Evaluation of a Reduced Modified Rhyme Test for Assessing Speech Intelligibility in Radio Communications

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Psychology
in the
College of Graduate Studies
University of Idaho
by
Patricia A. Dunavold

Major Professor: Brian P. Dyre, Ph.D.
Committee Members: Steffen Werner, Ph.D; James D. Brownlow, Ph. D
Department Administrator: Todd J. Thorsteinson, Ph.D.

May 2016

Authorization to Submit Thesis

This thesis of Patricia A. Dunavold, submitted for the degree of Master of Science with a Major in Psychology and titled "Evaluation of a Reduced Modified Rhyme Test for Assessing Speech Intelligibility in Radio Communications," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____
Brian P. Dyre, Ph.D.

Committee Members: _____ Date: _____
Steffen Werner, Ph.D.

_____ Date: _____
James D. Brownlow, Ph.D.

Department Administrator: _____ Date: _____
Todd J. Thorsteinson, Ph.D.

Abstract

It is essential that unmanned aerial vehicle (UAV) operators in one location communicate clearly with UAV operators in other locations as well as Command, Control, Communication, Computer, and Intelligence/Information (C4I) forces. In 2008 the Air Force used a Modified Rhyme Test (MRT) to evaluate the Speech Intelligibility (SI) of a new voice communications system on the Global Hawk UAV. The raw data from these tests were analyzed for differences in SI between a full MRT of 50 words and subsets of progressively decreasing size, hereafter referred to as a reduced Modified Rhyme Test (rMRT), using five less words for each subset. If no differences in SI exist between the MRT and a subset of size $n < 50$, then it could be concluded that test time and resources could be saved by using an rMRT. The results of this study concluded that an rMRT of 30 words would be sufficient.

Acknowledgments

I would like to thank Dr. Brian P. Dyre for his endless patience and guidance, and whose years'-long commitment helped ensure that this project was finally completed.

Thanks also to Dr. Steffen Werner whose suggestions served to improve the analysis of the study and the formatting of the thesis.

And I humbly offer my endless gratitude to Dr. James D. Brownlow whose expert statistical knowledge and advice were essential in the design of the analysis and the preparation of this thesis.

I also owe a debt of gratitude to the thirteen Air Force officers, civilians and contractors who participated in the original test effort and who generously signed waivers allowing me to use the data in the subsequent analysis in support of this thesis.

Dedication

First and foremost I would like to thank God who makes all things possible.

I'd also like to thank my son, Jon Thomas Dunavold, and his wife,

Danyelle Angelina Bossardet Dunavold. Their quiet support and gentle prodding

kept me going even when I was disillusioned and ready to quit.

Without their unwavering faith in me this project would never have been completed.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ACC	Airworthiness Certification Criteria
AFB	Air Force Base
AI	Articulation Index
AM	Amplitude Modulation
ANSI	American National Standards Institute
ATC	Air Traffic Control
AV	Aerial Vehicle
BB	Baseband
BLOS	Beyond-Line-of-Sight
C2	Command and Control
C4I	Command, Control, Communication, Computer and Intelligence/Information
CL	Command Link
CRM	Coordinate Response Measure
CV	Consonant-Vowel
CVC	Consonant-Vowel-Consonant
DRT	Diagnostic Rhyme Test
FM	Frequency Modulation
GH	Global Hawk
ICAO	International Civil Aviation Organization
INMARSAT	International Maritime Satellite
KU	Ku SATCOM
LOS	Line-of-Sight
LRE	Launch and Recovery Element
MCE	Mission Control Element
MRT	Modified Rhyme Test
NS	Non-Secure
PB	Phonetically Balanced
RL	Return Link

rMRT	Reduced Modified Rhyme Test
RT	Rhyme Test
RV	Recreational Vehicle
S	Secure
SATCOM	Satellite Communication
SI	Speech Intelligibility
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
STI	Speech Transmission Index
SUT	System Under Test
UAV	Unmanned Aerial Vehicle
UHF	Ultra High Frequency
USAF	United States Air Force
VC	Vowel-Consonant
VCET	Voice Communications Effectiveness Test
VHF	Very High Frequency

Testing Radio Equipment

The ability to communicate is a vitally important skill among all humans. For most humans the simplest form of communication consists of the ability to produce sounds and the ability to hear those sounds. Arguably, of our five senses, hearing is the sense we count on most often to ensure the communication of all our needs. One of the most famous examples of this is when Alexander Graham Bell spoke the now famous phrase, “Mr. Watson, come here! I want you!” over his newly invented telephone (Bell Telephone Magazine, 1947). Since Bell’s uttering of that famous phrase electronic communication equipment has become increasingly sophisticated but still requires careful selection and testing to ensure that the right equipment is placed in the proper environment and that the resulting communications are clear and intelligible.

Testing communication equipment is something that is done fairly frequently in a military environment. Speech Intelligibility (SI) tests are time consuming and expensive but need to be accomplished whenever any part of the electronic communication system is changed or upgraded in military aircraft in order to meet Airworthiness Certification Criteria (ASC/EN, 2005) and attain Airworthiness Certification. The standard test used to measure speech intelligibility is a Modified Rhyme Test (MRT). Modified rhyme tests require a huge number of assets, from the labor of multiple participants to lengthy and expensive flight time and the costs associated with a single MRT can be thousands of dollars. The most recent MRT conducted on an RQ-4B Global Hawk (GH) drone was accomplished in 2008 and required approximately 4 hours of costly flight time. However the operational cost

per hour didn't include the labor costs of the 10 participants and the two Human Factors proctors that were required to run the MRT. And, relative to other aircraft, the RQ-4B Global Hawk Drone is fairly inexpensive to operate. Other aircraft, such as the B-52 Stratofortress Bomber ($69,708) or the C-5B Galaxy Cargo Plane ($78,817) cost even more per hour to operate (Thompson, 2013). Multiply that by the need to run an MRT every few years on multiple different aircraft platforms and the costs can quickly run into hundreds of thousands of dollars.

Additionally, many participants complain that the MRT that is used to measure speech intelligibility is mind-numbingly boring due to the repetitive nature of the test. This makes it difficult to train and maintain a corps of experienced test participants since many participants are reluctant to participate in additional MRTs after their first experience. Anything that can be done to reduce the costs of conducting MRTs, and enhance the experience for the participants, would be beneficial to the United States Air Force (USAF) and other entities that rely on reliable and accurate radio communications.

In 2008 the USAF conducted a series of four separate 50-word MRTs during flight test using a new communications system on the GH unmanned aerial vehicle (UAV) (Dunavold & Herrera, 2009). For the present study, and using the raw data obtained in the 2008 GH MRT, a bootstrap procedure (Hastie, Tibshirani, & Friedman, 2009) was used to compare the SI scores of the full 50-word MRT set to the SI scores of seven different subsets of various word sizes to determine if an MRT using fewer than 50 words would yield the same results, thereby allowing the

USAF to continue testing new communications equipment by using fewer word sets while saving resources and maximizing the tight flight test budget.

Human Production and Perception of Speech

To understand speech concerns in radio communication it is useful to first understand how humans produce and perceive the spoken word. Although no two humans are exactly alike we can still produce sounds that are so similar that we can understand each other and therefore engage in meaningful communication.

Perception of spoken words has been studied for decades (Rosenzweig & Postman, 1958). Three main findings resulted from those early studies: (1) listeners can identify words more accurately if they know the list from which the words will be chosen and the shorter the list the more intelligible the words will be, (2) a list of alphabetic equivalents will be an aid to intelligibility only when the speaker and listener agree to them, and (3) the greater the length of a word the more intelligible it tends to be when frequency of usage is held constant and words in a list will be more intelligible if they cannot be readily confused with one another.

One particular study (Rosenzweig & Postman, 1958) reviewed the results of some of these early audition studies and compared them with the results of visual word recognition experiments. They found that while both vision and audition share the beneficial effects of frequency of past usage and restriction of alternatives, the length of the stimulus words had an inverse effect on the two sense modalities. Speech intelligibility increased with the length of a word when it was presented auditorily but word recognition decreased with the length of a word when it was presented visually. These, and other, studies support the International Civil Aviation Organization (ICAO) requirement for aviators to use a small set of official aviation phraseology for routine operations and the requirement to use a standard list of

internationally agreed upon alphabetic equivalents, commonly known as the ICAO alphabet, in an effort to improve SI (ICAO, 2007). An ICAO alphabet and pronunciation guide is presented in appendix A.

