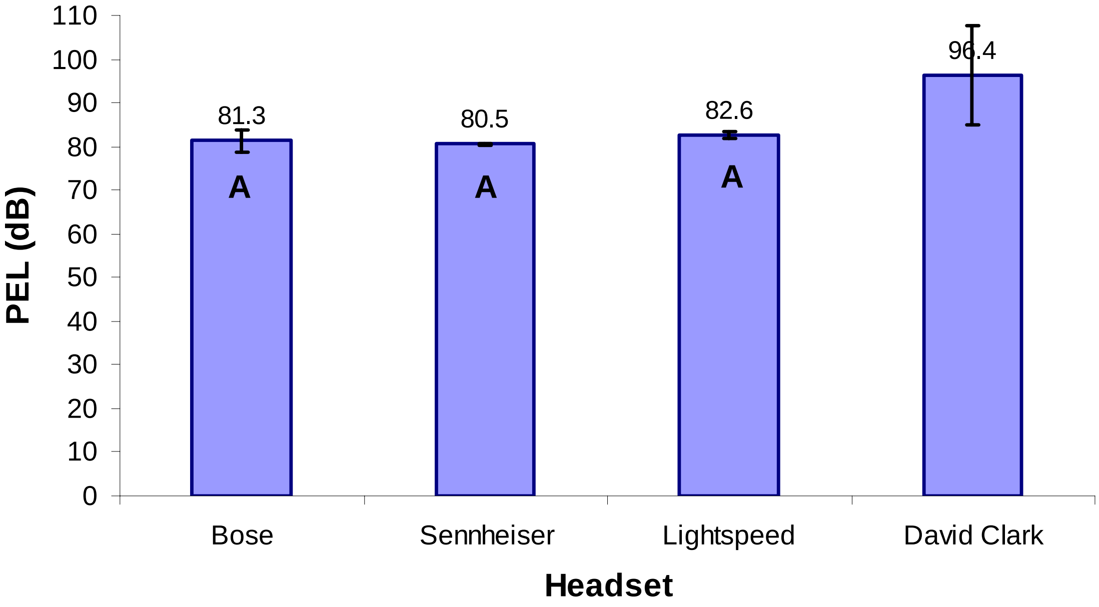


Figure 53. Headset protected exposure levels (PELs) in the aircraft engine noise at 125 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

PELs at 250 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets and resulted in no significant differences found between the PEL measurements for the four headsets at 250 Hz, *F*(3, 12) = 3.51, *p* = 0.087 (Table 23; Figure 54).

PELs at 500 Hz. A one-way ANOVA was again performed to compare the PEL data for the four headsets resulting in significant differences found between the PEL measurements for the four headsets at 500 Hz, *F*(3, 12) = 3.51, *p* =0.049 (Table 24). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise level under the Sennheiser headset was significantly lower than the LightSPEED headset. There were no significant differences found between the Sennheiser headset, the David Clark headset, and the Bose headset. Additionally, no significant differences were found between the David Clark headset, the Bose headset, and the LightSPEED headset (Figure 55).

PELs at 1000 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets resulting in significant differences found between the PEL measurements for the four headsets at 1000 Hz, *F*(3, 12) = 8.37, *p* = 0.003 (Table 25). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise levels under the Sennheiser headset and David Clark headset were significantly lower than the LightSPEED headset. However, no significant differences were found between the Sennheiser headset, the David Clark headset, and the Bose headset. Additionally, no significant differences were found between the Bose headset and the LightSPEED headset (Figure 56).

Table 23

ANOVA summary table for the headsets' protected exposure levels at 250 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	65.312	2.78	0.087
Within Group (S/H)	12	23.469		

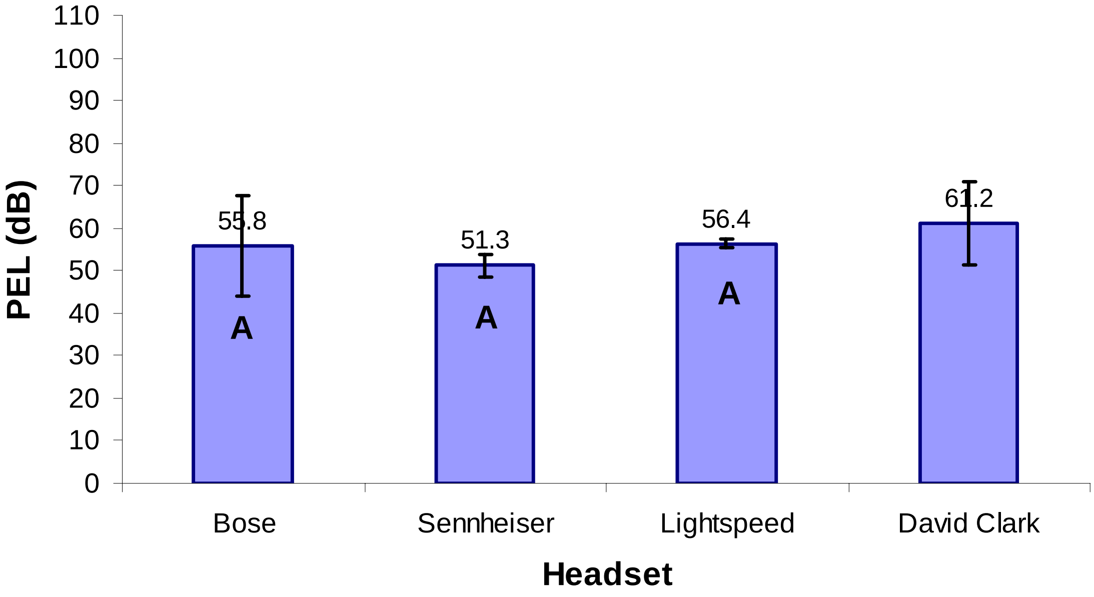


Figure 54. Headset protected exposure levels (PELs) in the aircraft engine noise at 250 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 24

ANOVA summary table for the headsets' protected exposure levels at 500 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	22.806	3.51	0.049
Within Group (S/H)	12	6.495		

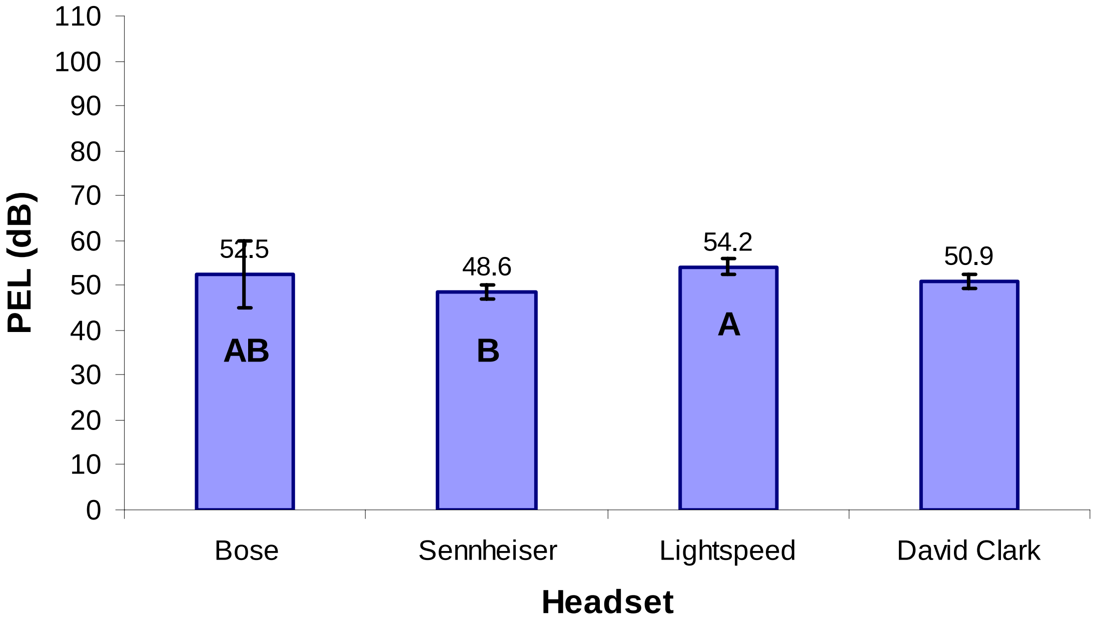


Figure 55. Headset protected exposure levels (PELs) in the aircraft engine noise at 500 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 25

ANOVA summary table for the headsets' protected exposure levels at 1000Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	33.266	8.37	0.003
Within Group (S/H)	12	3.974		

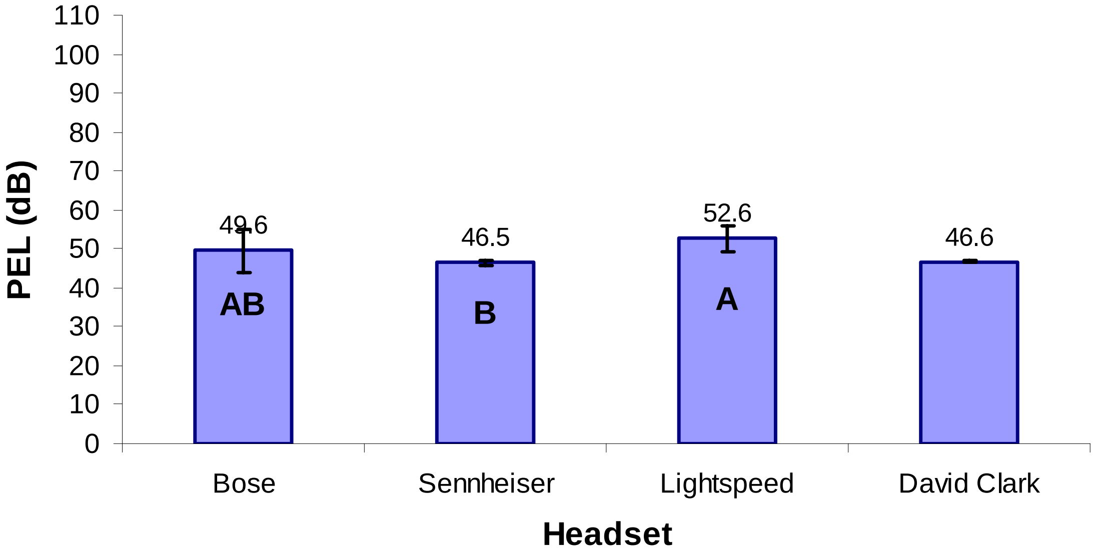


Figure 56. Headset protected exposure levels (PELs) in the aircraft engine noise at 1000 Hz. Different letters represent significant differences at the p < 0.05 level. Vertical range bars represent 95% confidence intervals about the means.

PELs at 2000 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets resulting in significant differences found between the PEL measurements for the four headsets at 2000 Hz, $F(3, 12) = 5.18$, $p = 0.016$ (Table 26). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise level under the David Clark headset was significantly lower than the Bose headset. There were no significant differences found between the David Clark, LightSPEED, and Sennheiser headsets. Lastly, no significant differences were found between the LightSPEED, Sennheiser, and Bose headsets (Figure 57).

PELs at 4000 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets resulting in significant differences found between the PEL measurements for the four headsets at 4000 Hz, $F(3, 12) = 5.05$, $p = 0.017$ (Table 27). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise level under the David Clark headset was significantly lower than those under the Sennheiser and Bose headsets. No significant differences were found between the LightSPEED headset and the David Clark headset. Additionally, no significant differences were found between the LightSPEED headset, Sennheiser headset, and Bose headset (Figure 58).

Table 26

ANOVA summary table for the headsets' protected exposure levels at 2000 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	1.24	5.18	0.016
Within Group (S/H)	12	0.239		

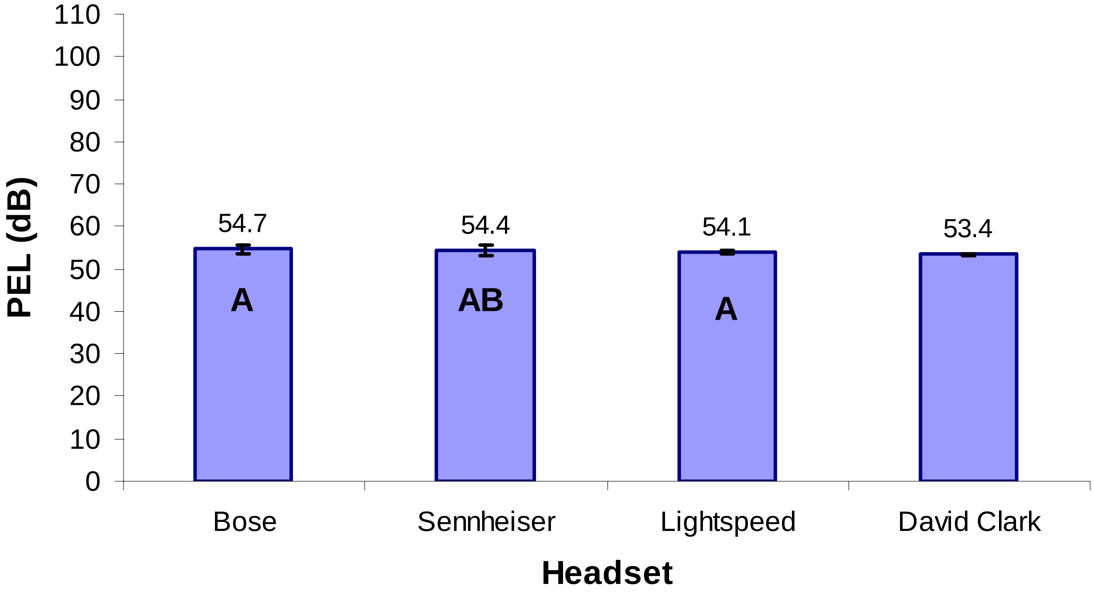


Figure 57. Headset protected exposure levels (PELs) in the aircraft engine noise at 2000 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 27

ANOVA summary table for the headsets' protected exposure levels at 4000 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	1.588	5.05	0.017
Within Group (S/H)	12	0.315		

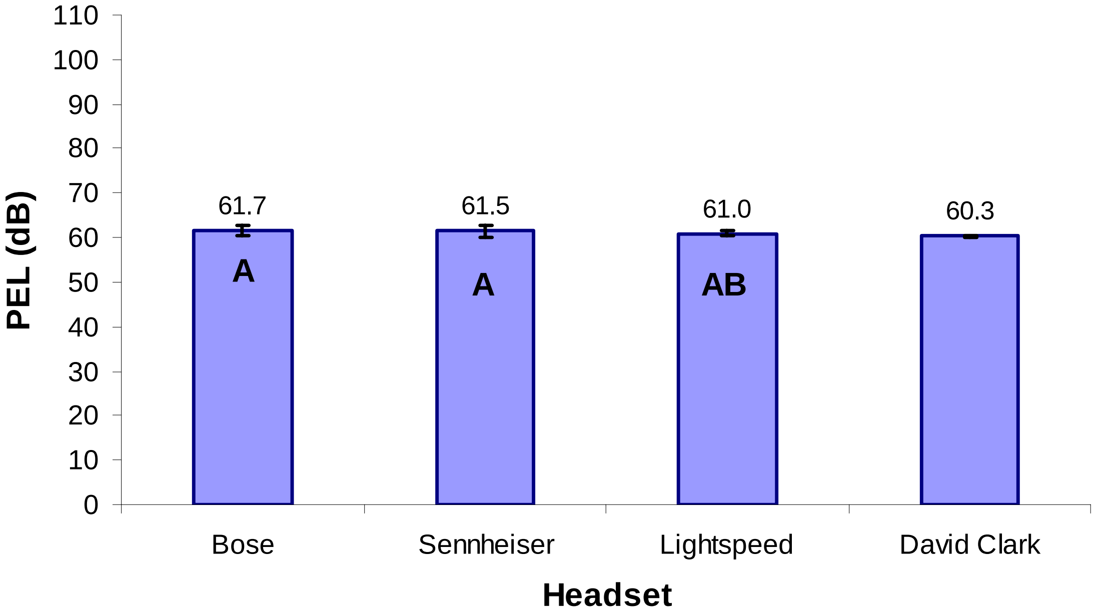


Figure 58. Headset protected exposure levels (PELs) in the aircraft engine noise at 4000 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

PELs at 8000 Hz. A one-way ANOVA was again performed to compare the PEL data for the four headsets resulting in significant differences found between the PEL measurements for the four headsets at 8000 Hz, *F*(3, 12) = 4.96, *p* = 0.018 (Table 28). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise level under the David Clark headset was significantly lower than under the Bose headset, with no significant differences between the David Clark, Sennheiser, and LightSPEED headsets. Furthermore, post hoc comparisons realized no significant differences between the LightSPEED headset, the Sennheiser, and the Bose headsets (Figure 59).

PELs at 12500 Hz. A one-way ANOVA was performed to compare the PEL data for the four headsets. The ANOVA analysis resulted in significant differences found between the PEL measurements for the four headsets at 12500 Hz, *F*(3, 12) = 5.49, *p* = 0.013 (Table 29). Further post-hoc comparisons were carried out using the Tukey HSD procedure. Pairwise comparisons showed that the noise level under the David Clark was significantly lower than the Bose headset. There were no significant differences found between the David Clark headset, the Sennheiser headset, and the LightSPEED headset. Also, no significant differences were found between the LightSPEED headset, the Sennheiser headset, and the Bose headset (Figure 60).

Table 28

ANOVA summary table for the headsets' protected exposure levels at 8000 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	1.509	4.96	0.018
Within Group (S/H)	12	0.304		

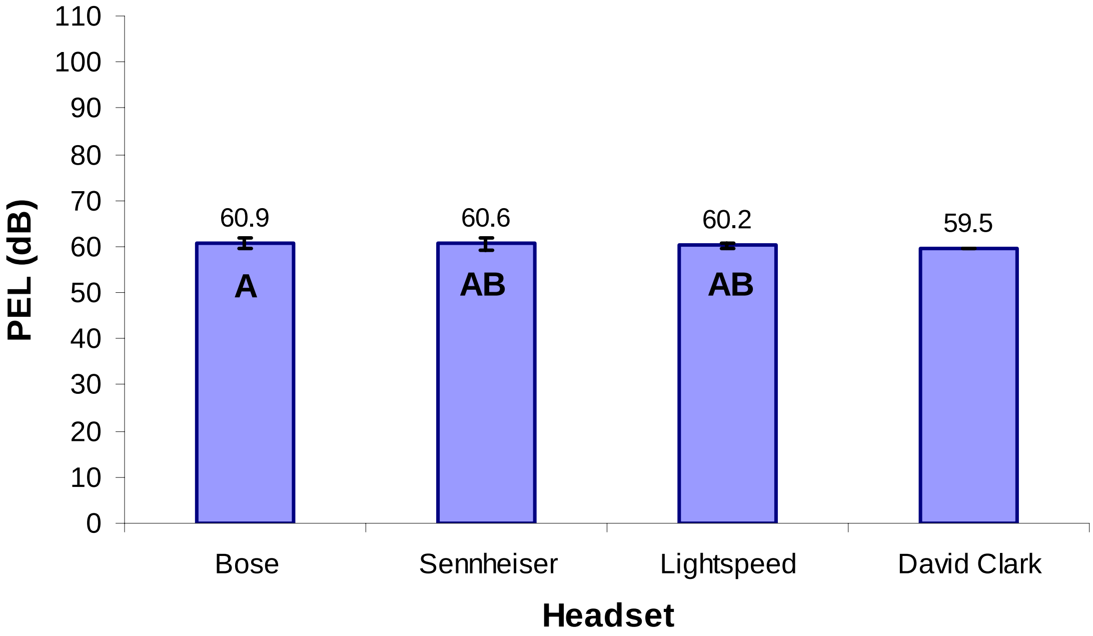


Figure 59. Headset protected exposure levels (PELs) in the aircraft engine noise at 8000 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Table 29

ANOVA summary table for the headsets' protected exposure levels at 12500 Hz.

Source	df	MS	*F*	*p*
Headset (H)	3	1.557	5.49	0.013
Within Group (S/H)	12	0.283		

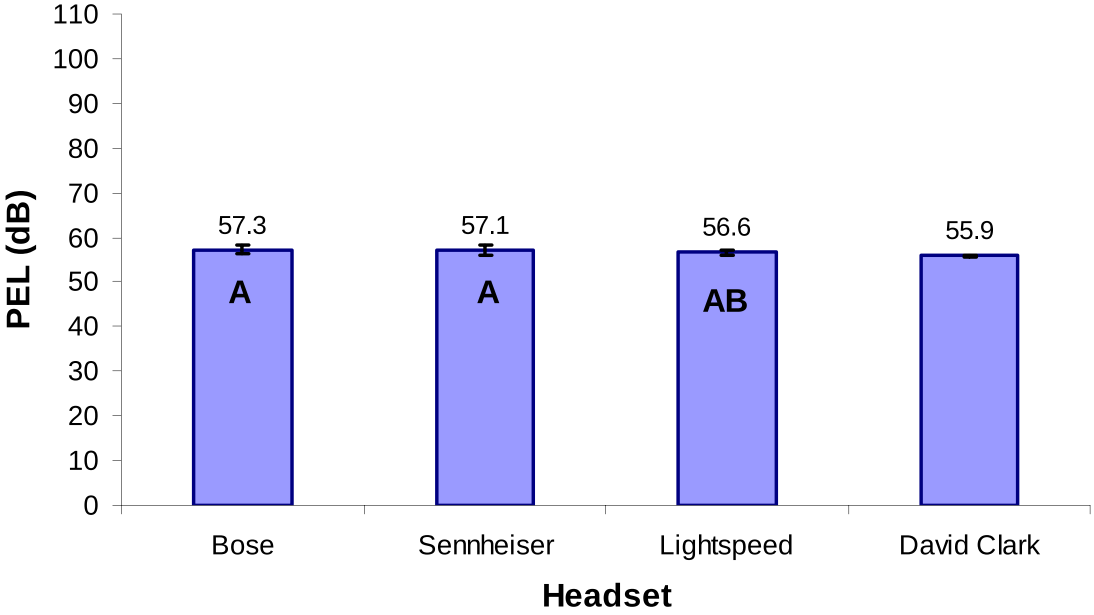


Figure 60. Headset protected exposure levels (PELs) in the aircraft engine noise at 12500 Hz. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

The headset PEL comparisons illustrate some of the well documented characteristic differences between ANR and passive attenuation. For example, PEL measurements for the passive David Clark headset consistently resulted in the lowest mean levels at and above 1000 Hz. However, this performance was often not significantly different from some of the ANR headsets. Likewise, with many of the PEL measurements below 500 Hz, where it is known that ANR outperforms passive attenuation, there were no significant differences found between the ANR and passive attenuation headsets. The ambiguities of these findings are in agreement with some past studies which have measured little difference between the performances of ANR versus passive attenuation (Nixon et al., 1992).

However, there is one very interesting finding at 125 Hz. The peak sound pressure levels in the spectrum, measured at 106-107 dB for the Cessna engine noise, were situated at the frequency range of 80 Hz – 125 Hz. At this point, the PELs of all three ANR headsets are significantly lower (14 – 16 dB lower) than the PEL of the passive David Clark headset. It is quite possible that this frequency range of the most intensive low frequency noise is responsible for upward spread of masking. The ANR headsets are designed to combat this type of high intensity, low frequency noise and therefore may reduce the upward spread of masking. On the other hand, the passive attenuating headset does not have the same attenuation capabilities to prevent the upward spread of masking. This interaction between each headset's attenuation capabilities and high intensity noise at 80 Hz – 125 Hz may be a contributing factor behind the increased speech intelligibility experienced by pilots while wearing an ANR headset in this study, as measured by the previously mentioned speech intelligibility metric.

Headset Protected Exposure Levels with Communications Input

Another (probably more influential) factor over the 125 Hz attenuation, behind the increased speech intelligibility afforded by the ANR headsets, was discovered serendipitously. As discussed within the Equalization of Speech Intelligibility Under Headset section, the headsets were calibrated so that the ATC communications played over each headset was at the same speech intelligibility value (STI=0.80) while in the 95 dBA aircraft engine noise. In calibrating each headset, it was determined that all three ANR headsets attained this speech intelligibility level at approximately 82 to 83 dBA. However, the passive David Clark headset required 93 dBA under the headset to attain a speech intelligibility level of STI = 0.80. In this analysis, there was a single between-subjects independent variable, headset. The four different models specified in the previous section entitled Experimental Design constituted the four levels of this analytical design. Also, for this analysis there was one dependent variable, the at-ear amplitude level.

It was decided to analyze these data using a one-way ANOVA procedure. Before the ANOVA analysis began, analysis to confirm that the data did not violate any assumptions was conducted. A visual inspection of the univariate normality plot of the residuals, outputted by SAS, was conducted. The slope of the data values did indeed follow the slope of the normality values calculated by SAS without deviation. Another visual analysis was conducted on the scatterplot of the predicted versus residual values from the PEL with communications input data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption.

The following ANOVA procedure resulted in significant differences across the four headsets, *F*(3, 8) = 1169.13, $p < 0.0001$ (Table 30). Further post hoc comparisons were conducted using a Tukey HSD method to make all pairwise comparisons. The results showed that the protected exposure level with communications input for the LightSPEED headset was significantly lower than the other three headsets. The comparisons also showed that the PEL levels with communications of the Bose headset and Sennheiser headset were significantly lower than the David Clark headset, with no significant differences found between the Bose headset and the Sennheiser headset (Figure 61).

The finding that the David Clark required a higher signal amplitude to achieve the same speech intelligibility rating was not so astonishing. However, the magnitude of this increase was unexpected, and somewhat alarming. The David Clark headset required an at-ear sound amplitude of 93 dBA to achieve the same speech intelligibility that the ANR headsets accomplished at approximately 10 dB lower. This 93 dBA could be of hazard to a pilot's hearing, depending on the exposure time. The Occupational Safety and Health Administration (OSHA) regulations state that a worker cannot be exposed to sound levels of 93 dBA for greater than approximately 6 hours per 24 hour day. This study's exposure times were only 3.5 hours for each of 4 days, well within the OSHA limitations. The exposure time is dependent upon the type of aviation environment in which the pilot is flying. If the pilot is flying across rural, uncontrolled airspace, then communications are sparse and the pilot will most likely not exceed the exposure limits.

Table 30

ANOVA summary table for the headsets' protected exposure levels with communications input.

Source	df	MS	*F*	*p*
Headset (H)	3	82.813	1169.13	<0.0001
Within Group (S/H)	8	0.071		

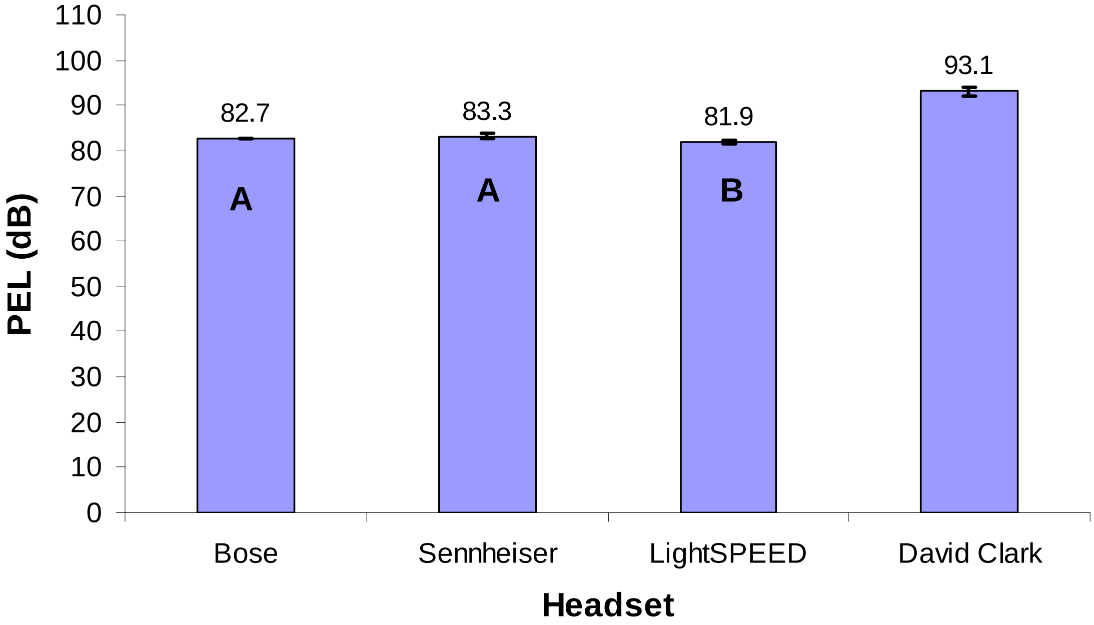


Figure 61. Headset protected exposure levels (PELs) in the aircraft engine noise with ATC communications input included. Different letters represent significant differences at the p < 0.05 level. Vertical range bars represent 95% confidence intervals about the means.

However, if the pilot is flying through congested, controlled airspace, such as the airspace along the east coast, the near-constant communications with air traffic controllers will set the exposure limit at 6 flight hours which could be exceeded in a simple cross-country flight, such a trip from Boston to Washington, D.C.

It should also be noted that the exposure level might be less in actual flight than it was in this experiment. In the engine noise realism ratings, the pilots criticized the realism of the engine noise as being of higher amplitude than would be experienced in a Cessna 172 in actual flight conditions. It is recommended that in-flight measurements be taken of the at-ear exposure levels during communication exchanges between the pilots and the air traffic controllers for long duration flight.

Final Headset Rankings

Following the completion of all four cross-country flight sessions, the pilots were asked to rank the headsets according to overall comfort, noise reduction, and communications performance. The ranking of the four headsets which each pilot gave were a composite score of overall comfort, noise reduction, and communications performance. This ranking data was collected and analyzed using a Fisher's Exact Test. The Fisher's Exact Test compared the categorical variable of headset with the pilots' rankings of those headsets (Table 31). The test resulted in a significant difference in ranking among the four headsets (p = 0.0002). Therefore, further pairwise contrasts were conducted using the Fisher's Exact Test procedure. These contrasts showed that the Bose was ranked significantly higher than the LightSPEED and David Clark headsets.

Table 31

Fisher's Exact Test contingency table comparing headsets via final pilot rankings.

Frequency Percent Row Percent Column Percent	Ranked First	Ranked Second	Ranked Third	Ranked Fourth	Total
Bose	6 18.75 75.00 75.00	2 6.25 25.00 25.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	8 25.00
Sennheiser	2 6.25 25.00 25.00	3 9.38 37.50 37.50	3 9.38 37.50 37.50	0 0.00 0.00 0.00	8 25.00
LightSPEED	0 0.00 0.00 0.00	2 6.25 25.00 25.00	4 12.50 50.00 50.00	2 6.25 25.00 25.00	8 25.00
David Clark	0 0.00 0.00 0.00	1 3.13 12.50 12.50	1 3.13 12.50 12.50	6 18.75 75.00 75.00	8 25.00
Total	8 25.00	8 25.00	8 25.00	8 25.00	32 100.00

However, no significant difference was found between the Bose and Sennheiser headset rankings. The ranking of the Sennheiser headset was significantly higher than that of the David Clark headset, but was not found to be significantly different from the LightSPEED ranking. Lastly, no significant difference in ranking was found between the LightSPEED and David Clark headsets (Figure 62).

It was surprising that the analysis did not find a significant difference between the Bose headset which was rated first by 75% of the pilots, and second by the remaining 25% of the pilots, as compared to the next highest ranked headset, the Sennheiser headset, ranked first by 25% of the pilots, second by 37.5% of the pilots, and third by the remaining 37.5% of the pilots. It seems that this result must be a function of the small sample size. The Fisher's Exact Test would most likely be more sensitive to the trend implicated by these results if a larger sample size could have been used. When recording their rankings of the four headsets, pilots were also asked to record the major factors, or headset characteristics, which led them to their final rankings. The most common comments are listed below:

- Bose – "*Effective ANR,*" "*Clearest communications,*" "*Lightest,*" "*Most Comfortable,*" "*Least fatiguing to wear,*" "*Well-padded*."
- Sennheiser – "*Good passive attenuation,*" "*Comfortable,*" "*Slightly bulkier than Bose,*" "*Good ANR*."
- LightSPEED – "*Unpleasant fit of ear pieces,*" "*Annoying constant buzz,*" "*Big,*" "*Bulky,*" "*Heavy*."
- David Clark – "*Average fit,*" "*Almost unintelligible with poor radio communications,*" "*Loud background noise,*" "*Clamped head*."

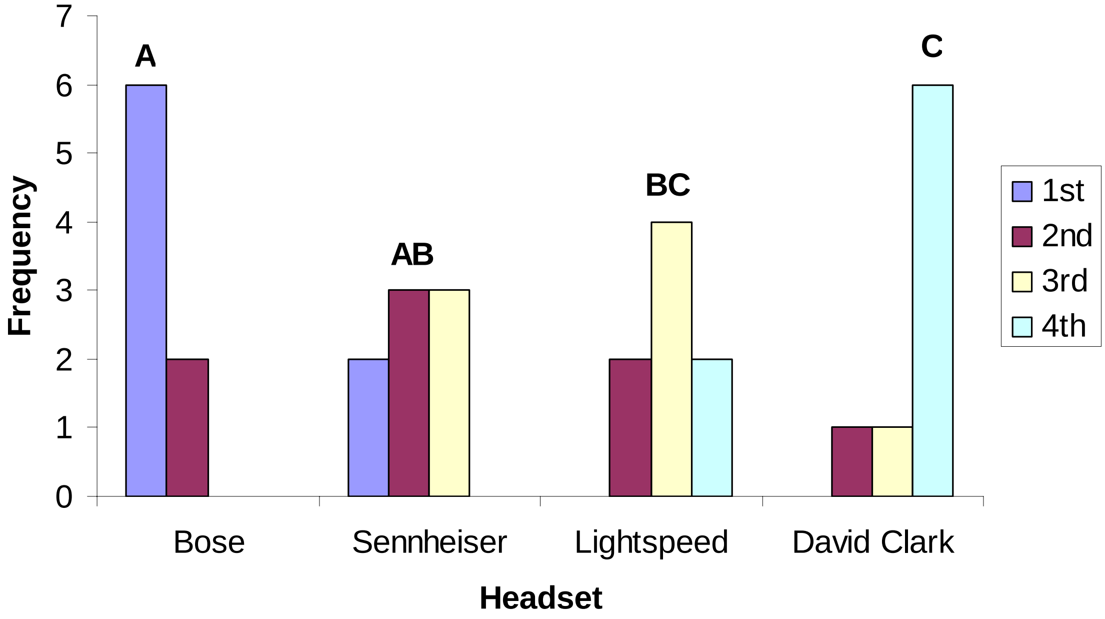


Figure 62. Pilots' final rankings of aviation headsets. Each ranking was a composite measure of the criteria: overall communications quality, comfort, and overall noise reduction. Different letters specify significant differences at the $p < 0.05$ level.

It is very interesting to note that the majority of the most common comments concerning the headsets deal with the comfort the pilots experienced while conducting the 3.5 hour cross-country simulations. This is evidence that comfort does make a contribution to the determination of which headsets are perceived as superior to others. Furthermore, these comments support the earlier conclusion that the non-significant comfort rating scales were more likely an indication that the rating scales were not sensitive enough for the small sample size, rather than an indication that comfort had no influence in the interface between headset and pilot. The area of comfort in regards to aviation communications headsets is definitely an elusive topic which warrants more research attention in hopes to gain an understanding of the factors dominating the pilot's perception of comfort and lead to such opinions as those expressed above.

Primary Task Performance

Primary Task Performance MANOVA. As stated in the Methodology section, data were collected for six measures of flight performance which defined a pilot's primary task performance. The flight performance deviations from ATC commands were calculated by taking the absolute value of the actual performance subtracted from the assigned performance. This was done for magnetic heading, altitude, airspeed, vertical speed, localizer tracking, and the altimeter pressure setting. These six measures were first analyzed using a 4 x 4 x 3 repeated measures MANOVA. The MANOVA indicated significance for the main effects of headset, $F(15, 895) = 6.31$, $p <0 .0001$, and workload, $F(15, 895) = 16.22$, $p < 0.0001$. As Table 32 shows, the main effect for STI values, along with all interactions proved to be non-significant.

Table 32

Wilk's Lambda MANOVA summary table for flight performance measures.

Source	Wilk's Lambda Value	Num df	Den df	*F*	*p*
Headset	0.758	15	895	6.31	<0.0001
STI	0.986	10	648	0.44	0.925
Workload	0.515	15	895	16.22	<0.0001
Headset*STI	0.929	30	1298	0.81	0.756
Headset*Workload	0.908	45	1452	0.70	0.932
STI*Workload	0.924	30	1298	0.86	0.678
Headset*STI*Workload	0.813	90	1576	0.76	0.952

Further analysis was conducted on the headset and workload main effects by separating the six performance measures and conducting individual 4 x 4 x 3 repeated measure ANOVAs. Before the ANOVA analyses began, analysis to confirm that all data did not violate any ANOVA assumptions was conducted. Visual inspections of the univariate normality plots of the residuals, outputted by SAS, were conducted. The slope of the data values did not follow the slope of the normality values calculated by SAS, indicating a violation of the normality assumption for all six performance measures. A logarithmic transformation was applied to each of the six sets of performance data separately, then reanalyzed for normality. The transformation was successful as the slope of the performance data sets then followed the normality values calculated by SAS. Another visual analysis was conducted on the scatterplots of the predicted versus residual values from the comfort rating data showed the tell-tale horizontal band of values, indicating the data did not violate the homogeneity of variance assumption.