Other researchers were interested in speech production. House and Stevens (1955) were interested in exactly how humans form the different vowel sounds. They used a series of x-ray pictures to map the specific physical dimensions of the vocal tract by which humans consistently create vowel sounds. Based on the x-ray data House and Stevens concluded that the articulation of vowel sounds are dependent on three different dimensions: (1) the position of the tongue constriction (or height of the tongue), (2) the size of the constriction formed by the tongue, and (3) the dimensions of the mouth opening. For example, the sound of the vowel /u/ is mainly determined by the rounding of the lips while the production of the vowel /i/ is fairly insensitive to mouth opening changes, but is primarily produced by the interaction of the tongue position and the degree of the constriction of the tongue. All the vowel sounds are similarly produced by a combination of the three different dimensions described above. Using this early research House and Stevens developed of a set of parameters that yielded a simple, yet reasonably accurate, description of the articulation of vowel sounds.

These findings are consistent with the findings of other researchers (Peterson & Barney, 1952; Nearey, 1989) and support the determination that these three dimensions are descriptive of human vowel articulation. House and Stevens (1955) used the information about the production of vowel sounds to reproduce the sounds

using an electrical analog of the vocal tract which were deemed, by students of speech, as being characteristic of human speech.

House and Stevens (1955) continued their research with a set of experiments using the electrical vocal-tract analog they had developed to produce simulated sounds. They found that the human participants could correctly and appropriately identify the different vowel sounds produced by the vocal-tract analog equipment. The results were used to produce contour plots of the frequency of correct and incorrect responses and their results compared favorably with the contour plots of the results produced from an earlier study done by different researchers (Peterson, G. E. & Barney, H. L., 1952).

These early studies helped psychologists understand speech production and led to further research into speech intelligibility testing.

Speech Intelligibility Testing Using Communications Equipment in an Operationally Representative Environment

Speech intelligibility (SI) testing has been accomplished for the last sixty or seventy years. As early as 1947, Kryter, Licklider and Stevens were researching methods to improve intelligibility in voice transmissions using amplitude modulation (AM) radio telephony. They tested the theory that the intelligibility of speech heard over communications equipment of limited power capability could be improved if the intensity of the weak speech sounds (consonants) were increased relative to that of the intense speech sounds (vowels). The method used to induce this effect was peak clipping which involved passing the speech signal through symmetrical nonlinear circuits that limit or clip the high-amplitude waves.

Their experiments simulated an operational flight environment and included quiet conditions for both speakers and listeners, speakers in quiet conditions and the listeners exposed to simulated airplane noise of 120 decibels to emulate a Navy PBY patrol bomber, simulated intense thunderstorm static, and exposing both the speakers and listeners to the simulated airplane noise to emulate the airplane noise picked up by the microphones. In all conditions except one the speech intelligibility was improved with peak clipping. However, in the experiment where both speakers and listeners were exposed to the simulated airplane noise speech intelligibility scores were lower. The researchers concluded that interference entering the communication system after the premodulation clipper had little effect on the amount of improvement and that it was not possible to recommend a standard amount of clipping for all situations.

Both physical and physiological problems occur when using an interphone or radio voice-communication system at high altitudes. Reduced ambient pressure at high altitudes can make breathing difficult, talking arduous, and loud speech impossible. Oxygen masks must cover the face and nose requiring the use of special microphones which are located inside the facial mask and which present the user with a set of special acoustical problems such as inadvertent clipping, which is when part of the message doesn't get transmitted, or *stepping on* other radio transmissions, which occurs when radio transmissions on the same frequency are transmitted at the same time and one transmission obscures or *masks* the other transmission. Additionally, microphones and earphones change in sensitivity and frequency at the different altitudes which increases the complexity of using radio voice-communication systems at high altitudes.

Miller and Mitchell (1947) developed methods for improving speech intelligibility while using an interphone or radio voice-communication system at high altitudes. Using an altitude chamber to simulate the pressure conditions of high altitude, they conducted a set of experiments using two types of articulation tests. Miller and Mitchell found that premodulation clipping was advantageous both at sea level and at high altitudes. When no clipping was used, the peaks of the sea level speech had to exceed the noise level by at least 6 decibels before 50% of the words were correctly received. At high altitude a signal-to-noise ratio of 15 decibels was required for a 50% articulation rate. However, when peak clipping was used the speech became intelligible at lower signal-to-noise ratios.

Auditory masking was also a phenomenon that many scientists were interested in understanding (Miller, G. A., 1947). Humans actually have the ability to hear two sounds at once and this ability can be beneficial. However, it can also be problematic such as when one sound masks a fainter but more important sound which sometimes occurs when radio transmissions get *stepped on*. Miller (1947) conducted articulation testing using three thresholds of speech: (1) the threshold of detectability, or the point when a sound or noise can be detected but cannot be perceived as a word, (2) the threshold of perceptibility, which is the point when the sound can be perceived as a word but not understood, and (3) the threshold of intelligibility, which is the point when the listener can understand the word well enough to perceive the meaning of the word. All three thresholds are distinct and reliable, and different listeners will agree on their value. Any one of the three thresholds can be used to determine the shift in the threshold due to the presence of a masking sound.

Several techniques were used to mask sound such as tones, noise, and other voices. For all three masking techniques, the stimulus-dimensions determining masking were: (1) the intensity, (2) the frequency or spectrum, and (3) the temporal pattern of the sound. Masking depends primarily on the speech-to-noise ratio over the range of frequencies involved in speech. Sounds of low frequency mask this range more effectively than sounds of high frequency. Additionally, interruptions in the sound decrease the masking effectiveness.

In 1951 Miller, Heise, and Lichten defined an *articulation test* as lists of syllables, words, or sentences read by an announcer to a group of listeners who

report what they hear. They defined *articulation score* as the percentage of discrete test units reported correctly by the listeners. This method was used to derive a quantitative evaluation of the performance of a specific speech communication system. Miller, Heise and Lichten described the three different classes of variables involved in this type of research: (1) personnel, the talkers and listeners involved in the test, (2) the test materials, which consists of sets of syllables, words, and sentences, or a continuous discourse; and (3) the communication equipment, which consists of the rooms, microphones, amplifiers, radios, earphones, etc. used in the tests.

Their 1951 study evaluated the items that make up the test materials and specifically addressed three types of contexts: (1) the context that is supplied by the knowledge that the test item is one of a small set of items, (2) the context that is supplied by the items that precede or follow a given item in a word or sentence, and finally, (3) the context that is supplied by the knowledge that the item is a repetition of the item that immediately preceded it. Miller, Heise, and Lichten (1951) were attempting to determine why an item (word) is heard correctly in one context but not in another by comparing the intelligibility of three different kinds of test materials: (1) words in sentences, (2) individual words from a pre-selected vocabulary, and (3) repeated words.

They concluded that their 1951 study supported the argument that it is not the particular item (word) but the context in which the item occurs that determines its intelligibility.

Another series of tests were conducted where the listener was given a target word which was automatically repeated once or twice against a set of tests where the target word was repeated only if the listener requested it be repeated. Again, the accuracy was increased with each repetition across all noise levels; however, the highest accuracy was recorded when the listener was able to select the target word from a set of pre-determined words.

These tests would seem to support the underlying philosophy for the use of the ICAO phonetic alphabet (Alpha, Bravo, Charlie, etc.) where the listener knows in advance the 26 letter designators.

Several years later Miller and Nicely (1955) conducted extensive testing on the overall effects of noise and frequency distortion on the intelligibility of human speech. The results of these types of tests had usually been reported as an articulation score which is the percentage of words the listener heard correctly. But this scoring method gives all errors equal weighting and does not discriminate among the different errors a listener could make. In fact, before 1955 little testing had been accomplished to determine the kinds of errors and their frequency. Miller and Nicely theorized that knowing the types of errors that occur and their frequency would help to minimize such errors and perhaps lead to improved communication.

In an effort to make the analysis manageable Miller and Nicely (1955) decided to compare only 16 consonants. These 16 consonants were chosen because they make up almost three quarters of consonant use in normal speech and represent about 40% of all phoneme use including vowels. A phoneme is the smallest unit of speech which, if changed, will change the meaning of the word. A

list of 200 nonsense syllables formed with the 16 consonants under investigation was used in the tests. Five female participants served as both speakers and listeners.