Magnetic Heading Performance Deviations ANOVA. A Mauchly's test of sphericity showed that the magnetic heading data did conform to the sphericity assumption, $p = 0.37$. A 4 x 4 x 3 repeated measures ANOVA was conducted on the magnetic heading deviation performance data. This analysis showed significant main effects for headset, $F(3, 21) = 9.85$, $p < 0.0001$, and also for workload, $F(3, 21) = 55.14$, $p < 0.0001$ (Table 33). Main effects for STI value were not significant at the $p < 0.05$ level, nor were any interactions in this analysis. Further analysis of the main effects of headset and workload were carried out using a Tukey HSD method to make all pairwise comparisons.

Table 33

ANOVA summary table for magnetic heading performance deviations.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	0.204		
Within				
Headset (H)	3	9.653	9.85	<0.0001
H x S	21	0.980		
STI	2	0.009	0.03	0.973
STI x S	14	0.005		
Workload (W)	3	8.492	55.14	<0.0001
W x S	21	0.154		
H x STI	6	1.074	1.66	0.131
H x STI x S	42	0.647		
H x W	9	0.344	1.17	0.315
H x W x S	63	0.294		
STI x W	6	0.100	0.30	0.935
STI x W x S	42	0.050		
H x STI x W	18	0.137	0.65	0.855
H x STI x W x S	125	0.211		

Post hoc comparisons of the main effect of headset showed that magnetic heading deviations were significantly lower for pilots while wearing the Bose headset, than when wearing the Sennheiser, LightSPEED, or David Clark headsets. There were not any significant differences found between the Sennheiser, LightSPEED, or David Clark headsets (Figure 63). Post hoc comparisons of the main effect for workload showed that magnetic heading deviations were significantly higher for the high level of psychomotor workload than low psychomotor and both levels of perceptual workload (Figure 64).

Altitude Performance Deviations ANOVA. A Mauchly's test of sphericity showed that the altitude data did conform to the sphericity assumption, $p = 0.24$. A 4 x 4 x 3 repeated measure ANOVA was conducted on the altitude deviation performance data. This analysis showed significant main effects for headset, $F(3, 21) = 12.86$, $p < 0.0001$, and also for workload, $F(3, 21) = 4.44$, $p = 0.005$. Corroborating the prior MANOVA, main effects for STI value were not significant at the $p < 0.05$ level, nor were any interactions in this analysis (Table 34). Further analysis of the effects of headset and workload were carried out using a Tukey HSD method to make all pairwise comparisons. Post hoc comparisons of the main effect of headset showed that altitude deviations were significantly lower for pilots while wearing the Bose headset, than when wearing the Sennheiser, LightSPEED, or David Clark headsets. There were no significant differences found between the Sennheiser, LightSPEED, or David Clark headsets (Figure 65). Post hoc comparisons of the main effect for workload showed that altitude deviations were significantly higher for the high level of psychomotor workload than the low perceptual workload condition.

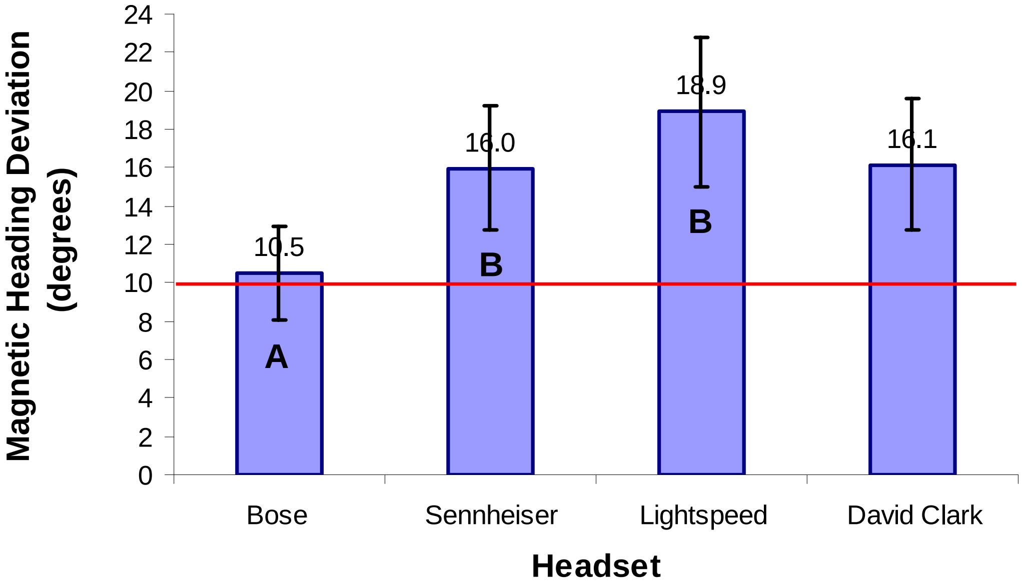


Figure 63. Mean differences in pilots' magnetic heading deviations between aviation headsets. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means. Horizontal red line indicates the maximum tolerance for heading deviation as stated in the FAA Practical Test Standards.

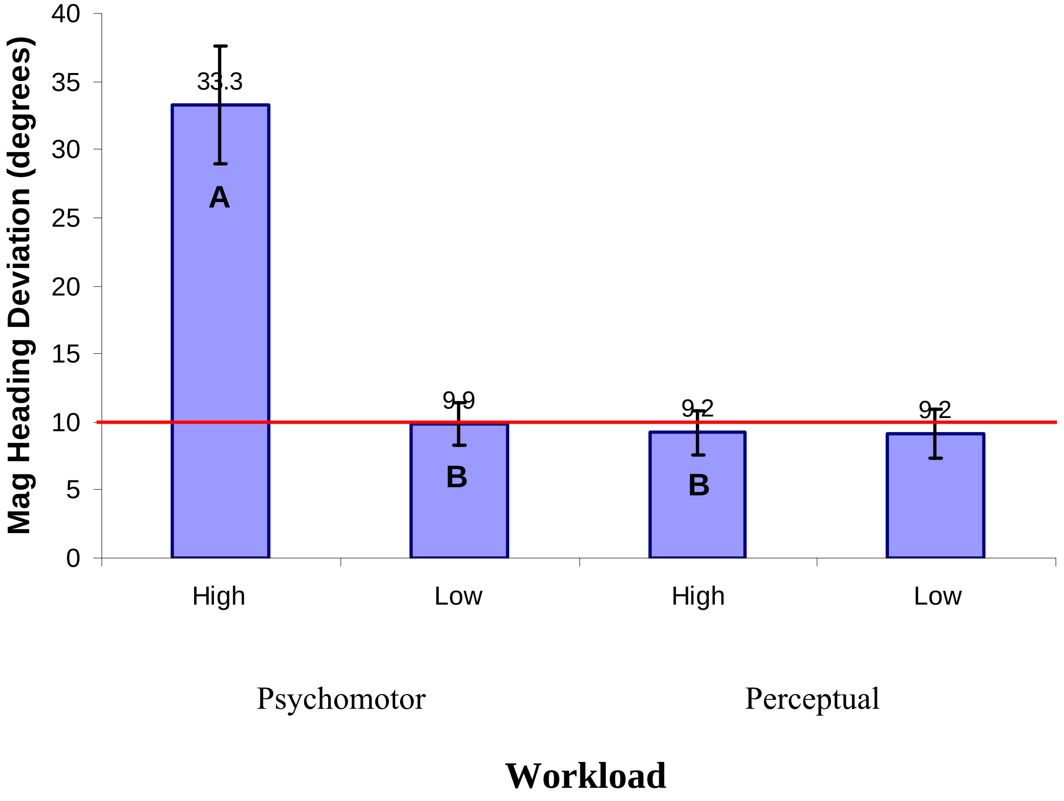


Figure 64. Mean differences in pilots’ magnetic heading between workload conditions. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means. Horizontal red line indicates the maximum tolerance for heading deviation as stated in the FAA Practical Test Standards.

Table 34

ANOVA summary table for altitude performance deviations.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	1.543		
Within				
Headset (H)	3	29.051	12.86	<0.0001
H x S	21	2.259		
STI	2	0.439	0.48	0.812
STI x S	14	0.915		
Workload (W)	3	2.384	4.44	0.005
W x S	21	0.537		
H x STI	6	0.616	0.72	0.635
H x STI x S	42	0.855		
H x W	9	0.173	0.30	0.976
H x W x S	63	0.578		
STI x W	6	0.250	0.72	0.637
STI x W x S	42	0.347		
H x STI x W	18	0.394	0.94	0.531
H x STI x W x S	125	0.419		

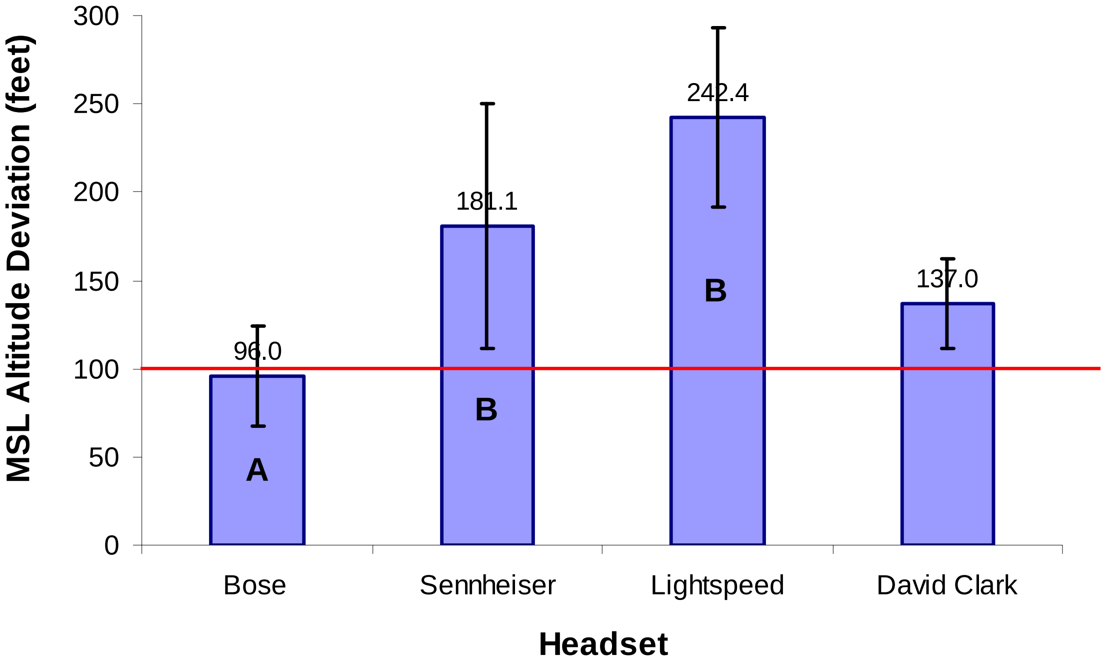


Figure 65. Mean differences in pilots' mean sea level (MSL) altitude deviations between aviation headsets. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means. Horizontal red line indicates the maximum tolerance for altitude deviation as stated in the FAA Practical Test Standards.

There were no significant differences in altitude deviations between the high psychomotor workload, the low psychomotor workload, and the high perceptual workload conditions. There were also no significant differences between the low and high perceptual workload conditions (Figure 66).

Indicated Airspeed Performance Deviations ANOVA. A Mauchly's test of sphericity showed that the indicated airspeed data did conform to the sphericity assumption, $p = 0.19$. A 4 x 4 x 3 repeated measures ANOVA was conducted on the indicated airspeed deviation performance data. This analysis showed significant main effects for headset, $F(3, 21) = 9.78$, $p < 0.0001$, and also for workload, $F(3, 21) = 8.55$, $p < 0.0001$. Corroborating the prior MANOVA, the main effect for STI value was not significant at the $p < 0.05$ level, nor were any interactions in this analysis (Table 35). Further analysis of the effects of headset and workload were carried out using a Tukey HSD method to make all pairwise comparisons. Post hoc comparisons of the main effect of headset showed that indicated airspeed deviations were significantly lower for pilots while wearing the Bose headset, than when wearing the Sennheiser, LightSPEED, or David Clark headsets. There were not any significant differences found between the Sennheiser, LightSPEED, or David Clark headsets (Figure 67). Post hoc comparisons of the main effect for workload showed that indicated airspeed deviations were significantly higher for the high level of psychomotor and high perceptual workload than low psychomotor. There were no significant differences between low psychomotor workload and any other the other workload conditions. Furthermore, there were no significant differences in indicated airspeed deviations between the high psychomotor workload and high perceptual workload (Figure 68).

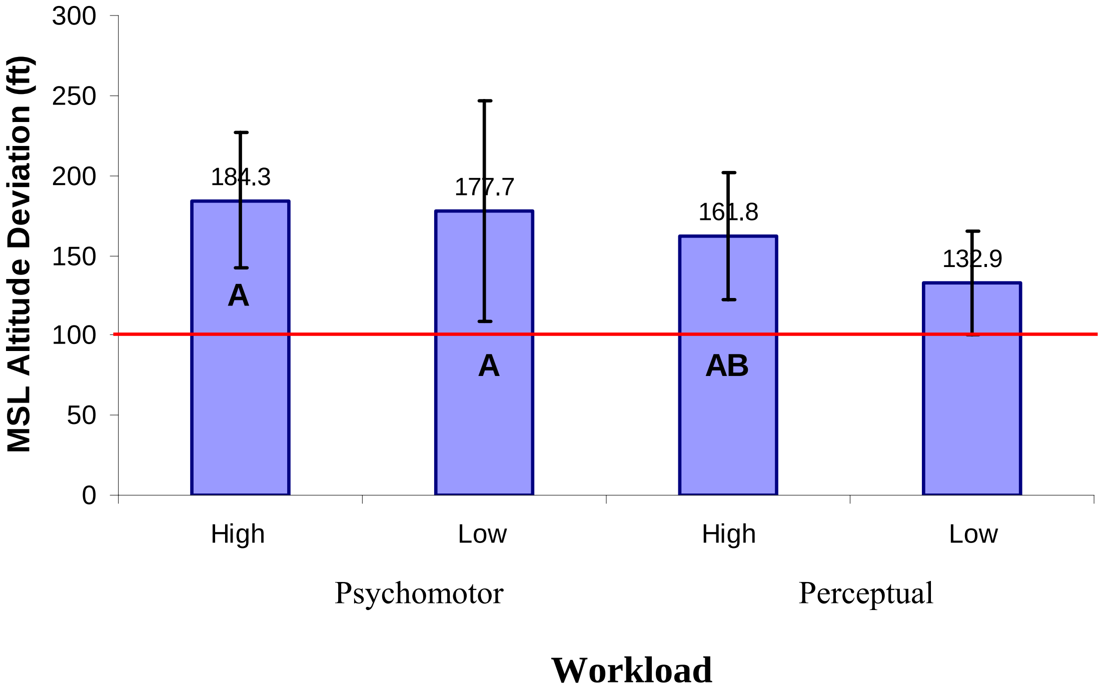


Figure 66. Mean differences in pilot's mean sea level (MSL) altitude deviations between workload conditions. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means. Horizontal red line indicates the maximum tolerance for altitude deviation as stated in the FAA Practical Test Standards.

Table 35

ANOVA summary table for indicated airspeed performance deviations.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	0.785		
Within				
Headset (H)	3	4.538	9.78	<0.0001
H x S	21	0.464		
STI	2	0.019	0.17	0.847
STI x S	14	0.110		
Workload (W)	3	1.573	8.55	<0.0001
W x S	21	0.184		
H x STI	6	0.024	0.36	0.904
H x STI x S	42	0.067		
H x W	9	0.184	1.03	0.414
H x W x S	63	0.179		
STI x W	6	0.132	1.37	0.225
STI x W x S	42	0.096		
H x STI x W	18	0.038	0.49	0.962
H x STI x W x S	125	0.077		

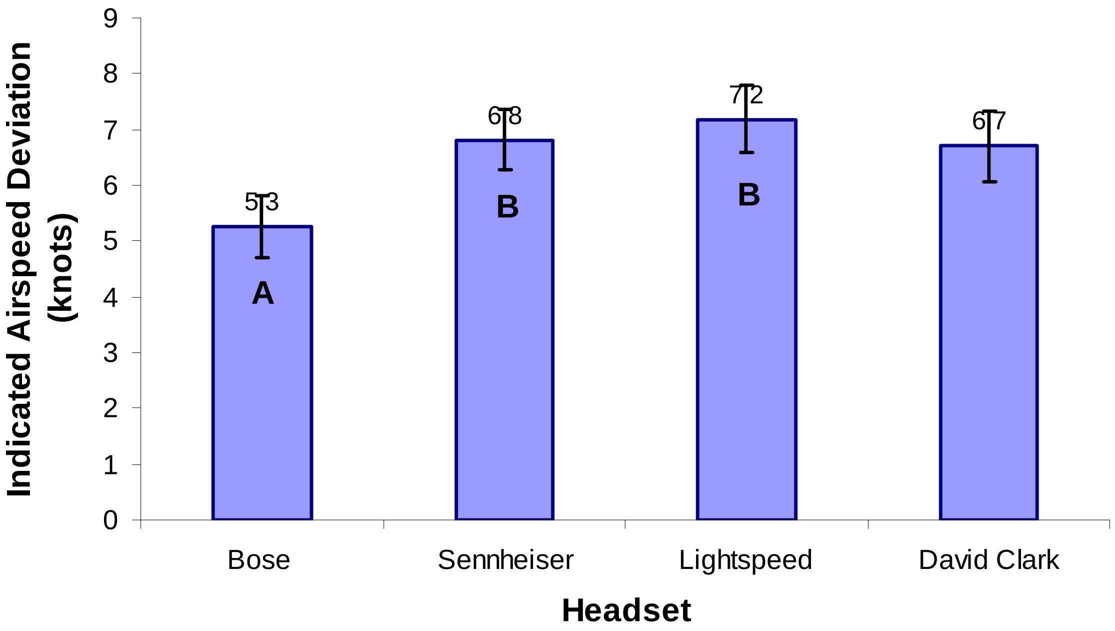


Figure 67. Mean differences in pilots' indicated airspeed deviations between aviation headsets. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means. The maximum tolerance for airspeed deviation, as stated in the FAA Practical Test Standards, is 10 knots.

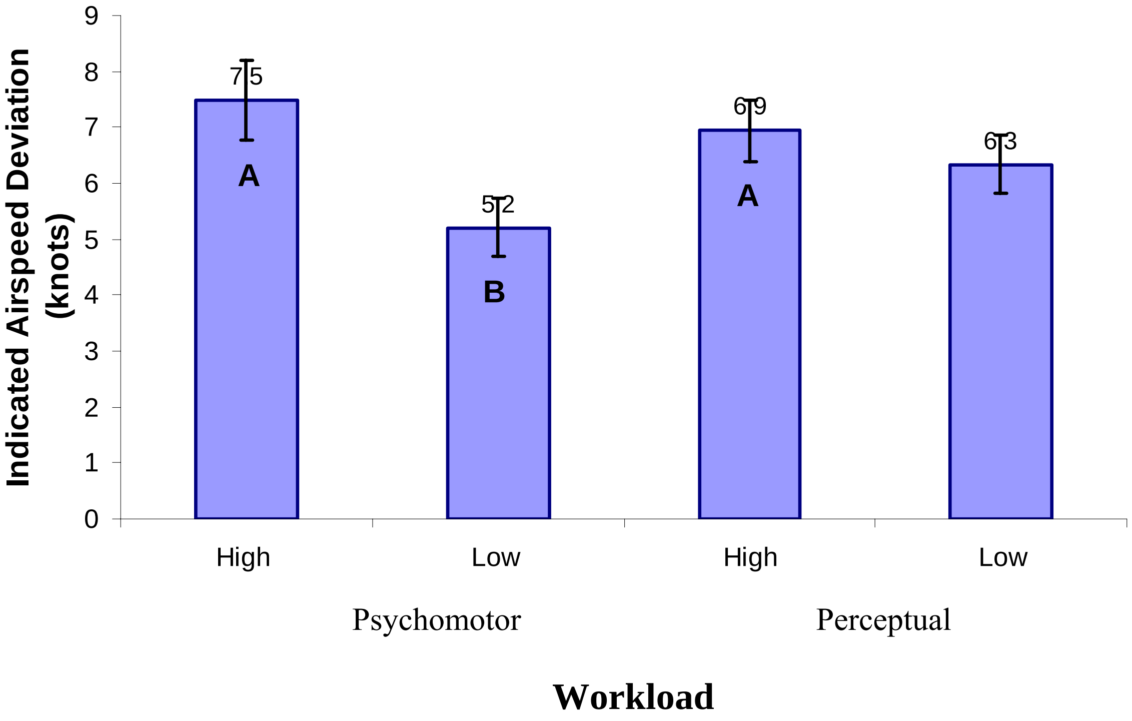


Figure 68. Mean differences in pilots' indicated airspeed deviations between workload conditions. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means. The maximum tolerance for airspeed deviation, as stated in the FAA Practical Test Standards, is 10 knots.

Vertical Speed Performance Deviations ANOVA. A Mauchly's test of sphericity showed that the vertical speed data did conform to the sphericity assumption, $p = 0.35$. A 4 x 4 x 3 repeated measures ANOVA was conducted on the vertical speed deviation performance data. This analysis showed significant main effects for headset, $F(3, 21) = 6.87$, $p = 0.002$, and also for workload, $F(3, 21) = 17.40$, $p < 0.0001$. Corroborating with the prior MANOVA, main effects for STI value were not significant at the $p < 0.05$ level, nor were any interactions in this analysis (Table 36). Further analysis of the effects of headset and workload were carried out using a Tukey HSD method. Post hoc comparisons of the main effect of headset showed that vertical speed deviations were significantly lower for pilots while wearing the Bose headset, than when wearing the Sennheiser, LightSPEED, or David Clark headsets. Additionally, there were not significant differences found between the Sennheiser, LightSPEED, or David Clark headsets (Figure 69). Post hoc comparisons of the main effect for workload showed that vertical speed deviations were significantly highest for the high level of psychomotor. The high perceptual workload condition was significantly lower than high psychomotor workload, but significantly higher than both the low psychomotor and low perceptual workload conditions. There was no significant difference found between low psychomotor workload and low perceptual workload (Figure 70).

Localizer Tracking Performance Deviations ANOVA. A Mauchly's test of sphericity showed that the localizer tracking data did conform to the sphericity assumption, $p = 0.07$. A 4 x 4 x 3 repeated measures ANOVA was conducted on the localizer tracking deviation performance data.

Table 36

ANOVA summary table for vertical speed performance deviations.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	0.918		
Within				
Headset (H)	3	4.046	6.87	0.002
H x S	21	0.589		
STI	2	0.009	0.09	0.918
STI x S	14	0.100		
Workload (W)	3	3.184	17.40	<0.0001
W x S	21	0.183		
H x STI	6	0.014	0.17	0.984
H x STI x S	42	0.085		
H x W	9	0.085	0.63	0.772
H x W x S	63	0.135		
STI x W	6	0.031	0.41	0.871
STI x W x S	42	0.076		
H x STI x W	18	0.048	0.70	0.808
H x STI x W x S	125	0.069		

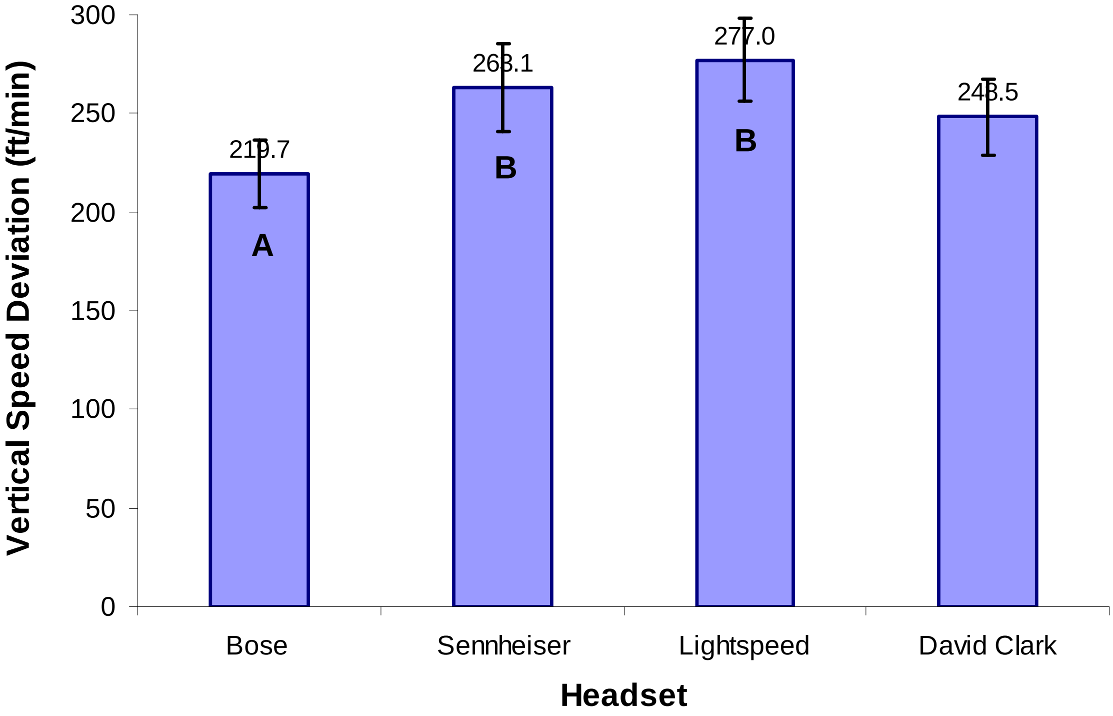


Figure 69. Mean differences in pilots' vertical speed deviations between aviation headsets. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

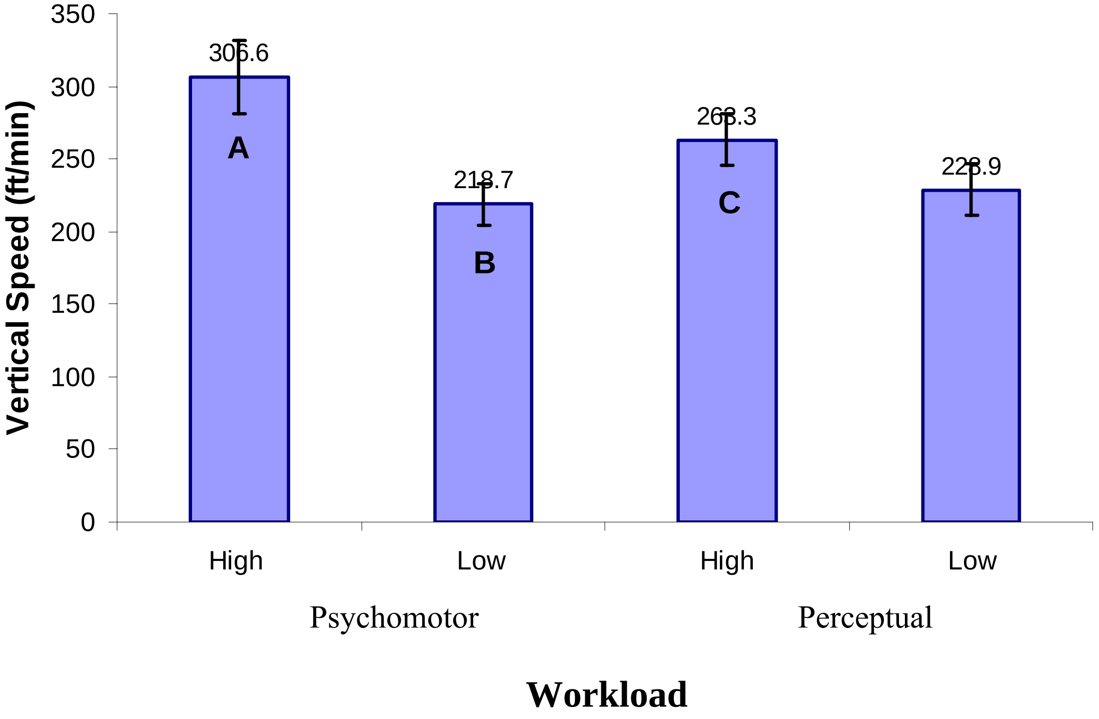


Figure 70. Mean differences in pilots' vertical speed deviations between workload conditions. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Although the prior MANOVA has shown a significant headset main effect, this ANOVA analysis did not find this main effect to be significant in the localizer tracking performance data. Furthermore, this analysis corroborated the prior MANOVA in finding no significance for the STI main effect, and the interaction for headset by STI (Table 37, Figure 71).

During the design of this experiment, there was concern about including the localizer tracking analysis in the MANOVA was it creates an unbalanced analysis. The localizer tracking flight task was limited to one specific flight task, the Instrument Landing System approach, which was found in only one workload condition: high psychomotor workload. Therefore, an analysis of workload and its accompanying interactions could not be conducted for this measure because it only appeared in one level of the workload independent variable.

It was determined acceptable to include the localizer tracking performance measure in the flight performance MANOVA for two main reasons. First, the composition of the task was identical to other tasks included in the MANOVA. The pilot had to use a combination of skills such as navigation, perception of the aircraft's flight attitude through the flight instruments, and manual manipulation of the flight controls to maintain the desired flight attitude to achieve the proper flight performance during the localizer tracking task. Secondly, it was deemed appropriate to include this measure in the MANOVA because of the nature of the MANOVA analysis method. The MANOVA analysis technique is very robust to unbalanced designs and therefore inclusion of this measure (with its limitation of one workload level) would not negatively affect the MANOVA analysis in any way.

Table 37

ANOVA summary table for localizer tracking performance deviations.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	0.969		
Within				
Headset (H)	3	1.132	1.88	0.150
H x S	21	0.602		
STI	2	0.097	1.16	0.326
STI x S	14	0.084		
H x STI	6	0.084	0.93	0.484
H x STI x S	42	0.090		

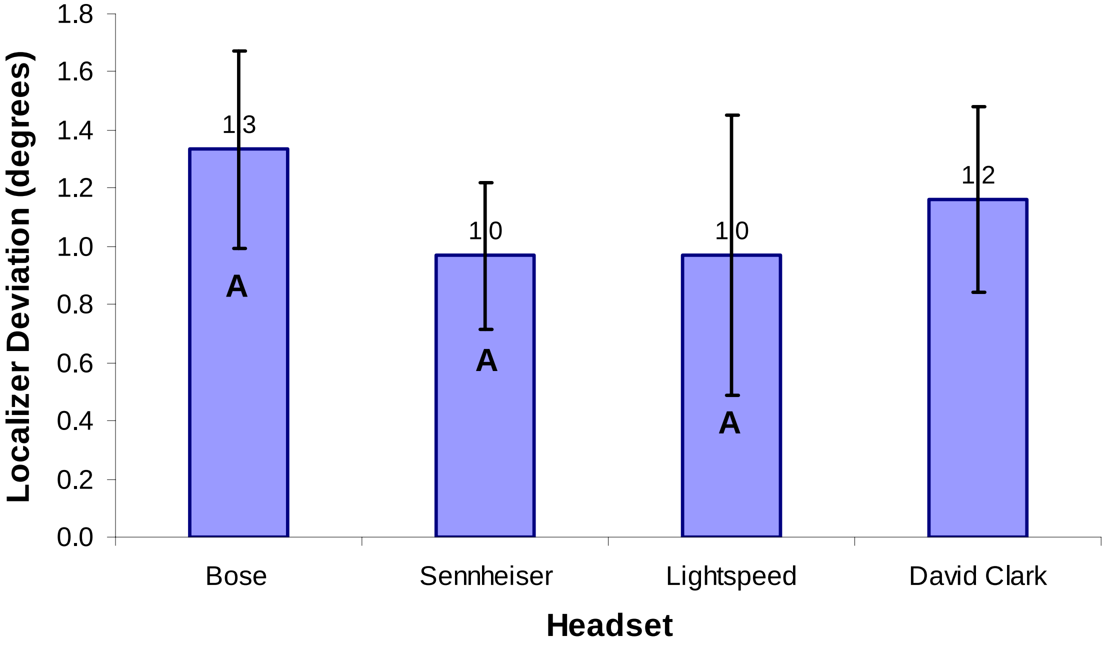


Figure 71. Mean differences in pilots' localizer tracking deviations between aviation headsets. Though means are reported, ANOVA analysis showed no significant main effect for headset at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

The differences between the analysis of the localizer tracking data and the data analyzed for the other 5 flight performance measures is apparent when looking at the localizer tracking ANOVA summary table (Table 37). The localizer tracking ANOVA summary table does not include the main effect for workload, any workload interaction, or associated error terms. Again, this is due to the fact that the localizer tracking task was only found in one level of the workload variable and therefore the workload variable could not be included in the analysis.

Altimeter Pressure Setting Performance Deviation ANOVA. A Mauchly's test of sphericity showed that the altimeter pressure setting data did conform to the sphericity assumption, $p = 0.54$. A 4 x 4x 3 repeated measures ANOVA was conducted on the altimeter pressure setting deviation performance data. This analysis showed significant main effects for headset, $F(3, 21) = 5.13$, $p = 0.002$, and also for workload, $F(3, 21) = 4.90$, $p = 0.002$. Corroborating the prior MANOVA, the main effect for STI value was not significant at the $p < 0.05$ level, nor were any interactions in this analysis (Table 38). Further analysis of the effects of headset and workload were carried out using a Tukey HSD method to make all pairwise comparisons. Post hoc comparisons of the main effect of headset showed that altimeter pressure setting deviations were significantly lower for pilots while wearing the Bose or LightSPEED headsets, than when wearing the Sennheiser, or David Clark headsets. There were not any significant differences found between the Bose and LightSPEED headsets. There were also no significant differences between the Sennheiser and David Clark headsets (Figure 72).

Table 38

ANOVA summary table for altimeter pressure setting performance deviations.

Source	df	MS	*F*	*p*
Between				
Subject (S)	7	9.916		
Within				
Headset (H)	3	38.321	5.13	0.002
H x S	21	7.470		
STI	2	4.037	1.31	0.271
STI x S	14	3.082		
Workload (W)	3	9.359	4.90	0.002
W x S	21	1.191		
H x STI	6	1.094	0.30	0.936
H x STI x S	42	3.648		
H x W	9	1.094	0.52	0.862
H x W x S	63	2.104		
STI x W	6	0.611	0.45	0.846
STI x W x S	42	1.357		
H x STI x W	18	1.160	0.78	0.722
H x STI x W x S	125	1.487		

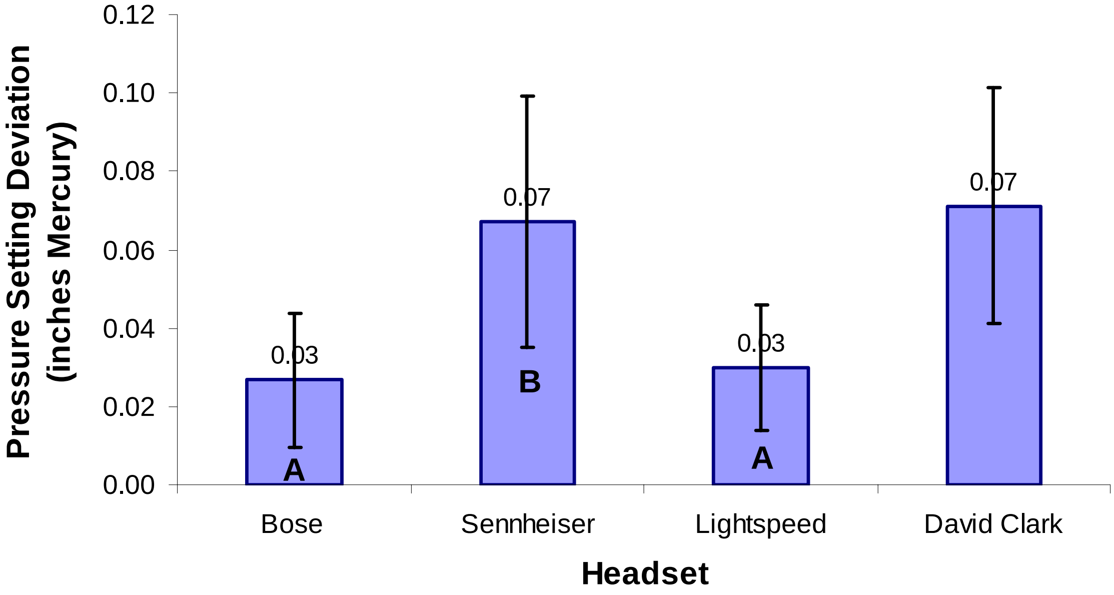


Figure 72. Mean differences in altimeter pressure setting deviations between aviation headsets. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

Post hoc comparisons of the main effect for workload showed that altimeter pressure setting deviations were significantly higher for the low level of perceptual workload than both levels of the psychomotor workload and the high level of perceptual workload. There were no significant differences found between both levels of psychomotor workload and the high level of perceptual workload. Additionally, there were no significant differences in altimeter pressure setting deviations between the high and low levels of psychomotor workload (Figure 73).