Miller and Nicely (1955) concluded that when consonant sounds were spoken over a voice communication system with noise or low-pass filtering several consonant sounds were reliably confused. But with high-pass filtering the errors (confusions) were random. These, and other studies, led to the development of the Modified Rhyme Test (MRT).

Development of the Modified Rhyme Test (MRT)

In 1958 Grant Fairbanks published a paper on research he had accomplished with a set of phonetically balanced words that rhymed (Fairbanks, 1958). The research described the development of a method for testing word identification in which the cues for response were confined to the initial consonant sounds and consonant-vowel (CV) transitions known as *phonemic differentiation*. This test was developed to fill a need for experimental materials that fit four criteria: (1) the stimulus unit would be a spoken word, (2) the response would be recognition of the word, (3) the response would depend on the initial consonant sound and the CV transition, and (4) it would closely follow the discrimination demands of real speech. Additionally, it was desired that the test be short, easy to administer, and suitable for groups as well as having several different, but equal versions so that the tests could be administered several times to the same participants while allowing them to remain naïve to the vocabulary.

Fifty sets of five rhyming words were selected as the stimulus words. All 250 words were three-, four-, and five-letter words which differed only in the initial consonant within each set of five words. All words were strictly *orthographic* in that each set of five words had the same *stem*, meaning that all the letters after the initial consonant was the same for all five words in each set. For example the words hot, got, not, pot, and lot form one set and the words wake, take, make, cake, and lake form another set. The participant's response sheet had a list of stems with the initial letter missing in the order that the stimulus words would be presented. The

participant recorded the letter that they thought they heard in the front of each of the corresponding stems.

Great care was taken in the construction of this test to ensure that the stimulus words were words of high familiarity within the English language. All 250 stimulus words were among the 9000 most frequently used words of the English language, while 200 of the words were among the 3000 most frequently used, and 112 were among 1000 of the most frequently used words. Additionally, the stimulus vocabulary contained 18 consonant phonemes and 13 vowels and *diphthongs*. A diphthong is formed when two adjacent vowel sounds occur within the same syllable creating a different and unique sound such as the /oy/ in boy and the /ou/ in about. The 18 consonants used account for approximately 90% of all consonant occurrences in the English language but the vowel distribution was slightly flatter than in normal usage.

The test stimulus words were recorded individually by Fairbanks. Care was taken to vocalize the words clearly and with the same amount of effort. The recorded words were then presented experimentally to 40 university students who served as the participants. The students were broken into 2 groups of 20. Each participant was exposed to the complete vocabulary once but not at all signal-to-noise levels. An in-depth analysis of the results showed that the phonemic differences strongly followed a regular hierarchy with certain sounds being correctly identified on a consistent basis.

Speech intelligibility testing is an expensive and time consuming process. Subjective testing methods require the use of trained personnel who act as speakers

and listeners. Oftentimes the required number of trained personnel is just not available. The Articulation Index (AI) is a method for calculating from physical and acoustical measurements made on a communication system a measure that is indicative of the intelligibility scores that would be obtained for that system under actual test conditions. The AI is computed from acoustical measurements or estimates, at the ear of a listener, of the speech spectrum and of the effective masking spectrum of any noise which may be present. (Kryter, 1962)

Kryter (1962) described the two methods commonly used for computing AI: (1) the 20-band method and (2) the one-third-octave-band and the octave-band method. The 20-band method uses measurements or estimates of the spectrum level of the speech and noise present in each of 20 contiguous bands of frequencies. At equal signal-to-noise ratios these bands contribute equally to speech intelligibility. The 20-band method is a five-step process that involves plotting the spectrum levels of the speech peaks that reach the listener's ear and then applying several algebraic calculations. The result is an AI number for that particular speech communication system at the noise conditions and level of speech specified in the test conditions.

The second method is a derivation of the 20-band method that uses measurements or estimates of the speech and noise present in certain one-third-octave bands or certain full octave bands. The octave-band method is not as sensitive to variations in speech and noise as the 20-band or one-third-octave-band method and is therefore not as precise. The one-third-octave-band and the octave-band methods are a six-step process that also involves plotting various band levels

or speech peaks and applying algebraic calculations to derive the AI score. The AI scores can then be converted to estimated speech-intelligibility scores. However, these methods don't actually measure SI and require complex, multi-step analysis methods making them unsuitable for use in an operational environment.

This early research lead to the development and evaluation of a new speech-intelligibility test designed to be used by operational personnel in an operational setting to evaluate the performance level of speech-communication systems. The format of the new test is similar to the classic Rhyme Test (RT) developed by Grant Fairbanks (1958) but uses a closed-response answer set instead of an open-response format (House, Williams, Hecker, & Kryter, 1965).

Since the test was designed to be used by operational personnel it was desired that the procedures and methods be quick, easy to administer and score, utilize little or no special equipment, except for the system under test (SUT), utilize relatively untrained personnel, and be reliable. This is in sharp contrast to earlier tests that were time consuming and often tedious, required specialized equipment and thoroughly trained speakers and listeners, and which were usually conducted in a laboratory setting.

Ultimately known as the Modified Rhyme Test (MRT) this test consists of 50 sets of 6-word ensembles comprised of 300 monosyllabic words. The majority of the words are in the consonant-vowel-consonant (CVC) form. In one half (or 25) of the word sets the initial consonant phoneme is the same while the final consonant phoneme or phonemic cluster changes, i.e., bat, bad, back, bass, ban, and bath. In the other 25 word sets the initial consonant phoneme differs but the final consonant

phoneme or phonemic cluster remains the same, as in meat, feat, heat, seat, beat, and neat. The stimulus words are always provided by means of a carrier sentence in the form, “Number blank is _____”, where the word “blank” is replaced with the set number and the blank space after the word “is” is replaced with the stimulus word, i.e., “Number one is meat”.

The main method in which the MRT differs from the RT is that the listener is given a closed-response answer set (as opposed to the open-response format used in the RT). This means that the listener is given a list of six words (usually grouped in two rows of three words each and enclosed in a rectangular box) from which to choose the correct response. The stimulus word (the word verbalized by the speaker) is always among the six potential answer choices. Each set of six words makes up a response ensemble and any of the six words in an ensemble can serve as the stimulus word for that particular ensemble. This means that the 50 response ensembles can be randomized to provide different test lists and various word arrangements within the ensemble can be used to prevent possible spatial response bias by the listeners.

This format reduces the amount of listener training considerably since the listener is now only required to be able to read English at a basic level and only has to be trained in the proper flow (box numbers 1 through 50 sequentially, and usually down the page and then across), the speed in which the words are given (usually one every 3 seconds), and the simple procedure of circling (or somehow marking) their word choice. The majority of the training falls upon the speakers who must practice the words until they can pronounce every word correctly, must learn the

carrier sentence, must ensure that they are not placing undue emphasis on the stimulus word, must ensure that the carrier sentences are verbalized at the same speed, and must ensure that there is approximately a 3-second pause between each stimulus presentation.

Several other ways in which the MRT differs from the RT are that the words are not phonetically balanced, the strict orthographic constraints imposed by Fairbanks (1958) are not followed, and word familiarity and relative frequency of occurrence rules are not imposed. Specifically, the frequency of occurrence of the variable consonantal elements in the stimulus word lists and the frequency of occurrence of these same sounds in the English language is about 60%. Only 100 of the 300 words are among the 1000 most common words and 39 of the words occur only five times or less in one million words. However, the word lists retain a high degree of phonemic similarity from test form to test form and contain representatives from the major classes of speech sounds. Additionally, care was taken to exclude any exotic or potentially offensive words.

After much testing, House, et al. (1965) concluded that when the responses are confined to given English words, the tests result in stable responses with no learning effect across the different test lists. However, House et al. did find a definite difference between the results across the two different speakers used for the tests suggesting that the speakers provide a larger experimental variable than had been anticipated. Furthermore, it was not immediately apparent what caused the differences in the scores between the two speakers and only an in-depth analysis revealed that the less-intelligible speaker's words were characterized by greater

vowel length than those of the more-intelligible speaker. This suggests that great care should be taken in the selection and training of speakers and that a larger number of speakers should help to reduce the variability in the resulting test scores. Additionally, House, et al. (1965) concluded that since their stimuli consisted of sets of words, primarily in the form of CVC, which differ only in the initial consonant phoneme or final consonant phoneme or phonemic cluster within each ensemble, an error can be regarded as a phonemic confusion and the tests become useful for diagnostic testing.