One of the main criticisms of primary task performance measures in the workload literature is that it is nearly impossible to control and separate the influences of the many variables on this single measure. As a result, it was necessary to analyze the preceding measures (speech intelligibility, workload, comfort ratings, etc.) and understand those variables separately, then to see the aggregate of those influences through the performance measures, which will be done in Conclusions section.

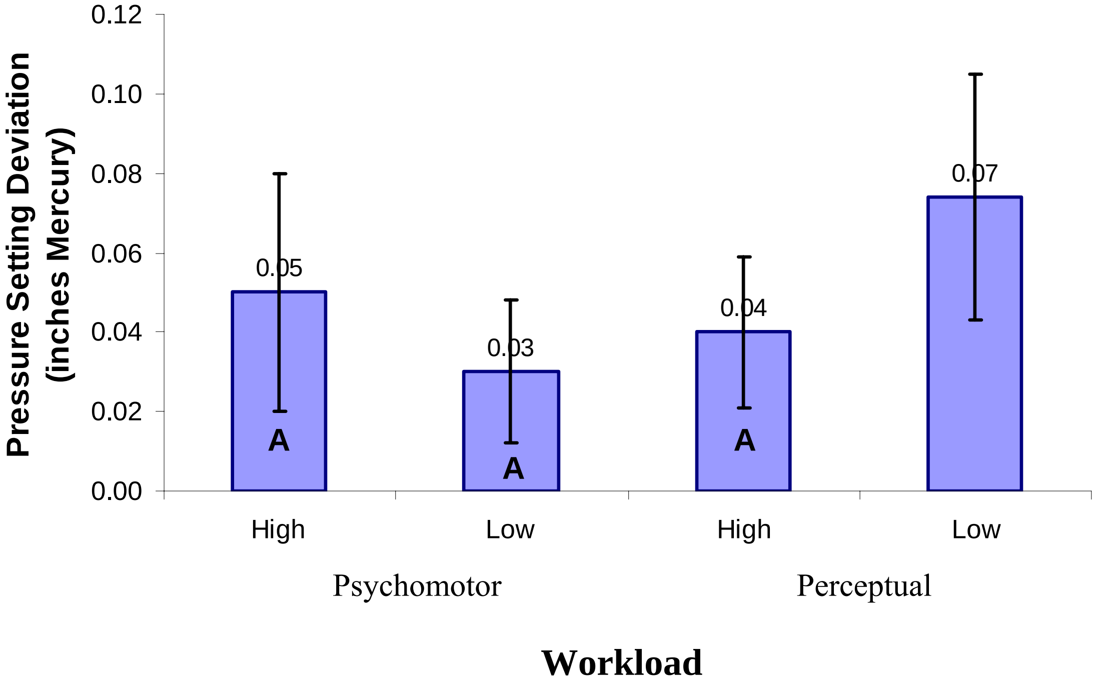


Figure 73. Mean differences in altimeter pressure setting deviations between workload conditions. Different letters represent significant differences at the $p < 0.05$ level. Vertical range bars represent 95% confidence intervals about the means.

CONCLUSIONS

Voice Analysis

The use of the pilot's voice amplitude and formant frequency recorded during the pilot's readbacks of ATC commands did not provide any findings that support a relationship between voice analysis and workload. However, there were several factors which could have influenced the data and led to the non-significant results, two factors especially. The first is the Cessna engine noise that was always present in the background of each voice recording. This noise was of such high intensity that it is entirely plausible that the engine noise masked any subtle differences in an individual pilot's voice amplitude and formant frequency. The second factor which could have influenced the voice data was the difference in microphone characteristics between the AvCom headset used to collect the pilot's baseline voice amplitude and baseline formant frequency and the and the microphone characteristics of the four headsets used in the experimental sessions. These different microphones may have transduced the pilots' voices differently, resulting in differences in the headsets, but not reliably reflecting the different levels of workload the pilots were experiencing, as was probably the case in this study. Research into voice analysis as a measure of workload requires efforts dedicated solely to exploring the potential of a voice-workload relationship before it should be incorporated as measure into an applied study.

Comparison of ANR versus Passive Attenuation

While previous studies investigating ANR and speech intelligibility have focused on the attenuation characteristics of headsets which incorporate active noise reduction when comparing them against passive attenuating headsets (Gower & Casali, 1994; Anderson & Garinther, 1997), this investigation clearly shows that the interface between headset and operator, and the mechanisms that drive this interface are far more complex then sole reliance on attenuation. Additionally, the benefits which can be derived from an ANR headset surpass simply greater protection from environmental noise.

The results of this study support the conclusion reached by Anderson & Garinther (1997) that ANR headsets do increase speech intelligibility as compared to passive attenuation headsets. Furthermore, this investigation shows that, as hypothesized, speech intelligibility affects workload and pilot performance throughout various flight environments and flight tasks. Aviation headsets which incorporate active noise reduction technology facilitate an increase in speech intelligibility and the corresponding benefits of reduced workload and in some cases increased performance.

Speech Intelligibility – Workload Relationship. The results of the speech intelligibility test show that pilots required air traffic control commands to be repeated fewer times when using ANR-based headsets. These same pilots required ATC commands to be repeated a greater number of times when using the David Clark headset, which incorporated only passive attenuation. The conclusion put forth here is that ANR increased the speech intelligibility to the point that pilots required fewer ATC commands to give a 100% correct readback. This same ANR vs. passive relationship is also seen in the workload ratings. Pilots rated the workload of flight modules to be lower while wearing an ANR headset than when wearing the passive headset. These two tests support the theory that a direct inverse relationship exists between speech intelligibility and workload. For example, the data show that an increase in speech intelligibility is directly linked to a decrease in the workload perceived by the pilots. It should be noted that this relationship is not a simple linear relationship. Figure 22 shows that workload is highest at the lowest STI value, and then decreases as the speech intelligibility of the ATC commands increase. It then hits a ceiling effect where the speech intelligibility is sufficient enough to comprehend all aspects of the command and correctly read the ATC command back to the controller. At this point (STI = 0.50), any additional increase in speech intelligibility will not result in a significant additional decrease in workload.

ANR has a direct impact on this speech intelligibility – workload relationship. The data for both the speech intelligibility and workload measures were analyzed for the effect which the different types of headsets may have. In both cases, it was found that the type of headset had a significant impact. Headsets which incorporated active noise reduction as the primary method of noise attenuation accounted for significant increases in speech intelligibility and significant decreases in workload across all experimental conditions. Therefore, the results of this study support the theory that the amount of mental resources needed to understand a communications message is heavily influenced by the speech intelligibility of such a message, and is reflected by the mental workload the pilot perceives. ANR technology increases speech intelligibility freeing mental resources which would have been dedicated to the comprehension of the communications message. These additional mental resources are now available to be reallocated to another task. The freed mental resources, which are now available due to the ANR – induced increase in speech intelligibility, are reflected as a decrease in mental workload in the pilots. It would not be appropriate to surmise that the mental resources freed by ANR headsets are not used by the pilot, which could be misconstrued as an indication that the pilot does not need these resources for other tasks (an argument often given by proponent of replacing pilots with automation). The results of this study seem to show just the contrary. The complex relationship between all variables presented in this study and the results of performance measures show that for certain ANR headsets the freed mental resources must be reallocated to another yet unknown task. However, the results for a headset such as the Bose headset, the mental resources freed from the communications task have been successfully reallocated to the flight control task, as evidenced by the decreased flight deviations. In other words, the pilot's resources are presently taxed by the division of resources between the tasks of aviation, navigation, and communication. A well-designed headset which facilitates a pilot's performance in the communications task frees resources which are reallocated to the aviation task, making for a better performing, safer pilot. This shows researchers and designers that they should concentrate on human-centered design principles to create tools in the cockpit which facilitate pilots in their jobs, not take the entire job away leaving the pilot bored and complacent.

PEL without communications input. Possible contributions to the speech intelligibility-workload relationship may stem from two sources. First, the ANR headsets have significantly more attenuation of the Cessna engine noise at low frequencies (i.e., less than 500 Hz). At some frequencies, the difference between the protected exposure levels of an ANR headset and a passive headset can exceed 10 dB. For example, the PEL without communications input for the passive headset at 125 Hz is 14 dB higher than the closest ANR headset. However, at many other low frequencies this difference shrinks to only a 2 to 5 dB difference. This fluctuation in PEL differences could be the result of the interaction between the intensity of the engine noise and the method of attenuation each headset relies upon. The peak intensity of the Cessna engine noise is situated between the 80 Hz to 125 Hz frequency range. Within this range, the engine noise can reach as high as 106 dBA. The intensity of this noise is of such a level as to allow the sound to more easily pass through the solid material with which the David Clark is constructed and uses as its method of attenuating sound. The ANR-based headsets, on the other hand, do not rely upon materials as their primary method of attenuation, and therefore seem to perform better at attenuating these high intensity, low frequency noises-especially at the peak intensity levels. An ANR-based headset's ability to better attenuate this high intensity, low frequency noise is theorized to play a major role in preventing the upward spread of masking which deteriorates and potentially could completely disrupt speech communications. A more detailed explanation of ANR's influence on the upward spread of masking is forthcoming.

PEL with communications input. In addition to the results of the speech intelligibility metric, the results of the PEL with communications input support the aforementioned theory of ANR's influence on the upward spread of masking. Prior to the experimental sessions, an equalization of speech intelligibility was conducted for the four aviation headsets. The purpose of this procedure was to control for each headset's volume setting, as the subject-set volume setting paradigm was determined to be inadequate for this investigation. However, something quite interesting was noticed during the equalization process. The three ANR devices were equalized to the speech intelligibility value of STI = 0.80. Each ANR headset achieved this speech intelligibility value at an at-ear sound level of approximately 82 to 83 dBA. The passive headset, on the other hand, required an at-ear sound level of 93 dBA to reach the speech intelligibility value of STI = 0.80. This is a large difference to achieve the same speech intelligibility value and one that is present any time a radio transmission comes through the communication system of the headset throughout the flight simulations. With these lower sound levels, the ANR headsets, as previously stated, counteract the upward spread of masking. This is a crucial point to be made, and deserves a reiteration.

The upward spread of masking is the vehicle that allows high intensity, low frequency noise to mask higher frequency signals, such as speech. By attenuating the low frequency noise, ANR prevents the upward spread of masking, thereby increasing the ability of the pilots to perceive and comprehend air traffic control communications. Passive devices, such as the David Clark headset, do not attenuate the low frequencies as well, and therefore allow some upward spread of masking to occur. The effects of the advantages which ANR technology has over the passive attenuation technology are shown in the speech intelligibility measure, PEL with communications input, workload measures, pilot ratings, and performance measures.

The approximate 10 dB difference between the ANR-based headsets and the passive headset, shown by the PEL with communications input, has significant implications for the pilot within the operational environment. Not only does a 10 dB difference indicate that a passive headset requires a doubling in the subjective loudness of the at-ear sound levels for a pilot to attain the same speech intelligibility as with an ANR-based headset, but the OSHA allowable exposure level is reduced considerably. According to the OSHA 5 dB exchange rate, a 10 dB increase requires exposure to be reduced by a factor of four. This means that pilots in light aircraft who use a passive attenuating headset only be allowed exposure to the cockpit noise environment for ¼ the time of those pilots who use an ANR-based headset. This is a drastic reduction in the allowable exposure time for a passive headset.

Further reaching implications of the PEL with communications results focus on the EPA regulation for adding a noise reduction rating (NRR) to hearing protection devices (EPA, 1979). Presently, this regulation only allows passive attenuation HPDs to receive an NRR and even be officially designated as hearing protection devices. Any device which outputs a sound (e.g., ANR reduction sound or communications signal) is not permitted to be tested and given a NRR. The results of this research show that the EPA regulation is incomplete at best. The PEL with communications results show that sound exposure is significantly increased with the addition of communications input. This exposure in the cockpit noise environment is considerably worse while wearing a passive HPD with a NRR, than if the pilot were to wear an ANR-based headset. However, it bears reiteration that the ANR-based headset is not eligible for a NRR or to be designated as an official HPD. The EPA regulation is obviously incomplete and outdated, and must be revised to reflect the two compelling facts that communications input and ANR have significant impact on an operator's sound exposure levels.

Flight Performance Measures. Although the effect that ANR has on the speech intelligibility-workload relationship is fairly straightforward, the relationship between the aviation headset and pilot performance is much more complex. It is in the flight performance measures that the effects of all variables tested converge. Therefore, any interpretation of these performance results would be of little use to interpret the performance measures individually. These measures never exist separately in the actual flight environment, but rather the pilot must be constantly aware of his or her performance in each of the six categories measured to maintain the safety of the flight. However, before any interpretation is attempted yet, the results will be individually compared against the FAA Practical Test Standards.

Comparison of the flight performance results with the FAA Practical Test Standards was believed desirable to ground these results in a practical application. For example, a difference of two degrees in magnetic heading deviation between two headsets could possibly be statistically significant in an analysis. However, in the flight environment this difference would hold no significance at all. Therefore, the FAA Practical Test Standards were chosen to set the thresholds between safe and unsafe flight performance deviations (where they were applicable). It should be noted that these safety standards have not been designed or proven through scientific validation, but rather have evolved over the almost 100 years that the U.S. government has required licensure for pilots. It should also be noted that these are the standards used to judge whether or not a pilot is safe in terms of flight performance during a licensure flight.

The first significant relationship discovered was that of differences found between the individual headsets and the performance deviations made throughout the flight by the pilots. The results show that pilots perform consistently better through all flight conditions while wearing the Bose ANR headset. Not only did pilots perform consistently better while wearing the Bose headset, but they also performed consistently safer. The results of the magnetic heading measurement show that the mean deviation for the Bose headset was 10.5 degrees, while the other three headsets each showed a mean deviation of 16 degrees. Mean deviations from maintaining an assigned cruise altitude show that while wearing the Bose headset, pilot deviation performance averaged 96 feet, whereas performance deviations for the Sennheiser, LightSPEED, and David Clark averaged 180 feet, 170 feet, and 136 feet, respectively. The mean airspeed deviation for the Bose headset was 5 knots, while the mean deviation for the Sennheiser, LightSPEED, and David Clark, were each approximately 7 knots.

As previously stated, the FAA has set standards which are used as the safety and performance standards for awarding pilots an instrument rating. It would then be practically appropriate to compare the mean performance deviations to the allowable performance deviations specified by the FAA. Of course, it should be noted that zero performance deviations would be the safest and most highly desirable result of any pilot's performance. Even so, the FAA allows 10 degrees of deviation in magnetic heading from the assigned magnetic heading during a final examination of a student-instrument pilot. The performance deviation attributed to the Bose headset hung just over this boundary. The other three headsets were well outside of the FAA stated safe performance tolerances. The same can be seen for the altitude deviation results. The mean performance deviation for the Bose headset was within the safe tolerances. However, the mean primary task measures obtained under Sennheiser, LightSPEED, and David Clark headsets were well outside the boundaries of what the FAA considers safe flight.

Airspeed deviations have shown a result somewhat different result from the previous two. While the results did show that the airspeed deviation while wearing the Bose headset was significantly less than the other three headsets, the deviations for all four headsets remained within the FAA safe performance tolerances. As a result, all four headsets displayed safe behavior, while the Bose again displayed the safest behavior.

The following three measures do not have set FAA standards by which to judge the safety of the results; therefore the smallest deviations will be considered the safest. The vertical speed data, which reflects the speed in feet per minute that the aircraft travels at while transitioning between assigned cruising altitudes, showed that pilots deviate least from their assigned vertical speed while wearing the Bose headset or the David Clark. Additionally, the result of the altimeter pressure setting measure showed that pilots set the altimeter pressure more accurately while wearing the Bose or LightSPEED headsets.

In the case of altimeter pressure setting, this measurement is of practical significance to instrument flight as the pilot has no visual reference with the ground to double-check the altimeter reading. Therefore, the smallest deviation in pressure setting is potentially a disastrous error because a deviation in 0.10 in/Hg equates to a significant absolute deviation of 100 feet from the assigned altitude. This deviation is even more dangerous due to its insidious nature. By erroneously setting the pressure window of the altimeter, the pilot's altimeter will show that he or she is at the correct altitude, but the real altitude above ground level will be 100 feet off per 0.10 in/Hg. The deviation from the assigned pressure setting was smallest for the Bose headset (0.03) and the LightSPEED headset (0.03), while the Sennheiser and David Clark deviations were both 0.07. This equates to actual altitude deviations of 30 feet for the Bose and LightSPEED headsets, and a 70 foot deviation for the Sennheiser and David Clark headsets.

Interpretation of the results of the pressure setting data is difficult because the use of the altimeter pressure setting as a performance metric has not been used as a performance metric for the evaluation of communication headsets. Therefore, it was considered to an exploratory technique. This factor leads to the conclusion that the results of the headset and workload main effects analysis of the pressure setting measurement show that it requires more validation before an interpretation can be put forth.

Attention in the interpretation will only be paid to trends which can be seen throughout the majority of the six measures, as these are of greatest importance to advancing safety in general aviation. It is quite obvious that flight performance, as shown by these measures, is superior when pilots are wearing the Bose headset than when they are wearing any of the other headsets in this study. This trend is the comprehensive result of all previously examined variables. For example, previous results have shown pilots perform consistently better in speech intelligibility and workload with all ANR headsets as opposed to the passive headset. However, when all variables are factored in pilots performed significantly better when wearing one headset, Bose. This is due to the fact that the Bose ANR headset performed superior in the speech intelligibility, workload, communications performance, and was rated in the mid-range for the protective exposure levels. The synergistic effect of all these variables is of such quality that 75% of the pilots tested ranked the Bose headset as their number one choice of the four headsets, and the remaining 25% ranked it as second. The rest of the headsets were ranked as a mixture of second, third, and fourth (Sennheiser was ranked first by two pilots). It can be seen that the same collection of factors that caused the pilots to rate the Bose headset high in relation to the other four headsets, also contributed to the increased performance and safety, which pilots achieved while wearing the Bose headset.

The overall trend shows that pilot performance was consistently in the highest group while wearing the Bose headset; in four of the six performance measurements pilots performed significantly better while wearing the Bose headset than while wearing the Sennheiser, LightSPEED, or David Clark headsets. The reason for the pilots to perform better while wearing the Bose headset is due to the fact that the Bose headset ranked highest, or in the highest grouping, in almost every measure incorporated into this investigation, with the only exception of the PELs where the Bose headset was placed consistently in the mid-range during post-hoc comparisons.

Comfort and Communication System Ratings. The same general trend found in the pilot performance measures was also seen in the pilots' ratings of various aspects describing the quality of the headsets' communication systems. Analysis of the comfort ratings and communication system operational performance ratings showed that the communication system probably plays a more influential role in pilot performance.

The non-significant comfort ratings are attributed to a lack of sensitivity to such factors as headband clamping force, which was shown to be significantly different by the headband clamping force measurement. Furthermore, it is surmised that a small sample size and the pilots' mental models of comfort could have been factors in the non-significant comfort rating results. Therefore, the study seemed to lack the power, most likely due to a combination of the three previously stated factors, to realize any significant differences in comfort across the four headsets, which were consistently described by the pilots in their final rankings.

However, most ratings regarding performance of the communication system did show significant differences indicating that the effect sizes for the operational performance ratings are potentially larger than those of the comfort ratings. It is surmised that the larger effect sizes may be a tell-tale indication for the larger influence on pilot performance, and could be used in comparison with comfort ratings. This relationship will have to be tested in the future to make a final decision on the causality between communication system quality, comfort, and pilot performance.

For now, the operational performance ratings in this study show the Bose headset to be consistently rated as a headset with one of the highest, or the highest, quality communication systems. These findings are further supported by the anecdotal comments made by the pilots that communications with the Sennheiser headset sounded "staticy" and "scratchy," the LightSPEED headset had a "constant and annoying background hum," and the David Clark headset had a "loud background noise" and "clamped (the pilot's) head." When the advantages of ANR attenuation, mid-range or better PELs, and a high quality communications system are combined, this investigation shows that a properly-designed headset could provide the pilot with a tool that facilitates their performance.

Due to the results on all measures throughout this investigation, it has been shown that the data support a theoretical relationship between speech intelligibility and workload, upon which active noise reduction has a significant positive impact. However, when this relationship is applied to operator performance, many more factors influence their performance than the three already mentioned. Factors such as the quality of the communication system play a very influential role in determining whether the headset will be a tool to facilitate pilot performance, hinder pilot performance, or have a negligible influence.

Limitations of this Study. Several limitations have been identified within this study. Internally, some of the metrics, such as the voice analysis and the altimeter pressure setting performance measure, were considered exploratory and will require further study until they can be regarded as validated workload and performance measures.

Externally, limitations constrain the generalization of the previously stated results and conclusions. Due to the limitations of the iGATE simulator, only IFR pilots and IFR flight conditions were used in this study. Therefore, the conclusions of this study can only be generalized to IFR flight conditions. Furthermore, the use of the Cessna 172 cockpit noise environment allows generalization of the conclusions to be extended to piston engine aircraft. The conclusions can, however, be extended from single engine aircraft to multi-engine aircraft based on the research conducted by Tobias (1968). Tobias (1968) showed that multi-engine piston driven aircraft exhibit similar cockpit noise environments as their single engine counterparts and therefore the results of this study can also be applied to this class of aircraft. The results of this study cannot, and should not, be applied to any type of turbine engine aircraft, even the turboprop class of aircraft.

RECOMMENDATIONS and FUTURE RESEARCH

There were some obvious limitations in the study which will be reviewed in future replications or smaller studies which investigate specific relationships and questions identified in this experiment. First, this study identified the relative effects of comfort and communications operational performance. Operational performance played a more influential role in the effect individual headsets had on pilot performance. Although, the ratings identified some significant aspects of the communications system, the aviation communications headset system components which affect pilot performance most were not conclusively determined. More research is necessary to identify the system components which drive the operational performance of the communications system and which most positively impact pilot performance. The research findings should then be incorporated into future headset development.

This study did not find any significant differences between headsets for any of the comfort ratings. However, when pilots were asked why they ranked headsets in the order they did in the final ranking, the answers were invariably a mixture of comfort and operational performance that they liked or disliked. Therefore, an effect of comfort may be a contributing factor, although a smaller factor relative to the significant communication system factors, but still one which requires investigation with a larger sample containing the power to tease out the differences between headsets in terms of comfort and the pilot's final rankings.

As stated in the 'Participants' section, IFR-rated pilots were only used in this experiment because of the limitations of the iGATE flight simulator. The iGATE flight simulator is certified by the FAA for instrument training, however, the available visual reference for visual flight (VFR) is so small (12” x 3”) that a valid visual flight simulation was not possible. It would be very interesting to run this same experimental protocol with less experienced pilots on a simulator capable of realistically simulating VFR (e.g. adding the 120 degree field of view accessory to the iGATE simulator). This experiment was run on a simulator rated only for instrument flight (IFR), and therefore more experienced instrument-rated pilots were used. However, the findings of ANR positively impacting speech intelligibility and workload may be even more influential, and beneficial, to *less experienced* pilots (e.g. student pilots). During flight training, student pilots are exposed to an environment unlike any other they have experienced previously. The fast-paced, dynamic nature of this environment, combined with the difficulty of controlling the aircraft, increases workload on the student pilot. This overload condition makes for a difficult environment for human learning and training. ANR technology may reduce workload on the student pilot allowing them to retain more information and skills, potentially accelerating the training process.

The last recommendation regards the voice analysis as a measure of workload. The results in this investigation did not show that voice analysis as an indicator of workload level was viable within the single engine light aircraft environment. However, once the confounding variables of piston engine noise and the varying types of headsets are removed, voice analysis may prove to be useful in measuring pilot workload. To conclusively determine whether voice analysis is, or is not, a viable measure of operator workload, an investigation simulating the lower background noise of a commercial airliner, combined with the same headset incorporated across all treatments could possibly impose the level of control needed to effectively evaluate voice analysis as a measure of workload. Another method to evaluate voice analysis might be the use of archived air traffic control recording which were taken during an aircraft accident. The analysis of the pilot's speech amplitude and formant frequency may show simultaneously whether voice analysis effectively evaluates workload and stress in this case, but also whether it could be used in exactly the process for which it is intended, to evaluate in-situation recordings of pilot voiced speech.

REFERENCES

Acton, W.I. (1970). Speech intelligibility in a background noise and noise-induced hearing loss. *Ergonomics, 13(5),* 546-554.

Acton, W.I., Lee, G.L., & Smith, D.J. (1976). Effects of headband forces and pressure on comfort of earmuffs. *Annals of Occupational Hygiene, 19,* 357-361.

Air-safety.net. (1996). Accident description – 27 March, 1977. Retrieved August 2, 2004 from http://aviation-safety.net/database/1977/770327-0.htm.

Albuquerque, E.S.J., da Gama, A. D., & Macedo, M.V. (1991). Carotid angiodynographic studies in individuals occupationally exposed to noise and vibration. *Aviation, Space, and Environmental Medicine, 62,* 134-140.

Anderson, B.W., and Garinther, G.R. (1997). Effects of active noise reduction in armor crew headsets. In *AGARD conference proceedings* (pp. 20-1 – 20-6). Neuilly-Sur-Seine, France: AGARD.

ANSI S3.5 –1969. (1969). *American national standard methods for calculation of the articulation index (AI).* American National Standards Institute, New York, New York.

ANSI S3.5 –1997. (1997). *American national standard methods for calculation of the speech intelligibility index (SII).* American National Standards Institute. New York, New York.

ANSI S3.19 –1974. (1974). *Method for the measurement of real-ear protection of hearing protectors and physical attenuation of earmuffs.* American National Standards Institute. New York, New York.

Backs, R.W. (1998). A comparison of factor analytic methods of obtaining cardiovascular autonomic components for the assessment of mental workload. *Ergonomics, 41,* 733-745.

Backs, R.W., Lenneman, J.K., & Sicard, J.L. (1999). The use of autonomic components to improve cardiovascular assessment of mental workload in flight simulation. *International Journal of Aviation Psychology, 9,* 33-47.

Bartholomae, R.C., & Stein, R.R. (1990). Active noise cancellation – performance in a hearing protector under ideal and degraded conditions. In *Proceedings of Noise-Con 90* (pp. 151-155). Austin: University of Texas.

Berger, E.H. (2000). Hearing protection devices. In E.H. Berger, L.H. Royster, J.D. Royster, D.P. Driscoll, & M. Layne (Eds.), *The Noise Manual* (pp. 379-454). Virginia: American Industrial Hygiene Association.

Berger, E.H., & Mitchell, I. (1989). Measurement of the pressure exerted by earmuffs and its relationship to perceived comfort. *Applied Acoustics, 27(2),* 79-88.

Blix, A.S., Stromme, S.B., & Ursin, H. (1974). Additional heart rate – an indicator of psychological activation. *Aerospace Medicine, 45,* 1219-1222.

Boersma, P., & Weenink, D. (n.d.) *Praat: doing phonetics by computer.* Retrieved on January 23, 2005 from http://www.praat.org.

Bronkhorst, A.W., & Plomp, R. (1989). Binaural speech intelligibility in noise for hearing-impaired listeners. *Journal of the Acoustical Society of America, 86,* 1374-1383.

Brown, D.L., Vitenese, H.S., Wetzel, P.A., & Anderson, G.M. (2002). Instrument scan strategies of F-117 pilots. *Aviation, Space, and Environmental Medicine, 73,* 1007-1013.

Brungart, D.S. (2001). Evaluation of speech intelligibility with coordinate response measure. *Journal of the Acoustical Society of America, 109*, 2276-2279.

Casali, J.G. (1989). Multiple factors effect speech communication in the workplace. *Occupational Safety and Health, 58,* 32-42.

Casali, J.G. (1992). Technology advances in hearing protection: Active noise reduction, frequency/amplitude-sensitivity, and uniform attenuation. In *Proceedings of the Human Factors Society Annual Meeting* (pp. 258-262). Santa Monica, CA: Human Factors and Ergonomics Society.

Casali, J.G., & Grenell, J.F. (1990). Noise-attenuating earmuff comfort: A brief review and investigation of band-force, cushion, and wearing-time effects. *Applied Acoustics, 29,* 117-138.

Casali, J.G., & Horylev, M.H. (1987). Speech discrimination in noise: The influence of hearing protection. In *Proceedings of the Human Factors Society Annual Meeting* (pp. 1246-1250). Santa Monica, CA: Human Factors and Ergonomics Society.

Casali, J.G., Lam, S.T., Epps, B.W. (1987). Rating and ranking methods for hearing protector wearability. *Sound and Vibration, 21(12),* 10-18.

Casali, J.G., & Robinson, G.S. (1994). Narrow-band digital active noise reduction in a siren-cancelling headset: Real-ear and acoustical manikin insertion loss. *Noise Control Engineering Journal, 42(3),* 101-115.

Casali, J.G., & Wierwille, W.W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load, *Human Factors, 25,* 623-641.

Casali, J.G., & Wierwille, W.W. (1984). On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics, 27,* 1033-1050.

Cargart, R. (1946). Monitored live voice as a test of auditory acuity. *Journal of the Acoustical Society of America, 17,* 339-349.

Connell, L.J., and Reynard, W.D. (1993). Emergency medical service helicopter incidents reported to the aviation safety reporting system. *International Symposium on Aviation Psychology*. Columbus, OH.

Damongeot, A., Tisserand, M., Krawsky, G., Grosdemange, J.P., & Lievin, D. (1981). Evaluation of the comfort of personal hearing protectors. In P.W. Alberti (Ed.), *Personal Hearing Protection in Industry* (pp. 151 –161). New York: Raven Press.

Dau, T., Puschel, D., & Kohlrausch, A. (1996). A quantitative model of the effective signal processing in the auditory system I. Model structure. J*ournal of the Acoustical Society of America, 99*, 3615-3622.

DeJoy, D.M. (1984). The nonauditory effects of noise: review and perspectives for research. *Journal of Auditory Research, 24,* 123-150.

Di Nisi, J. Muzet, A., Ehrhart, J., & Libert, J.P. (1990). Comparison of cardiovascular response to noise during waking and sleeping in humans. *Sleep, 13*, 108-120.

Di Nisi, J., Muzet, A., & Weber, L.D. (1987). Cardiovascular responses to noise: effects of self-estimated sensitivity to noise, sex, and time of day. *Journal of Sound and Vibration, 114*, 271-279.

Dirks, D.D., Morgan, D.E., and Dubno, J.R. (1982). A procedure for quantifying the effects of noise on speech recognition. *Journal of Acoustical Society of America, 76,* 87-96.

Drullman, R., Festen, J., & Plomp, R. (1994). Effect of envelope smearing on speech perception. *Journal of the Acoustical Society of America, 95*, 1053-1064.

Dubno, J.R., Dirks, D.D., and Morgan, D.E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *Journal of Speech and Hearing Disorders, 47,* 114-123.

Elhilali, M., Chi, T., & Shamma, S.A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication, 41,* 331-348.

Eschenbrenner, A. J., Jr. (1971). Effects of intermittent noise on the performance of a complex psychomotor task. *Human Factors, 13,* 59-63.

Eggemeier, F.T., & Wilson, G.F. (1991). Subjective and performance-based assessment of workload in multi-task environments. In D.L. Damos (Ed.) *Multiple Task Performance.* London: Taylor & Francis.

Endsley, M.R. (1995). Measurment of situational awareness in dynamic systems. *Human Factors, 37,* 65-84.

Environmental Protection Agency (EPA). (1979). Noise labeling requirements for hearing protectors. Federal Register. 44(190), 40CFR Part 211, 56130 56147.

Eves, F.F., & Gruzelier, J.H. (1984). Individual differences in cardiac response to high intensity auditory stimulation. *Psychophysiology, 21,* 342-352.

Federal Aviation Administration (FAA). (1979). *General aviation and air taxi activity survey / calendar year 19--.* Washington, D.C.: Department of Transportation.

Flugarth, J.M., & Wolfe, B.N. (1971). The effectiveness of selected earmuff-type hearing protectors. *Sound and Vibration, 5(5),* 25-27.

French, N. R. and Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19,* 90-119.

Froelich, G. (1969). The effects of ear defenders on speech perception in military transport aircraft. In *Proceedings of AGARD conference on aeromedical aspects of radio communication and flight safety*. Neuilly-Sur-Seine, France: AGARD.

Gasaway, D.C. (1986). Noise levels in cockpits of aircraft during normal cruise and considerations of auditory risk. *Aviation, Space and Environmental Medicine, 57,* 103-112.

Gelfand, S.A., Ross, L., and Miller, S. (1988). Sentences reception in noise from one versus two sources: Effects of aging and hearing loss. J*ournal of the Acoustical Society of America, 83,* 248-256.

Griefahn, B. and DiNisi, J. (1992). Mood and cardiovascular functions during noise, related to sensitivity, type of noise, and sound pressure level. *Journal of Sound and Vibration, 155*, 111-123.

Goldstein, E.B. (1989). *Sensation and Perception.* Belmont, CA: Wadsworth.

Gomes, L.P., Pimenta, A.F., Branco, N.A. (1999). Effects of occupational exposure to low frequency noise on cognition. *Aviation, Space, and Environmental Medicine, 70,* A115-A118.

Gower, D.W., & Casali, J.G. (1994). Speech intelligibility and protective effectiveness of slected active noise reduction and conventional communications headsets. *Human Factors, 36(2),* 350-367.

Green, D.M. (1993). A maximum likelihood method for estimating thresholds in a yes-no task. *Journal of the Acoustical Society of America, 87*, 2096-2105.

Hack, J., Robinson, H., and Lathrop, R. (1965). Auditory distraction and compensatory tracking. *Perceptual and Motor Skills, 20,* 228-230.

Hancock, P.A., & Caird, J.K. (1993). Experimental evaluation of a model of mental workload. *Human Factors, 35,* 413-429.

Hart, S.G. (1975). Time estimation as a secondary task to measure workload. In *Proceedings of the 11th Annual Conference on Manual Control* (pp.64-77). Washington, DC: US GOP.

Hagerman, D. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology, 11,* 79-87.

Hagerman, D. (1984). Clinical measurements of speech reception thresholds in noise. *Scnadinavian Audiology, 13,* 57-63.

Hankins, T.C., & Wilson, G.F. (1998). A comparison of heart rate, eye activity, EEG, and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine, 69,* 360-367.

Hansen, C.H. (2001). *Understanding Active Noise Cancellation.* New York: Spon Press.

Henderson, D., and Hamernik, R. (1995). Biologic bases of noise-induced hearing loss. *Occupational Medicine: State of the Art Reviews, 10*, 513-534.

Hormann, H., Lazarus-Mainka, G., Schubeius,M., & Lazarus, H. (1984). The effect of noise and the wearing of ear protectors on verbal communication. *Noise Control Engineering Journal, 23*, 69-77.

House, A.S., Williams, C.W., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation testing methods: Consonant differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37,* 158-166.

Howell, K., & Martin, A.M. (1975). An investigation of the effects of hearing protectors on vocal communication in noise. *Journal of Sound and Vibration, 41(2),* 181-196.

Humphrey, D., & Krmaer, A. (1994). Towards a psychophysiological assessment of dynamic changes in mental workload. *Human Factors, 36,* 3-26.

International Standards Organization (ISO). (1996). Hearing protectors – safety requirements and testing – ear muffs (ISO/DIS 10449). Geneva, Switzerland.

Ivergard, T.B.K., & Nicholl, A.G. (1976). User tests of ear defenders. *American Industrial Hygiene Journal,* 139-142.

Jorna, P.G. (1992). Spectral analysis of heart rate and psychological state: a review of its validity as a workload index. *Biological Psychology, 34,* 237-257.

Keppel, G. (1973). *Design and Analysis: A Researcher's Handbook.* New Jersey: Prentice-Hall.

Kjellberg, A., Andersson, P., Skoldstrom, B., & Lindberg, L. (1996). Fatigue effects of noise on aeroplane mechanics. *Work & Stress, 10*, 62–71.

Kling, J.W., & Riggs, L.A. (Eds.). (1972). *Woddworth & Schlosberg's Experimental Psychology.* NY: Holt, Rinehart & Winston.

Kramer, A.F., Sirevaag, E.J., & Braune, R. (1987). A psychophysiological assessment of operator workload during simulated flight missions. *Human Factors, 29,* 145-160.

Kryter, K.D. (1962). Validation of the articulation index. *Journal of the Acoustical Society of America, 34,* 1698-1702.

Kryter, K.D. (1974). *The Effects of Noise on Man*. Orlando: Academic Press.

Kryter, K.D., & Poza, F. (1980). Effects of noise on some autonomic system activities. *Journal of the Acoustical Society of America, 67,* 2036-2044.

Lancaster, J.A., Robinson, G.A., & Casali, J.G. (2004). Comparison of two voice synthesis systems as to speech intelligibility in aircraft cockpit engine noise. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (p. 188-122), New Orleans, Louisiana: HFES.