Although the MRT was developed for use with operational personnel to operationally determine the performance level of speech-communication systems Griffiths (1967) has taken the MRT and further modified it for use as a diagnostic test. Previous studies (Miller & Nicely, 1955; Nearey, 1989) have shown that confusions in the manner of articulation, place of articulation, and voicing are all independent. Griffiths theorized that a procedure that tested for all these one-dimensional confusions could also be used to validly predict multi-dimensional confusions.

The revised word lists developed by Griffiths (1967) consisted of 250 English monosyllabic words. Primarily in the form of CVC, these words were arranged in five lists of 50 words. The words were grouped in sets of five words that had a consistent vowel pronunciation throughout but did not adhere to the strict orthographic standards insisted on by Fairbanks (1958) in his Rhyme Test (RT). One hundred and fifty of the words were from the word lists developed for the MRT by House, et al. (1965) and the remaining 100 words were new. Generally, the new

words were not as common as the words they replaced nor was any effort made to achieve phonetic balance similar to that in the English language. This test also used a closed-response format similar to that of the MRT but, unlike the MRT, no carrier sentence was used in the administration of this test.

The results of these tests were evaluated in a manner similar to the one used by House et al. (1965) for the MRT to see if the same results and conclusions could also be applied to this modification. Overall, the results were similar in that there was no apparent learning effect with repeated administration of the tests; however, House et al. found that the initial consonant sounds were more easily recognized in their MRT. This finding did not hold true with the modification devised by Griffiths (1967) as the final consonant sound was more easily recognized in two of the five word lists. Despite this finding Griffiths concluded that the modification was nearly as stable and repeatable as the original MRT and was preferable for use when a complete and detailed analysis of the confusions of the initial and final consonant sounds was desired.

Kryter and Whitman (1965) conducted an experiment with a crew of eight listeners utilizing both the phonetically balanced (PB) word test (1000 monosyllabic words in twenty 50-word lists) and the modified rhyme test (300 monosyllabic words in 50 six-word lists). The results were then compared to other experiments by different researchers (Miller, Heise, & Lichten, 1951; Nickerson, Miller, & Shyne, 1960; and Fairbanks, 1958) using the PB-word test, rhyme tests, and the modified rhyme tests. In this experiment the words were delivered through headphones over a high-quality speech system with various different signal-to-noise ratios.

Kryter and Whitman (1965) acknowledged that there are several differences in test difficulty between the rhyme and PB word tests. They attributed this difference to the joint effects of at least two factors: (1) the number of response alternatives and (2) the amount of acoustic and linguistic information in the test words. For example, contained within the PB word tests there are several linguistic and acoustic cues, that, if heard, are sufficient to allow the listener to correctly identify a particular word. However, the PB test is difficult because of the large number of response alternatives (1000). In contrast, the words in the rhyme tests are difficult because multiple linguistic and acoustic cues are not available to the listener. The listener must decide on the correct word based on accurately hearing the different initial or final consonant sound of each word. But this choice is made easier by the fact that the listener gets to choose the correct word from a list of five to six alternatives.

Kryter and Whitman (1965) concluded that, although the Fairbanks rhyme test and the modified rhyme tests have definite advantages in terms of ease of test administration, scoring, analysis of results, and minimal speaker and listener training required; these tests are not as efficient a measure of the effects of broad-band noise on word intelligibility as the PB-word test employing all 1000 PB words.

More recently a broad review of the literature on the speech intelligibility of competing messages and the masking of speech was conducted by Ericson and McKinley (1997). They also reviewed the literature on the detection of speech. This review was included to describe the factors that can affect speech intelligibility. Six general areas of literature were reviewed: (1) monaural aspects of speech

intelligibility, (2) multi-channel (left-eared and right-eared) presentations over headphones, (3) lateralized speech signals, (4) free-field talkers and maskers, (5) multi-path interference, and (6) headphone presentations via manikins and synthesizers.

After Ericson and McKinley (1997) reviewed the literature they conducted a series of five experiments. The experiments used 12 paid volunteers, six males and six females. The speakers were grouped into pairs consisting of two males, two females, or a male and a female. All participants served as both speakers and listeners. Test materials consisted of phrases and words that were typically used in a military environment. Phrases were comprised of six words that formed meaningful and sensible thoughts. Speech intelligibility performance was measured using the coordinate response measure (CRM) or the voice communications effectiveness test (VCET). The CRM is a nonstandardized test used to measure the speech intelligibility of simultaneous talkers. The VCET was specifically designed to measure the amount of information transfer in airborne communications that is typical in a military environment.

In all five experiments speech was presented diotically, the same signal was presented to both ears simultaneously; dichotically, different signals were presented to right and left ear simultaneously; or directionally over headphones in a controlled laboratory setting. Specifically, the five experiments included: (1) speech intelligibility in different directions, (2) diotic, directional, and dichotic presentations of speech in ambient noise, (3) information transfer and speech intelligibility, (4) four

competing messages, and (5) selective attention (speaker location) and speech intelligibility.

In all five experiments Ericson and McKinley (1997) found that female speakers tended to mask each other the most and produced the lowest levels of intelligibility. Male speakers masked each other less than the female speakers while mixed gender speakers masked each other the least. They also identified several parameters affecting directional speech intelligibility and concluded that although the cocktail-party effect (Cherry, 1953) cannot be measured with just one experiment several of the findings from their five experiments were consistent with this phenomenon.

Erickson and McKinley's (1997) research highlighted the fact that there are distinct differences between male and female speakers and underscored the importance of including female speakers in any study that intends to generalize the results to a population that includes females.

An even more recent review by Steeneken (2003) included two commonly used types of speech assessment methods. The first method, subjective assessment, is based on the use of speakers and listeners. The second method, objective assessment, is based on the use of physical parameters of the transmission channel. The subjective assessment method, which requires at least four speakers and four listeners, is labor intensive and the results will depend on the individual subject's responses. However, the objective assessment method does not actually measure speech intelligibility but only determines the physical parameters to predict intelligibility according to a certain model. Therefore this

method, the objective assessment method, is not suitable for use in an operationally representative environment.

Steeneken (2003) defined speech intelligibility, explained how it differed from speech quality, and described how subjective intelligibility assessment is conducted. Subjective intelligibility assessment is based on the measurement of the number of phonemes, words (either real words or nonsense words), or sentences heard and understood. Various techniques for the administration of these different stimuli are used and the response method utilized could be open or closed. An open-response method allowed the listener the freedom to respond to what they thought they heard while a closed-response method gave the listener a set of responses from which to choose. Steeneken described two of the various different subjective intelligibility assessment tests, i.e. Diagnostic Rhyme Test (DRT) and MRT, and the different methods of scoring that are typically used with each method.

Steeneken (2003) also described the objective intelligibility assessment process and the two most frequently used objective measures, the Speech Transmission Index (STI), and the Speech Intelligibility Index (SII). The STI measures an artificial test signal (instead of an actual speech signal) and the SII also includes the actual physical properties of the transmission channel which make these methods unsuitable for measuring SI in an operational environment.

Overall, as all these studies show, testing using voice communications equipment presents many different challenges but none more important than attempting to test in an operationally representative environment.

In 2007 Cole conducted an experiment to see if a reduced modified rhyme test (rMRT) would yield comparable results to a (full) MRT. The purpose of the study was to reduce the amount of costly flight time that is currently required to validate new and improved communications systems in military aircraft.

The 2007 study used the original word list from the House, et al. (1965) MRT and was administered to 39 participants in three different levels under three different conditions. A full MRT, using 50 words, a reduced MRT (rMRT) using 30 words, and an extremely shortened MRT using 10 words were administered three times each to all participants. The word lists were administered at three different signal-to-noise ratio (SNR) levels, 15db, 0db, and -5db, and were delivered in a laboratory setting using a voice recording of the same male voice for all conditions.

Cole (2007) concluded that the SI scores were equivalent for the 30- and 50-word lists at all SNR levels but that the 10-word list did not produce equivalent SI scores. Furthermore, the 30-word list resulted in a 39% time savings over the full MRT. Cole recommended further research be accomplished comparing the SI results of a 30-word MRT and a 50-word MRT in an actual flight test environment.