Lancaster, J.A., & Casali, J.G. (2005). Investigating pilot performance using mixed-modality simulated data link. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (p. 188-122), Orlando, Florida: HFES.

Lee, E.A., & Messerschmidt, D.G. (1994) *Digital communication* (2nd ed.). Boston: Kluwer Academic Publishers.

Lerner, E.J. (1983). The automated cockpit. *IEEE Spectrum, 20,* 57-62.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Levitt, H. (1978). Adaptive testing in audiology. *Scandinavian Journal of Audiology, Suppl. 6,* 241-291.

Levitt, H., & Rabiner, L.R. (1967). Use of sequential strategy in intelligibility testing. *Journal of the Acoustical Society of America, 39,* 609-612.

Lhuede, E.P. (1980). Earmuff acceptance among sawmill workers. *Ergonomics, 23,* 1162-1172.

Li, F.F., & Cox, T.J. (2003). Speech transmission index from running speech: A neural network approach. *Journal of the Acoustical Society of America, 113*, 1999-2008.

Lindeman, H.E. (1976). Speech intelligibility and the use of hearing protectors. *Audiology, 15,* 348-356.

Luczak, H., & Laurig, W. (1973). An analysis of heart rate variability. *Ergonomics, 16,* 85-97.

Macmillian, N.A., & Creelman, C.D. (1991). *Detection Theory: A User's Guide.* Cambridge University Press.

Meeker, W.F. (1957). *Active Ear Defender Systems: Component Considerations and Theory* (WADC Technical Report 57-368 Part I). Dayton, Ohio: Wright-Patterson Air Force Base, Aero Medical Laboratory.

Michon, J.A. (1966). Tapping regularity as a measure of perceptual motor load. *Ergonomics, 9,* 67-72.

Miller, G.A., Heise, G.A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*, 329-335.

Moore, T. (1981). Voice communication jamming research. In *AGARD conference proceedings* (pp. 2:1-2:6). Neuilly-Sur-Seine, France: AGARD.

Moroney, W.F., Biers, D.W, & Eggemeier, F.T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *International Journal of Aviation Psychology, 5(1),* 87-106.

Mulder, L.J.M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology, 34,* 205-236.

Nilsson, M., Soli, S.D., & Sullivan, J.A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America, 95*, 1085-1099.

Nixon, C.W., McKinley, R.L., Steuver, J.W., & McCavitt, A.R. (1992). What can active noise reduction headsets do for you? *1992 Hearing Conservation Conference* (pp. 107-110). Lexington, KY: National Hearing Conservation Association.

Nobili, R., Mammano, F., & Ashmore, J. (1998). How well do we understand the cochlea? *Trends in Neurosciences, 21*, 159-167.

Newby, H.A. (1979). *Audiology*. Englewood Cliffs, NJ: Prentice-Hall.

O' Donnell, R.D., & Eggemeier, F.T. (1986). Workload assessment methodology. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and performance.* New York: Wiley.

Ott, R.L., & Longnecker, M. (2001). *An Introduction to Statistical Methods and Data Analysis*. Pacific Grove, CA: Duxbury.

Pais, F.P., Araujo, A., & Ribeiro, C.S. (1996). Echocardiographic evaluation in patients with vibroacoustic syndrome. *Aviation, Space, and Environmental Medicine, 67*, 668.

Park, M.Y., & Casali, J.G. (1991). An empirical study of comfort afforded by various hearing protection devices: Laboratory versus field results. *Applied Acoustics, 34,* 151-179.

Peterson, E.A., Augenstein, J.S., Hazelton, C.L., Hetrick, D. & Levene, R.M. (1984). Some cardiovascular effects of noise. *Journal of Auditory Research, 24,* 35 – 62.

Pickett, J.M. (1959). Low frequency noise and methods for calculating speech intelligibility. *Journal of the Acoustical Society of America, 13*, 1259-1263.

Plomp, R., & Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18,* 43-52.

Plutchnik, R. (1959). The effects of high intensity intermittent sound on performance, feeling, and physiology. *Psychological Bulletin, 56*, 133-151.

Plutchnik, R. (1961). The effects of high intensity intermittent sound on compensatory tracking and mirror tracing. *Perceptual and Motor Skills, 12*, 187-194.

Rink, T.L. (1979). Hearing protection and speech discrimination in hearing-impaired persons. *Sound and Vibration, 13(1),* 22-25.

Robertson, R.M., & Williams, C.E. (1975). Effect of noise exposure during primary flight training on conventional and high-frequency hearing of students pilots. *Aviation, Space, and Environmental Medicine, 46,* 717-724.

Robinson, G.S., & Casali, J.G. (1995). *Empirical determination of insertion loss for the NCT PA-3000 active noise reduction headset* (Tech Report #9511 Audio Lab 9/29/95-5-HP). Blacksburg, Virginia: Virginia Polytechnic Institute and State University, Department of Industrial and Systems Engineering, Auditory Systems Laboratory.

Robinson, G.S., & Casali, J.G. (2000). Speech communication and signal detection in noise. In E.H.Berger, L.H. Royster, D.P. Royster, D.P. Driscoll, and M. Layne (Eds.), *The noise manual* (pp. 567-600). Fairfax, VA: American Industrial Hygiene Association.

Royster, L.H., & Holder, S.R. (1981). Personal hearing protection: problems associated with the hearing protection phase of the hearing conservation program. In P.W. Alberti (Ed.), *Personal Hearing Protection in Industry* (pp. 151 –161). New York: Raven Press.

Rubio, S., Diaz, E., Martin, J., & Puente, J.M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA TLX, and workload profile methods. *Applied Psychology: An International Review, 53,* 61-86.

Sammer, G. (1998). Heart period variability and respiratory changes associated with physical and mental load: non-linear analysis. *Ergonomics, 41,* 746-755.

Sanders, M.S., & McCormick, E.J. (1993). *Human Factors in Engineering and Design.* New York: McGraw-Hill.

Savich, M. (1981). Practical problems of hearing protector use in Canadian mines. In P.W. Alberti (Ed.), *Personal Hearing Protection in Industry* (pp. 151 –161). New York: Raven Press.

Scott, T.H. (1962). Varied sensory environment. *Psychological Bulletin, 59*, 257-272.

Stark, J.M., Scerbo, M.W., Freeman, F.G., & Mikulka, P.J. (2000). Mental fatigue and workload: Effort allocation during multiple task performance. In *Proceedings of the IEA 2000 / HFES 2000 Congress* (pp. 3-863 – 3-866). Washington, D.C.: International Ergonomics Association.

Stevens, S. S., Miller, J., and Truscott, I. (1946). The masking of speech by sine waves, square waves, and regular and modulated pulses. *Journal of the Acoustical Society of America, 18*, 418-424.

Steneeken, H.J.M., & Houtgast, T. (1971). Evaluation of speech transmission channels by using artificial signals. *Acustica, 25,* 355-367.

Steeneken, H.J.M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America, 67*, 318-326.

Steeneken, H.J.M., & Houtgast, T. (2002). Validation of the revised STIr method. *Speech Communication, 38,* 413-425.

Steeneken, H.J.M., & Verhave, J.A. (1996). Personal active noise reduction with integrated speech communication devices: development and assessment. In *AGARD conference proceedings* (pp. 18-1 – 18-8). Neuilly-Sur-Seine, France: AGARD.

Strother, J.B. (1999). Communication failures lead to airline disasters. In *IEEE Proceedings of Communication Jazz: Improvising the New International Communication Culture* (pp. 29-34). Melbourne, FL: IEEE.

Suter, A.H. (1992). Communication and job performance in noise: A review. *ASHA Monograph, 28.*

Tobias, J.V. (1968a). *Cockpit noise intensity: fifteen single-engine light aircraft.* (FAA Office of Aviation Medicine Reports No. 68-21). Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Tobias, J.V. (1968b). *Cockpit noise intensity: eleven twin-engine light aircraft.* (FAA Office of Aviation Medicine Reports No. 68-25). Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Tobias, J.V. (1970). Auditory processing for speech intelligibility improvement. *Aerospace Medicine, 41*, 728-733.

Tobias, J.V. (1972a). *Binaural processing of speech in light aircraft.* (FAA Office of Aviation Medicine Reports No. 72-31). Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Tobias, J.V. (1972b). *Auditory effects of noise on air-crew personnel.* (FAA Office of Aviation Medicine Reports No. 72-32). Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Tokhi, M. O. and Leitch, R. R. (1992). *Active noise control.* New York, NY: Oxford Science Publications.

Townsend, T.H. (1978). Speech intelligibility through communication headsets for general aviation. *Aviation, Space, and Environmental Medicine, 49,* 466-469.

Townsend, T.H., & Olsen, C.C. (1979). Effects of phase manipulation on speech intelligibility through communication headsets. *Aviation, Space, and Environmental Medicine, 50*, 355-356.

Urquhart, R.L. (2002). *The effects of noise on speech intelligibility and complex cognitive performance.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Van Wijngaarden, S.J., and Rots, G. (2001). Balancing speech intelligibility versus sound exposure in selection of personal hearing protection equipment for Chinook aircrews. *Aviation, Space, and Environmental Medicine, 72*, 1037-1044.

Veltman, J.A., & Gaillard, A.W.K. (1996). Measurement of pilot workload with subjective and physiological techniques. *Biological Psychology, 42,* 323-342.

Vitense, H.S., Jacko, J.A., & Emery, V.K. (2003). Multimodal feedback: An assessment of performance and mental workload, *Ergonomics, 46,* 68-87.

Ward, W.D. (1986). Anatomy and physiology of the ear: normal and damaged hearing. In E.H. Berger, W.D. Ward, J.C. Morrill, & L.H. Royster (Eds.), *Noise and hearing conservation manual* (pp.177-196). Akron, OH: American Industrial Hygiene Association.

Ward, W.D., Royster, L.H., & Royster, J.D. (2000a). Anatomy and physiology of the ear: normal and damaged hearing. In E.H. Berger, L.H. Royster, J.D. Royster, D.P. Driscoll, & M. Layne (Eds.), *The noise manual* (pp. 101-122). Fairfax, VA: American Industrial Hygiene Association.

Ward, W.D., Royster, L.H., & Royster, J.D. (2000b). Auditory and nonauditory effects of noise. In E.H. Berger, L.H. Royster, J.D. Royster, D.P. Driscoll, & M. Layne (Eds.), *The noise manual* (pp. 123-147). Fairfax, VA: American Industrial Hygiene Association.

Wagstaff, A.S., Tvete, O. & Ludvigsen, B. (1996). The effect of a headset leakage on speech intelligibility in helicopter noise. *Aviation, Space, and Environmental Medicine, 67,* 1034-1038.

Wagstaff, A.S., Tvete, O., & Ludvigsen, B. (1999). Speech intelligibility in aircraft noise as a function of altitude. *Aviation, Space, and Environmental Medicine, 70*, 1064-1069.

Wagstaff, A.S., & Woxen, O.J. (2001). Double hearing protection and speech intelligibility – Room for improvement. *Aviation, Space, and Environmental Medicine, 72,* 400-404.

Wheeler, P.D., & Halliday, S.G. (1981). An active noise reduction system for aircrew helmets. In *AGARD conference proceedings* (pp. 22-1 – 22-8). Neuilly-Sur-Seine, France: AGARD.

Whitaker, L., Peters, L. and Garinther, G. (1989). Tank crew performance: Effects of speech intelligibility on target acquisition and subjective workload assessment. In *Proceedings of the Human Factors Society Annual Meeting* (pp. 1411-1413). Santa Monica, CA: Human Factors Society.

Whitaker, L. A (1991). Performance as a function of communication in military vehicle simulators. In *Proceedings of the Human Factors Society Annual Meeting* (pp. 614-617). Santa Monica, CA: Human Factors Society.

Whitaker, L. A. and Peters, L. J. (1993). Communication between crews: The effects of speech intelligibility on team performance. In *Proceedings of the Human Factors Society Annual Meeting* (pp. 630-634). Santa Monica, CA: Human Factors Society.

Wickens, C.D. (1991). Processing resources and attention. In D.L. Damos (Ed.), *Multiple task performance*. London: Taylor & Francis.

Wickens, C.D., & Hollands, J.G. (2000). *Engineering Psychology and Human Performance.* New Jersey: Prentice Hall.

Wierwille, W.W., & Connor, S.A. (1983). Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. *Human Factors, 25,* 1-16.

Wierwille, W.W., & Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35,* 263-281.

Wierwille, W.W., Rahimi, M., & Casali, J.G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors, 27,* 489-502.

Zera, J. (2004). Speech intelligibility measured by adaptive maximum-likelihood procedure. *Speech Communication, 42,* 313-328.

Appendix A

Informed Consent Form

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Informed Consent for Participants in Research Projects Involving Human Subjects

Title of Project: The Effect of Active Noise Reduction on Speech Intelligibility and Pilot Performance in an Instrument Flight Simulation

Prinicpal Investigator: R. Brian Valimont, M.S.

Faculty Advisor: Dr. John G. Casali, Professor, ISE

I. THE PURPOSE OF THIS RESEARCH

The purpose of this study is to determine the performance and benefits of an active noise reduction aviation communications headset in the noise environment of a cockpit of a Cessna 172.

II. PROCEDURES

The procedures used in this research are as follows. If you wish to become a participant after reading the description of the study, then sign this form. If you have any questions about the study or this form, please feel free to ask them at any time.

The study consists of five sessions. For the first session, you will be screened to determine if you qualify for the experiment. Screening will consist of a review of your pilot's license with instrument-rating, your medical certificate, and your logbook. This will be done to verify that you are presently instrument current according to the Federal Aviation Regulation (FARs). Then a hearing test will be administered. To begin with, you will be asked several questions to assess the general health and condition of your ears. Then you will be given an examination in which the experimenter will look into your ears using an otoscope. Next, your right and left hearing will be tested with very quiet tones played through a set of headphones. You will have to be very attentive and listen carefully for these tones. Depress the button on the hand-held switch and hold it down whenever you hear the pulsed-tones and release it when you do not hear the tones. The tones will be very faint and you will have to listen carefully to hear them. No loud or harmful sounds will be presented over the headphones.

If you qualify and choose to participate in the study, you will be scheduled for four sessions of flight simulations. You will be allowed, and are encouraged, to bring any flight gear you choose for a cross-country flight (e.g. knee or lap board, flight computer, etc.), except your headset. For each cross-country flight simulation, you will be given a different headset to wear throughout the session. You will be given a pre-planned flight plan and all necessary information (flight log, sectional, etc.) to successfully complete the flight. You will be allowed 20 minutes to study the flight route and ask any questions you having pertaining to the flight. Then, you will be situated in the simulator room where the simulator will be set at the starting point of the cross-country flight. When you are ready, you may radio to begin the flight simulation. While you are flying you are to follow the pre-planned route, obey all ATC communications, and try to keep your aircraft within 10 degrees of the assigned heading, within 10 knots of the assigned airspeed, and 100 feet of the assigned altitude.

When the flight simulation is over you will asked some simple questions concerning the headset you wore throughout the flight simulation. Please answer those questions as honestly and accurately as you possibly can. At the end of the entire experiment you will be asked to rank each headset according to your personal preferences. Again, please be as honest and as accurate as you can. Each experimental session will take approximately 3.5 hours.

III. RISKS

During the hearing test, you will be in a soundproof booth with the experimenter sitting outside. The door to the booth will be shut but not locked; you may open it from the inside or the experimenter may open it from the outside. There is also an intercom system through which you may communicate with the experimenter by simply talking (there are no buttons to push). If you are or think you may be claustrophobic or if you are uncomfortable in the confined spaces, please tell the experimenter at this time. He/she will show you the rooms and let you enter them to see if they make you uncomfortable. The flight simulations will be conducted in another room and the experimenter will be in an adjacent room.

The Occupational Safety and Health Administration (OSHA) currently allows workers in the United States to be exposed to 90 dBA time-weighted average noise for 8 hours/day. Sound measurements have been conducted using an artificial head (ANSI 3.19-1974) within the aircraft noise (95 dBA) to be utilized in the experiment to determine the sound pressure levels (SPL) under the headsets. The results indicate a SPL of 84 - 94 dBA under the headset. Speech Transmission Index values (STI, a measure of how comprehendible speech is) will affect your performance during ATC radio communications. Different STI values will be utilized to see how intelligibility is affected. At no time will the exposure levels under each headset Occupational Safety and Health Administration's (OSHA) 4-hour exposure. Additional the duration of the flight simulations will extend, at most, to 3.5 hours, well below OSHA's 4 hour exposure limit for 95 dBA noise levels.

Given the short exposure times during ATC communications, it is felt that there is little or no potential for doing any harm to your hearing. (Stimulus levels presented during the experiment will be checked and adjusted before every experimental session.)

III. BENEFITS OF THIS PROJECT

Your participation in this experiment will provide information that will be used to determine the relative strengths and weaknesses of active noise reduction (ANR) technology in aviation communication headsets. The results of this study will help to determine if ANR will improve speech intelligibility relieving pilots of some of their workload, and indirectly improving flight performance.

IV. EXTENT OF ANONYMITY AND CONFIDENTIALITY

The results of your participation will be kept strictly confidential. At no time will the researchers release the results of your participation to anyone other than the individuals working on the project without your written consent. Your written consent is required for the researcher to release any data identified with you as an individual to

anyone other than personnel working on the project. The information you provide will have your name removed and only a subject number will identify you during analyses and any written reports of the research. All subject numbers will be secure and stored on the principal investigator's personal computer.

VI. COMPENSATION

You will be paid $20.00 per hour for your participation in the experiment. Payment will be made immediately after you have finished your participation.

VII. FREEDOM TO WITHDRAW

You are free to withdraw from this study at any time without penalty. If you choose to withdraw, you will be compensated for the portion of time you have spent in the study. There may also be certain circumstances under which the investigator may determine that you should not continue as a participant of this project. These include, but are not limited to, unforeseen health-related difficulties, inability to perform the task, and unforeseen danger to the participant, experimenter, or equipment.

VIII. APPROVAL OF RESEARCH

This research has been approved, as required, by the Institutional Review Board (IRB) for projects involving human subjects at the Virginia Polytechnic Institute and State University and by the Grado Department of Industrial and Systems Engineering.

_________________________ ______________________________
IRB Approval Date Approval Expiration Date

IX. PARTICIPANT'S RESPONSIBILITIES AND PERMISSION

I voluntarily agree to participate in this study and I know of no reason why I cannot participate. I have read and understand the informed consent and conditions of this project, and understand that I have the following responsibilities: (1) to listen attentively to the stimulus sounds presented during the tests, to respond appropriately and accurately, and to follow all instructions to the best of my ability, (2) to notify the experimenter at any time about discomfort or a desire to discontinue participation. I hereby acknowledge the above and give my voluntary consent for participation in this project. If I participate, I may withdraw at any time without penalty.