In 2008 the USAF conducted a series of four separate 50-word MRTs in an actual flight test environment using a new communications system on the Global Hawk (GH) unmanned aerial vehicle (UAV) (Dunavold & Herrera, 2009). All four tests used the same 10 speakers and listeners (however only 9 speakers and listeners were available for two of the four tests) and the same experimental test controls. Four different radio frequencies were used and all four of the tests passed with an SI score of 80% or higher. For this study, and using the raw data obtained in

the 2008 GH MRT, a bootstrap procedure (Hastie, Tibshirani, & Friedman, 2009) was used to compare the SI scores of the full 50-word MRT set to the SI scores of seven different subsets with a total of 13 different speakers.

Overall Methods and Conditions

The current system under test (SUT) consisted of the voice communications from the Mission Control Element (MCE) using Ku satellite communication (SATCOM) audio pass-through to the Air Traffic Control (ATC) radio in the GH aerial vehicle (AV), and subsequently line-of-sight (LOS) transmission to the external radio, representing an ATC center, and vice versa. Primarily, the test item was the ATC radio, AN/ARC-210 model RT-1794C, integrated within the GH AV segment. The GH used in the test was an unmanned AV capable of autonomous flight or it could be controlled from the ground by pilots in the launch and recovery element (LRE) and the MCE. During normal operations, the LRE pilot controlled the AV during taxi, takeoff, and landing. Once the AV reached a certain predetermined altitude, the LRE pilot handed off the AV to the MCE pilot to continue the flight. In preparation for landing, the MCE pilot returned control of the AV to the LRE pilot for landing and taxi operations.

An MRT was conducted to evaluate the SI of the communications system. The MRT has been adopted by the American National Standards Institute (ANSI) as the standard for the measurement of monosyllabic word intelligibility, ANSI S3.2-1989, *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, 1989) and has been approved for use with all speech intelligibility testing that is supported with government funds. ANSI S3.2-1989 describes how testing should be conducted and covers the scope and purpose of the document, applications of the standard, definitions, general guidance for testing, how to select and train the speakers and listeners, how to select the test materials,

how to conduct speech intelligibility tests, how to measure and analyze the results of the tests, and how to report the results of said tests. It also includes a list of references and an appendix of approved test materials, as well as a brief discussion of the three types of speech intelligibility tests; Phonetically Balanced (PB), Modified Rhyme Test (MRT), and Diagnostic Rhyme Test (DRT), and guidance for choosing from the three options along with complete word lists for each of the three options.

Since this study utilized military, civilian and contractor employees of the federal government, and therefore was supported by government funds, ANSI S3.2-1989 was used as the final authority on all design, methodological, and procedural decisions.

Participants

According to ANSI S3.2-1989 there should be a minimum of five speakers and five listeners for each MRT, however it is acceptable to have more listeners than speakers. Ultimately, 13 Edwards Air Force Base (AFB) employees were used as participants in the test, seven males (three civil servants, three active duty Air Force officers, and one contractor) and six females (three civil servants and three contractors). The age range for the participants was 24 to 55 years (mean = 36). All speakers and listeners were trained on MRT methods prior to the start of formal testing. All participants were required to achieve a 90% accuracy rate (both as a speaker and as a listener) before they were certified to be eligible to participate in the formal tests. All participants were military, civilian, or contractor employees of the Air Force whose hearing had been tested within the previous year and who had passed the standard H-1 hearing test which is required for military aircrew. A study

conducted by Erickson and McKinley (1997) found that female speakers tended to mask each other the most and produced the lowest levels of intelligibility. Additionally, in today's military, more and more females are serving as pilots, aircrew, and air traffic controllers, positions that require extensive communications using electronic communications equipment. Therefore, for this test, at least 40% of the speakers and listeners were female. Because this data was being collected during routine flight test as a part of already scheduled communications systems testing with military, civilian and contractor personnel, informed consent was not necessary. Military, civilian and contractor personnel who work for the federal government are presumed to have given consent to any job requirement when they take the oath of office and swear to uphold the constitution of the United States of America. However all participants signed a consent form after the fact to allow the data to be used in this analysis. A sample consent form is presented in appendix B.

Stimulus Materials

An MRT was administered to all participants during two different missions. The words used were the standard 300 words selected by House et al. in their 1965 study. A closed-format answer set was used. The word sets were grouped in ensembles of six words, composed of two lines of three words each, and enclosed in a box. The word verbalized by the speaker was always one of the six words in the box. The 50 response ensembles were randomized to provide different test lists, and various word arrangements within ensembles were used to prevent possible spatial biases by the participants. All the words have been loaded into a computer program, Visual Basic with an EXCEL overlay, which was used to randomize and

counterbalance the word lists. The 300 stimulus words arranged in response ensembles are shown in Figure 1.

1	went	sent	bent	18	way	may	say	35	heat	neat	feat
	dent	tent	rent		pay	day	gay		seat	meat	beat
2	hold	cold	told	19	pig	big	dig	36	dip	sip	hip
	fold	sold	gold		wig	rig	fig		tip	lip	rip
3	pat	pad	pan	20	pale	pace	page	37	kill	kin	kit
	path	pack	pass		pane	pay	pave		kick	king	kid
4	lane	lay	late	21	cane	case	cape	38	hang	sang	bang
	lake	lace	lame		cake	came	cave		rang	fang	gang
5	kit	bit	fit	22	shop	mop	cop	39	took	cook	look
	hit	wit	sit		top	hop	pop		hook	shook	book
6	must	bust	gust	23	coil	oil	soil	40	mass	math	map
	rust	dust	just		toil	boil	foil		mat	man	mad
7	teak	team	teal	24	tan	tang	tap	41	ray	raze	rate
	teach	tear	tease		tack	tam	tab		rave	rake	race
8	din	dill	dim	25	fit	fib	fizz	42	save	same	sale
	dig	dip	did		fill	fig	fin		sane	sake	safe
9	bed	led	fed	26	same	name	game	43	fill	kill	will
	red	wed	shed		lame	came	fame		hill	till	bill
10	pin	sin	tin	27	peel	reel	feel	44	sill	sick	sip
	fin	din	win		eel	keel	heel		sing	sit	sin
11	dug	dung	duck	28	hark	dark	mark	45	bale	gale	sale
	dud	dub	dun		bark	park	lark		tale	pale	male
12	sum	sun	sung	29	heave	hear	heat	46	wick	sick	kick
	sup	sub	sud		heal	heap	heath		lick	pick	tick
13	seep	seen	seethe	30	cup	cut	cud	47	peace	peas	peak
	seek	seem	seed		cuff	cuss	cub		peach	peat	peal
14	not	tot	got	31	thaw	law	raw	48	bun	bus	but
	pot	hot	lot		paw	jaw	saw		bug	buck	buff
15	vest	test	rest	32	pen	hen	men	49	sag	sat	sass
	best	west	nest		then	den	ten		sack	sad	sap
16	pig	pill	pin	33	puff	puck	pub	50	fun	sun	bun
	pip	pit	pick		pus	pup	pun		gun	run	nun
17	back	bath	bad	34	bean	beach	beat				
	bass	bat	ban		beak	bead	beam				

Figure 1: List of House, et al. (1965) MRT Words

Procedures

Before beginning the actual tests all participants were thoroughly trained as both speakers and listeners. Only trained participants who met the 90% speech intelligibility (SI) criteria as previously described were selected to participate in the actual tests. The higher passing score was required because training was conducted in a controlled environment with speakers and listeners all in the same room, thus higher scores would be expected in training than would be expected in the dynamic test environment. As per ANSI S3.2-1989 participants' native language had to be English.

Training

All of the training was accomplished in the same room. The room selected for training was a quiet room where all extraneous noises (i.e., air conditioning, conversation, aircraft noises) were controlled or eliminated. The room was maintained at a comfortable temperature and had an adequate level of lighting with sufficient seating for all participants. All participants were seated at an oblong table with chairs along each side and at one end. Each participant was assigned a number and a seat and given a package of test materials and a pencil. The participants were encouraged to read through all 300 words to gain familiarity with the words and to ensure that they knew how to correctly pronounce each word. Any questions as to the correct pronunciation of the words were answered at that time.