____________________________ ___________________________________
Print Name Subject's Signature Date

Should I have any pertinent questions about this research or its conduct, and research subjects' rights, and whom to contact in the event of a research-related injury to the subject, I may contact:

R. Brian Valimont (Principal Investigator) - 540-231-9086 / rvalimon@vt.edu

Dr. John G. Casali, (Faculty Advisor) - 540-231-9081 / jcasali@vt.edu

David M. Moore 540-231-4991 / moored@vt.edu
Chair, IRB
Office of Research Compliance
Research & Graduate Studies

This Informed Consent is valid from ____________to __________________.
[NOTE: Subjects will be given a complete copy (or duplicate original) of the signed Informed Consent.]

Appendix B

Pre-Flight Instruction Sheet

The Effect of Active Noise Reduction on Speech Intelligibility in a

High-Fidelity Flight Simulation

Thank you very much for your participation. You are about to embark on a cross-country simulation, but before starting the flight there are a few instructions and requirements to review. First, you will be receiving the flight plan for the cross-country route, a flight log, detailing the checkpoints along the route, the destination airport, appropriate radio frequencies, the cruise altitude, airspeed, etc. Any flight equipment you will need, or have opted not to bring, will also be provided at this time. You will have 20 minutes to review the flight and ask any questions you may have concerning this material. When you are familiar with the flight, you will step into the simulator bay and we will start the simulation. There are a few key instructions that are slightly out of the ordinary, but very important to the experiment.

First, you must **OBEY ALL ATC COMMANDS.** This is very important! All ATC commands were designed for special reasons to the experiment. You must obey all the ATC commands, not matter how unlikely they sound.

Second, you are **NOT ALLOWED** to respond to any ATC commands with the readbacks "ROGER," "WILCO," or "UNABLE." It is very important that you respond to every ATC command with a full readback.

Lastly, there are performance requirements that must be met to the best of your ability as you conduct the cross-country flight. These performance requirements are in accordance with the FAA instrument-rating practical test standards. The performance requirements are as follows:

- Heading: stay within 10° of route designated magnetic heading or ATC assignment.
- Altitude: stay within 100 ft. of route designated altitude or ATC assignment.
- Airspeed: stay within 10 kts. of either route cruise airspeed, published airspeed, ATC assignment.
- Rollouts: within 10° of assigned heading.
- Use only standard, and half-standard turns.
- ILS Approach: Do not descend below DH, unless allowable under FAR's.
- NDB tracking / approach: within 10° of bearing or heading.
- VOR tracking / approach: no more than ¾ scale deflection.
- Maintain MDA +100ft., -0 ft.
- Complete appropriate checklists at associated phase of flight.

Do you have any questions concerning these instructions and requirements?

Appendix C

Modified Cooper – Harper Instructions

Modified Cooper-Harper Instructions

Key Definitions of Terms

Mental Workload: For purposes of the survey, mental workload includes all processes which are purely mental such as (but not limited to) attention, visualization, spatial orientation, decision-making, and memory. Mental workload also includes the combination of mental processes such as perception and the resulting physical actions. For example, the perception of information displayed on the flight instruments and the resulting physical control inputs are considered psychomotor workload, a type of mental workload. The same is true for the perception of weather's effects on the aircraft and the pilot's resulting actions.

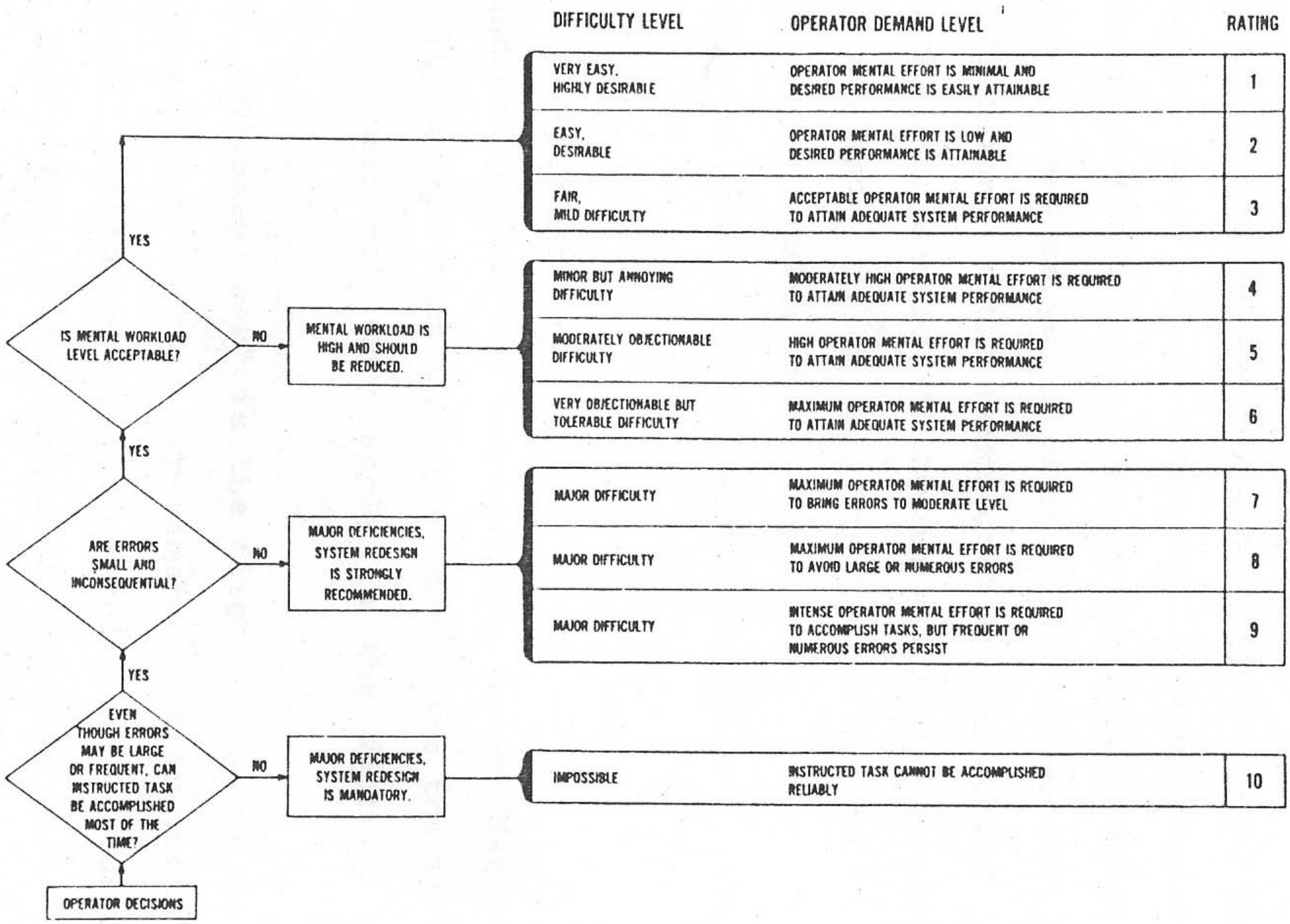
Rating Scale Steps

On the Modified Cooper-Harper scale you will notice that there is a series of decisions which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. Thus, you should follow the logic sequence on the scale for each of your ratings in this experiment.

Remember you are to circle only one number, and the number must be arrived at by following the logic of the scale. You should always begin at the lower level and follow the logic path until you have decided on a rating. In particular, do not skip any steps in the logic. Otherwise your rating may not be valid or reliable.

Appendix D

Modified Cooper – Harper Rating Scale


DIFFICULTY LEVEL
OPERATOR DEMAND LEVEL
RATING
VERY EASY, HIGHLY DESIRABIE
OPERATOR MENTAL EFFORT IS MINIMAL AND DESIRED PERFORMANCE IS EASILY ATTAINABLE
1
EASY, DESIRABLE
OPERATOR MENTAL EFFORT IS LOW AND DESIRED PERFORMANCE IS ATTAINABLE
2
FAIR, MILD DIFFICULTY
ACCEPTABLE OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE
3
MINOR BUT ANNOYING DIFFICULTY
MODERATELY HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE
4
MODERATELY OBJECTIONABLE DIFFICULTY
HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE
5
VERY OBJECTIONABLE BUT TOLERABLE DIFFICULTY
MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE
6
MAJOR DIFFICULTY
MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO BRING ERRORS TO MODERATE LEVEL
7
MAJOR DIFFICULTY
MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO AVOID LARGE OR NUMEROUS ERRORS
8
MAJOR DIFFICULTY
INTENSE OPERATOR MENTAL EFFORT IS REQUIRED TO ACCOMPLISH TASKS, BUT FREQUENT OR NUMEROUS ERRORS PERSIST
9
IMPOSSIBLE
INSTRUCTED TASK CANNOT BE ACCOMPLISHED RELIABLY
10
YES
IS MENTAL WORKLOAD LEVEL ACCEPTABLE?
NO
MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.
YES
ARE ERRORS SMALL AND INCONSEQUENTIAL?
NO
MAJOR DEFICIENCIES, SYSTEM REDESIGN IS STRONGLY RECOMMENDED.
YES
EVEN THOUGH ERRORS MAY BE LARGE OR FREQUENT, CAN INSTRUCTED TASK BE ACCOMPLISHED MOST OF THE TIME?
NO
MAJOR DEFICIENCIES, SYSTEM REDESIGN IS MANDATORY.
OPERATOR DECISIONS

Appendix E

Headset Comfort / Communications Quality Rating Scales

Below is a survey designed for you to rate several aspects related to the comfort of the headset you just wore. It also includes several topics regarding the quality of the integrated communication system in the headset. Please place an X in one of the seven spaces provided between each pair of descriptors that best shows your opinion. Please rate each scale as honestly and accurately as you possibly can.

Painless ___ : ___ : ___ : ___ : ___: ___ : ___ Painful

Uncomfortable ___ : ___ : ___ : ___ : ___: ___ : ___ Comfortable

No Uncomfortable ___ : ___ : ___ : ___ : ___: ___ : ___ Uncomfortable
Pressure Pressure

Intolerable ___ : ___ : ___ : ___ : ___: ___ : ___ Tolerable

Tight ___ : ___ : ___ : ___ : ___: ___ : ___ Loose

Not Bothersome ___ : ___ : ___ : ___ : ___: ___ : ___ Bothersome

Heavy ___ : ___ : ___ : ___ : ___: ___ : ___ Light

Cumbersome ___ : ___ : ___ : ___ : ___: ___ : ___ Not Cumbersome

Soft ___ : ___ : ___ : ___ : ___: ___ : ___ Hard

Feeling of Complete ___ : ___ : ___ : ___ : ___: ___ : ___ No Feeling of Complete
Isolation Isolation

Ear Open ___ : ___ : ___ : ___ : ___: ___ : ___ Ear Blocked

Ear Empty ___ : ___ : ___ : ___ : ___: ___ : ___ Ear Full

Low Fidelity ___ : ___ : ___ : ___ : ___: ___ : ___ High Fidelity
Communications Communications

Extraneous Noise ___ : ___ : ___ : ___ : ___: ___ : ___ No Extraneous Noise

Sound Distortion ___ : ___ : ___ : ___ : ___: ___ : ___ No Sound Distortion

Interferes with ___ : ___ : ___ : ___ : ___: ___ : ___ No Interference with
Communications Communications

Background Hum ___ : ___ : ___ : ___ : ___: ___ : ___ No Background Hum
Present Present

Low Overall ___ : ___ : ___ : ___ : ___ : ___: ___ : ___ : ___ High Overall
Communications Quality Communications Quality

Appendix F

Fatigue, Simulator Realism, and Engine Noise Realism Rating Scales

1) Please circle the number that best rates your present level of fatigue.

Exhausted ___ : ___ : ___ : ___ : ___ : ___ : ___ Well-Rested

2) Please mark the space that best rates the realism of the simulator engine noise as compared to a Cessna 172 engine noise.

Not Realistic ___ : ___: ___ : ___ : ___ : ____ : ___ Highly Realistic

3) Please mark the space that best rates the realism of the overall flight simulation as compared to real flight in a Cessna 172.

Not Realistic ___ : ___: ___ : ___ : ___ : ____ : ___ Highly Realistic

R. Brian Valimont

395 Warren St.
Christiansburg, VA 24073
Home: (540) 382-4433 Office: (540) 231-0462
Email: rvalimon@vt.edu

Education

Doctor of Philosophy, Industrial and Systems Engineering — Expected Graduation: May 2006
Specialization: Human Factors Engineering & Ergonomics,
Virginia Polytechnic Institute and State University, Blacksburg, VA

Master of Science, Human Factors and Systems — Graduation: Dec. 2002
Embry-Riddle Aeronautical University, Daytona Beach, FL,

Bachelor of Science, Human Factors — Graduation: Dec. 2000
Embry-Riddle Aeronautical University, Daytona Beach, FL

Research & Work Experience

Graduate Research Assistant / Dissertation Research — Aug. 2004 – Present
Virginia Tech

"The Effect of Active Noise Reduction on Speech Intelligibility and Pilot Performance in an Instrument Flight Simulation"

- Wrote numerous research proposals during my first year in the doctoral program pursuing funding independent of ongoing departmental projects.
- Co-wrote a successful proposal securing research funding from Bose Corporation to support a research experiment of my design
- Investigated the effects of active noise reduction on speech intelligibility, mental workload, pilot performance, and safety during an instrument cross-country flight simulation. For this experiment, four instrument cross-country flight simulations were designed which incorporated different workload levels induced by realistic flight tasks, changing weather patterns, and pre-recorded interactive air traffic control communications.
- In addition to experimental design, I was responsible for administrative and managerial duties. I completed all requirements to achieve approval from the Institutional Review Board (IRB) for use of human subjects, found and organized

human subjects who fit very specific requirements, and scheduled them for a long series of experimental trials.

- Motivated subjects to continue and complete a grueling experimental schedule which consisted of the first, one-hour, screening trial which established participation eligibility, then four, four-hour, exhausting experimental trials without losing one pilot.

- Developed new skills specifically for this experiment. Wrote two MATLAB programs, one which takes a wav file voice recording and degrades the speech intelligibility along a Speech Transmission Index function. Also, conducted pure-tone audiograms during the screening sessions.

Course Instructor Jan. 2006 – Present

Dept. of Engineering Education

- Instructor for undergraduate course in *Digital Technology* for electrical engineers, computer engineers, and computer science majors.
- Develop lectures, in-class exercises, homework projects, and MATLAB labs to provide students with a foundation in the creation and manipulation of digital audio, digital video, encryption, and networking.

Course Instructor Aug. 2005 – Dec. 2005

Dept. of Industrial & Systems Engineering

- Instructor for undergraduate / graduate course in *Occupational Safety and Hazard Control*
- Developed lectures, in-class exercises, homework projects, and exams to provide students with a foundation in safety and hazard control which later can be applied in private industry

Graduate Teaching Assistant Sept. 2003 – Aug. 2004

Virginia Tech

- Teaching assistant for Industrial Ergonomics.
- Conducted lectures in professor's absences, created homework assignments, as well as sections of coursework, tests and exams.
- Graded all coursework and maintained accurate grade spreadsheets.

Graduate Research Assistant / Thesis Research Aug. 2001 – Dec. 2002

Embry-Riddle Aeronautical University

- Investigated *augmented reality* as an instructional medium and compared human learning results with those of traditional instructional mediums (i.e. video, text, web-based)
- Led a small team of undergraduate and graduate students in the experimental design and experimental apparatus construction

Publications

- Valimont, R.B., Lancaster, J.A., & Casali, J.G. (submitted). ANR vs. Passive Communications Headsets: Investigation of Speech Intelligibility, Pilot workload, Comfort, and Hearing Protection in Long Duration Flight in an Aircraft Simulator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. San Francisco, CA: HFES.

- Valimont, R.B., Lancaster, J.A., & Casali, J.G. (2005). ANR vs. Passive Communications Headsets: Investigation of Speech Intelligibility, Pilot workload, Comfort, and Hearing Protection in Long Duration Flight in an Aircraft Simulator (ISE Department Technical Report 200503, Audio Lab Number 12/19/05-3-HP). Blacksburg, VA: Virginia Tech, Grado Department of Industrial and Systems Engineering, Auditory Systems Laboratory. (contractor's report).

- Vincenzi, D.A., Valimont, R.B., Macchiarella, N., Opalenik, C., & Gangadharan, S.N. (2003). The Effectiveness of Cognitive Elaboration Using Augmented Reality as a Learning and Training Paradigm.

- Valimont, R.B., Vincenzi, D.A., Majoros, A.E., & Gangadharan, S.N. (2002). The Effectiveness of Augmented Reality as a Facilitator of Information Acquisition. The IEEE 21st Digital Avionics Systems Conference. IEEE: Irvine, CA.

Honors & Awards

- United Parcel Service (UPS) Fellow, Jan. 2003- Aug. 2005
- Link Foundation Simulation and Training Fellow, 2002
- Irma Kirk Graduate Scholarship, 2001
- ISE Departmental Teaching Assistant of the Year, 2004

Computer Skills

- Microsoft: Word, Excel, PowerPoint, Access, and Visio
- SAS and SPSS statistical software
- MATLAB
- Raptor flowchart software
- Blackboard
- Hardware Skills

Technical & Specialized Skills

- Commercially Licensed Pilot with Instrument and Multi-Engine ratings
  - Accumulated a couple hundred hours of experience in visual and instrument flight conditions
- Automotive mechanics and restoration
  - Worked on the repair of various systems of automobiles
  - Performing a body-off restoration of a 1973 VW Super Beetle
- Read, write, and speak German
- Arc and Metal Inert Gas (MIG) welding
- Certified open water scuba diver
- Watercraft and propeller mechanics

Affiliations

- Human Factors and Ergonomics Society, 2002 – 2005
- Student HFES Chapter, Virginia Tech, 2003 – 2005
- American Society of Safety Engineers, 2003 – 2005
- Alpha Pi Mu, Industrial Engineering Honor Society, 2003- 2005
- Institute of Industrial Engineers, 2003 – 2005

Selected Graduate Coursework

- Auditory Display Design
- Human Factors System Design
- Human Physical Capabilities
- Human Factors Research Methods I, II
- Applied Multivariate Analysis
- Memory & Cognition
- Sensation & Perception
- Human Information Processing
- Applied Systems Engineering / System Dynamics
- Regression Analysis
- Occupational Safety & Hazard Control
- Usability Engineering
- Human Computer Systems
- Macroergonomics
- Systems Concepts, Theories, & Tools
- Operations Research
- Manufacturing Systems Engineering
- Aviation Psychology
- Aviation Accident Investigation
- Advanced Ergonomic Methods