During training the speakers performed their duties by standing at the head of the table. Participant number one served as the first speaker, participant number two served as the second speaker, and so on until all participants had verbalized

their word list. After a participant performed the duties of a speaker they took a seat at the table and served as a listener. After each speaker completed their assigned 50-word set they returned to the table and all the participants then moved over one seat in a circular manner around the table. This emulated the different crew work stations in the MCE and LRE and ensured each participant sat as far away from the speaker as possible and was positioned such that both the left and right ears were directed away from and towards the speaker during some portion of the training session.

Speakers were instructed to clearly verbalize each word with the following carrier sentence; “Number blank, mark the word _____ now”, where the word “blank” was replaced with the number of the set and the “blank” after the word “word” was replaced with the actual stimulus word. Similar to the study conducted by Kryter and Whitman (1965) this carrier sentence was used because it contained both a neutral vowel preceding the stimulus word and a voiceless stop plosive after the stimulus word. Speakers were instructed that the carrier sentence and the stimulus word must be said together without giving undue emphasis on the stimulus word. Speakers were also instructed to leave a 3-second delay between each sentence by stating the following phrase quietly in their mind between each sentence: “One one hundred, two one hundred, three one hundred” or “One Mississippi, Two Mississippi, Three Mississippi”, to ensure the 3-second delay.

Listeners were instructed to circle the word in each ensemble that they thought they heard the speaker verbalize. The listeners were instructed to circle a word for each ensemble; if they were unsure which word they heard then they were

instructed to guess. If a listener wanted to change their answer they were instructed to place an “X” over the first word circled and then circle the correct word. Using this correction method actually gives the test proctor additional information. If an excessive number of corrections have been made it can indicate that a participant was unsure of their answers, having difficulty discerning the correct word – which could be an indication of a degraded communication system – or that they were changing their answers after looking at a fellow participant’s answer sheet. Listeners were instructed not to look at the speaker (to ensure they did not engage in “lip reading”) and to keep their eyes on their own papers. Listeners were instructed as to the proper flow of the answer sheets; down the page, then across to the second column and down, and then across to the third column and down.

Since this was an operational test conducted on a test aircraft using operational aircrew, it was intended that the results will be generalized to the entire Air Force population of UAV operators. Since the entire UAV operator population hails from all across the United States some speakers may have regional dialects. During the training all listeners were exposed to all speakers who verbalized all 300 words during training. This ensured that all listeners were familiar with any idiosyncratic pronunciations or regional dialects that any speakers may have had. Any speaker who did not receive a score of at least 90% correct responses from all listeners was disqualified as a speaker. Any listener who did not score at least 90% accuracy across all speakers was disqualified as a listener.

Two training sessions for the MRT participants were conducted to complete the four specific SI test objectives. The first training session was conducted prior to

the first mission. Fourteen volunteers were trained. The second training session was conducted prior to the second mission. Fifteen volunteers were trained. Of the 15 volunteers trained for the second mission, eight were already trained from the first mission but due to the seven new participants it was necessary to conduct another training session. It is an ANSI S3.2-1989 requirement to have all the participants trained together to ensure that all participants have the opportunity to hear each other and to disqualify anyone who has a regional dialect/minor speech impediment that the other potential participants cannot understand. Ten participants are standard for an MRT with at least 40% of participants being female per technical information memorandum *Speech Intelligibility Evaluation in Aircraft Test and Evaluation: The Modified Rhyme Test* (Dunavold, Cole & Thomas, unpublished). The extra participants were trained for backup purposes. However, during the second mission, only nine participants were available due to scheduling constraints.

During training, participants were given enough time to familiarize themselves with all 300 words that made up the MRT word lists. The list of potential MRT/rMRT words is presented in appendix C. A list of 50 words with the carrier sentence was spoken by each speaker in the same manner used in actual test but without the radio. The answer sheets were scored in order to ensure that each potential speaker could enunciate clearly enough that all the potential listeners could understand them. All participants passed the baseline training with an MRT score of 90% or greater. This process eliminated any potential speakers who may have had minor speech impediments or pronounced regional accents.

Testing

The MRT required that participants be in crew positions in the MCE and at an external location that represented an external user such as an ATC center. To avoid disruption of real-world operations and to acquire sufficient data, participants at the external location were not positioned in an operational ATC environment. The external user communication took place in a mobile recreational vehicle (RV), equipped with a radio and five communication stations, which was used to represent an external location such as an ATC center.

During testing, Participants 1 through 5 were in the MCE, and Participants 6 through 10 were in the RV. All participants served as both speakers and listeners. The participants took turns being speakers. Speakers performed in numerical order using correspondingly numbered word lists. The participants in the MCE spoke from their word lists first to exercise the command link (CL), followed by the participants in the RV who spoke from their word lists to exercise the return link (RL).

Each speaker began by keying the radio and instructing the listeners, “I will begin speaking from word list number _____ in 10 seconds,” where the blank was filled in with the number (1 through 10) of the word list being used. The stimulus words were verbalized in a carrier sentence with a 3-second pause between each carrier sentence, until a total of 50 words were completed. For each stimulus word, the speaker used the carrier sentence: “Number ____. Mark the word _____ now.” The first blank was replaced with the word number and the second blank was replaced with the stimulus word to be marked. The stimulus word was always highlighted in yellow on the speaker’s answer key. To prevent the radio from

overheating, the speaker keyed the radio off between each carrier sentence. This was repeated until the each speaker's 50-word set was completed.

While the speaker verbalized each word from the list, the listeners circled one of six multiple choice words on the answer sheets. A sample answer key and answer sheet are presented in appendix D and E. The listeners were instructed to guess an answer if they were unsure of the word verbalized to avoid leaving any answers blank. The formula used for MRT scoring adjusted for chance/guessing.

During testing a trained proctor was at each location for the duration of the tests. The proctors oversaw the testing and ensured that the tests were conducted in accordance with ANSI S3.2-1989, controlled the flow of the tests, ensured the proper spacing between sentences was adhered to, and coordinated breaks for the participants. To coordinate bathroom breaks, snack breaks, and start and stop times the proctors maintained contact with each other by cell phone which did not interfere with the test frequencies.

Analysis Method

Each answer sheet was scored using the standard ANSI S3.2-1989 formula to determine the number of correct answers. Although all participants who were not speaking were instructed to circle one of six multiple choice words on their answer sheets, only the answer sheets of the participants in the RV were used to calculate the CL scores; and only the answer sheets of the participants in the MCE were used to calculate the RL scores. The purpose of having the other participants mark their answers while in the same location as the speaker was to keep all participants engaged during the test.

The formula used to calculate the scores was based on the standard formula presented in ANSI S3.2-1989. The following formula was applied to determine each participant's score, represented as a percentage:

$$R_A = R - (W / (n-1)) \times (100 / \text{number of test words})$$

Where R_A is the number of items correct adjusted for chance/guessing, R is the number of items correct, W is the number of items incorrect, and *n* is the number of alternative choices per item. (American National Standards Institute, 1989) So the MRT score = number correct – (number incorrect ÷ (*n*-1)) × (100 ÷ number of test words) where *n* = the number of words in a word ensemble (*n*=6). Using *n*-1 in the formula adjusted for chance/guessing on the part of the participant.

The MRT score for each mode was an average of all participant's scores during that mode. The overall MRT score was an average of the four aggregate scores from each mode tested. For this particular MRT the scores for each mode and the overall assessment were evaluated based on the listeners' MRT scores per the following criteria: Satisfactory if MRT score ≥ 80%, Marginal if 70% ≥ MRT score < 80%, and Unsatisfactory if MRT score < 70%.

Methods and Conditions – Mission 1

The first two objectives, to determine the SI of non-secure (NS) Very High Frequency (VHF) Amplitude Modulation (AM) mode and the non-secure (NS) Ultra High Frequency (UHF) AM mode, were accomplished during the first mission on October 30, 2008. The flight profile, AV configuration, MCE configuration, and the RV configuration were set up as indicated under Overall Methods and Conditions.

Five participants, two of whom were female, were stationed in the RV and the other five participants, two of whom were female, were stationed in the MCE.

Methods and Conditions – Mission 2

The third and fourth objectives, to determine the SI of the secure (S) UHF Frequency Modulation (FM) mode and the secure UHF AM baseband mode (BB), were accomplished during the second mission on November 24, 2008. The flight profile, AV configuration, MCE configuration, and the RV configuration were set up as indicated under Overall Methods and Conditions. Only nine MRT participants were available for this mission. Five participants, two of whom were female, were stationed in the RV. The four remaining participants, two of whom were female, were stationed in the MCE. Because we were short one participant, Male Participant No. 1 in the MCE spoke twice, once each from two different word lists and two different radio stations – word list No. 1 from radio station No. 1 and word list No. 5 from radio station No. 5. The data were included in the calculations for the CL and the overall results. However, since there were only four participants in the MCE, there were only four sets of data used to calculate the RL. This did not significantly affect the results, as the lowest score on any given test was 83%, which was above the minimum required satisfactory score of 80%.

Results – Missions 1 and 2

The overall speech intelligibility of the Global Hawk communication system using audio pass-through via Ku SATCOM from the MCE to the air traffic control AN/ARC-210 radio in the GH AV and subsequently line-of-sight transmission to the external radio and vice versa was satisfactory. The overall SI score, across all

frequency modes, was 84%, which met the Airworthiness Certification Criteria (ACC) requirement of at least 80%.

However, it remains to be seen if the SI score of the reduced MRT (rMRT) consisting of less than 50 responses equals the SI score of the full MRT.

rMRT Analysis

Design

To determine if the SI score of an rMRT is statistically equivalent to the SI score of a full MRT an analysis was performed using multiple sets of seven different subsets of the data that were collected during the GH UAV MRTs. Each data subset contained several randomly selected collections of 15, 20, 25, 30, 35, 40, and 45 words from each individual participant. The collections are not all independent; that is, each collection of *n* words was a subset of the responses from the full MRT (50 words) for each participant. The research question to be answered was, "How large a subset of the full MRT must be used to get an SI score that is statistically equivalent to the SI score of the full MRT?" The analysis was performed to test the following hypotheses:

- H_o 1: The SI score of a 45-word rMRT set = SI score of the Full MRT set
- H_a 1: The SI score of a 45-word rMRT set ≠ SI score of the Full MRT set
- H_o 2: The SI score of a 40-word rMRT set = SI score of the Full MRT set
- H_a 2: The SI score of a 40-word rMRT set ≠ SI score of the Full MRT set
- H_o 3: The SI score of a 35-word rMRT set = SI score of the Full MRT set
- H_a 3: The SI score of a 35-word rMRT set ≠ SI score of the Full MRT set
- H_o 4: The SI score of a 30-word rMRT set = SI score of the Full MRT set
- H_a 4: The SI score of a 30-word rMRT set ≠ SI score of the Full MRT set
- H_o 5: The SI score of a 25-word rMRT set = SI score of the Full MRT set
- H_a 5: The SI score of a 25-word rMRT set ≠ SI score of the Full MRT set
- H_o 6: The SI score of a 20-word rMRT set = SI score of the Full MRT set
- H_a 6: The SI score of a 20-word rMRT set ≠ SI score of the Full MRT set

- H_o 7: The SI score of a 15-word rMRT set = SI score of the Full MRT set
- H_a 7: The SI score of a 15-word rMRT set ≠ SI score of the Full MRT set

The subsets were grouped as shown in table 1.

Table 1 50 Word Sets Divided Into Reduced Word Sets

rMRT	rMRT	rMRT	rMRT	rMRT	rMRT	rMRT	MRT
15	15	15	15	15	15	15	15
20	20	20	20	20	20	20	20
25	25	25	25	25	25	25	25
30	30	30	30	30	30	30	30
35	35	35	35	35	35	35	35
40	40	40	40	40	40	40	40
45	45	45	45	45	45	45	45
50	50	50	50	50	50	50	50
SI of 15 Words	SI of 20 Words	SI of 25 Words	SI of 30 Words	SI of 35 Words	SI of 40 Words	SI of 45 Words	SI of Full 50 Words

For all analyses only the raw MRT scores from the 2008 GH UAV tests were used. A bootstrap procedure (Hastie, Tibshirani, & Friedman, 2009) was applied using the code presented in appendix F. Within each frequency mode, i.e., non-secure UHF AM, non-secure VHF AM, secure UHF FM and secure UHF AM BB, and within each participant's response set for that particular frequency mode, responses of various size *n* (15, 20, 25, 30, 35, 40, 45) were randomly selected from each participant's set of 50 responses and analyzed for score. This was repeated 100 times for each participant and for each *n* resulting in a total of 700 responses for each participant for each of the four frequency modes. These individual scores were then compared with the scores of 50 words for each participant for each frequency mode.

For example, for the non-secure UHF AM frequency mode, the scores that resulted from the 100 random samples of 15 words for a particular participant were compared to the score for the 50-word set for that participant. Then the scores that resulted from that participant's 100 random samples of 20 words were compared to the score for the 50-word set for that participant. The scores that resulted from the 100 random samples of 25 words were compared to the score for the 50-word set for that participant. The scores that resulted from the 100 random samples of 30 words were compared to the score for the 50-word set for that participant. The scores that resulted from the 100 random samples of 35 words were compared to the score for the 50-word set for that participant. The scores that resulted from the 100 random samples of 40 words were compared to the score for the 50-word set for that participant. And finally the scores that resulted from the 100 random samples of 45 words were compared to the score for the 50-word set for that participant. This was repeated for each participant for each subset of *n* words for each frequency.

A *z*-test statistic comparing the proportions correct (between the full sample of 50 and the subsample of size *n*) was computed and recorded. The *z*-values were sorted and the *p*-value for accuracy of at least 98% was found. If that *p*-value was less than 0.05, the null hypothesis of "no difference" in proportion correct was rejected for that particular value of *n*. Then confidence intervals for the *p*-values were estimated from the distribution of *p*-values for each condition.

Results

There was no significant difference between any of the *n*-word SI scores and the SI scores of the original 50-word analysis, when *n* equaled 30, 35, 40, and 45 for all frequency modes. Figures 2, 3, 4, and 5 show the p-values for each hypothesis test, and their associated 95% confidence intervals. In all frequency modes $p > .06$ when *n* equaled 30. Therefore we fail to reject the null hypotheses, H_o 1, 2, 3, and 4, and the results do not appear to support the alternative hypotheses, H_a 1, 2, 3, and 4.

Additionally, there was a significant difference between the SI scores with an *n* of 15 and the SI scores of the original 50-word MRT for all frequency modes. Therefore the null hypothesis, H_o 7, was rejected and the results do not appear to support the alternative hypothesis, H_a 7, (non-secure VHF AM, $p < .05$; non-secure UHF AM, $p < .04$; secure UHF AM BB, $p < .04$; and secure UHF FM, $p < .04$).

However, the analysis revealed that there were minor differences in the scores of the different frequency modes when *n* was less than 30 but greater than 15. For example, for the non-secure VHF AM, secure UHF AM BB, and secure UHF FM modes only 25 words were required to achieve the same results as a full 50-word MRT with 98% accuracy (see figures 2, 4, and 5). Therefore we fail to reject the null hypothesis, H_o 5, for three frequency modes (non-secure VHF AM, $p > .06$; secure UHF AM BB, $p > .05$; and secure UHF FM, $p > .05$) and the results do not appear to support the alternative hypothesis, H_a 5, for one frequency mode (non-secure UHF AM, $p < .05$). Additionally, one frequency mode, non-secure VHF AM, demonstrated that 20 words would be sufficient to achieve the same results as a full 50-word MRT with 98% accuracy (see Figure 2, $p = .06$). However, for three frequency modes,

non-secure UHF AM, secure UHF AM BB; and secure UHF FM, 20 words were insufficient to achieve the same results as a full 50-word MRT. Therefore we fail to reject the null hypothesis, H_o 6, for one frequency mode (non-secure VHF AM, $p = .06$) and the results do not appear to support the alternative hypothesis, H_a 6, for three frequency modes (non-secure UHF AM, $p < .05$; secure UHF AM BB, $p < .05$; and secure UHF FM, $p < .05$).

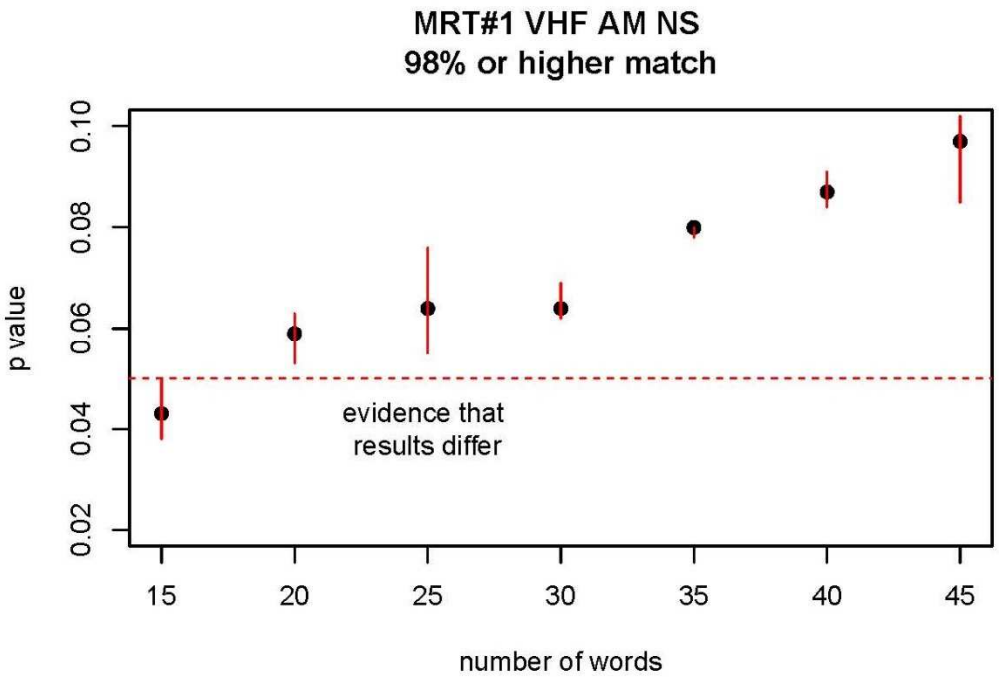

Figure 2: MRT 1 – Non Secure VHF AM with 95% Confidence Intervals

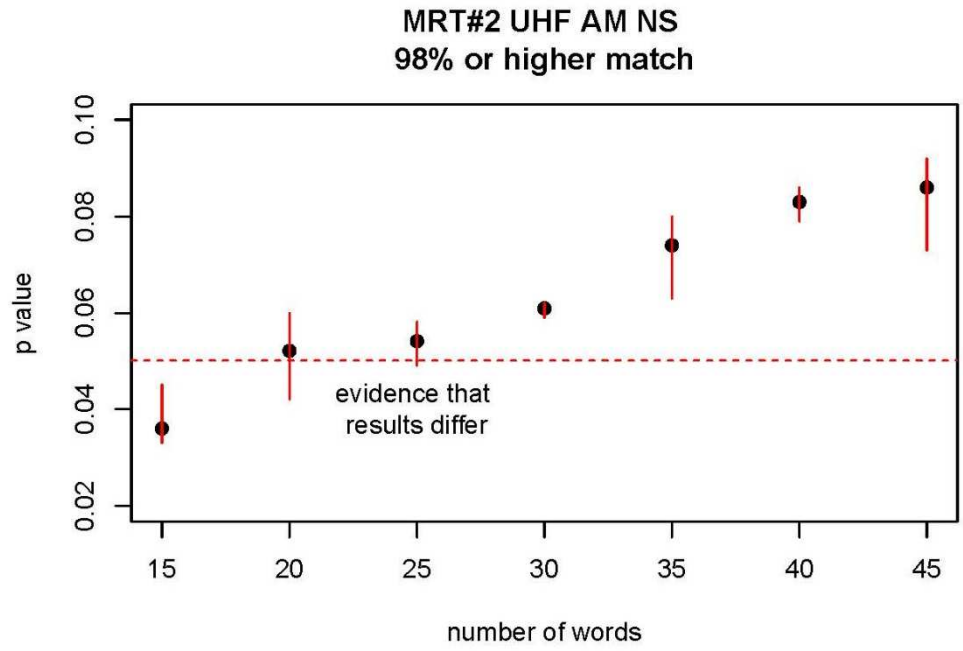

Figure 3: MRT 2 – Non Secure UHF AM with 95% Confidence Intervals

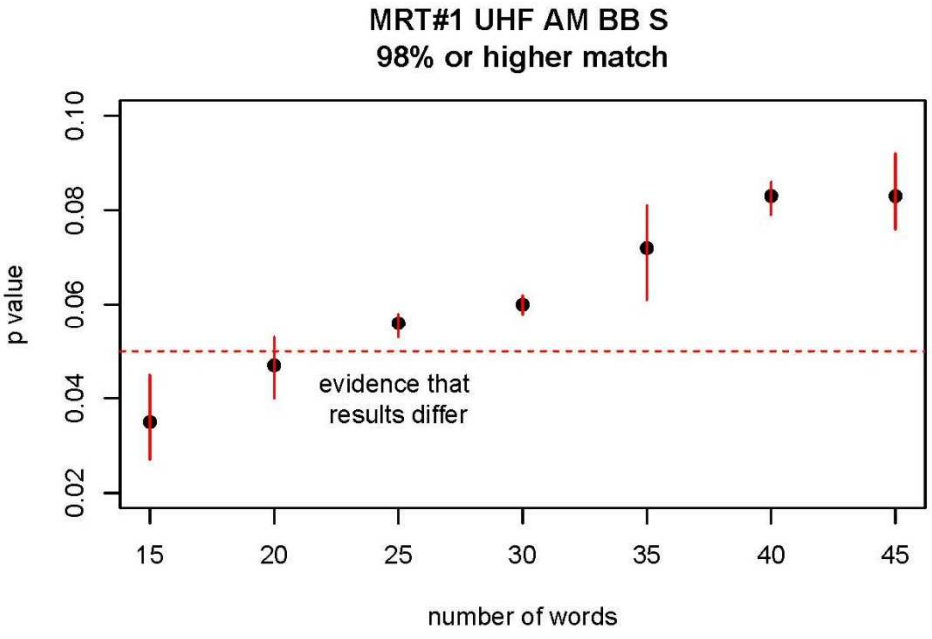

Figure 4: MRT 1 – Secure UHF AM BB with 95% Confidence Intervals

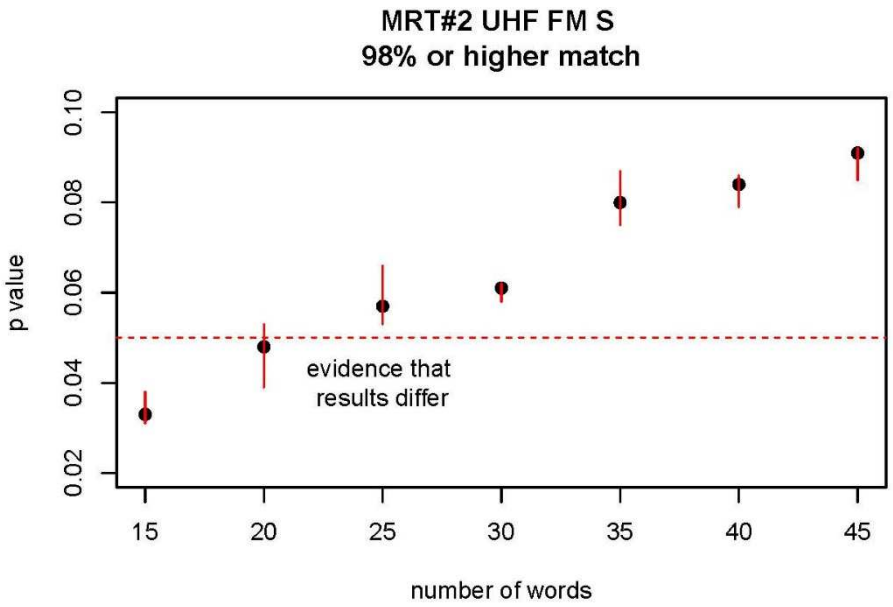

Figure 5: MRT 2 – Secure UHF FM with 95% Confidence Intervals

Therefore the analysis revealed that, across all frequencies, using an rMRT of 30 words would result in a 98% accuracy rate. However, it was found that it would require an *n* of greater than 45 words to achieve an accuracy rate of 99%.

Figures 2, 3, 4, and 5 show that for most frequency modes adding more than 30 words did not result in statistically significant improvements in the SI scores. The percentage of time that the SI score of less than 50 words matched the SI score of 50 words remained at 98% for 30, 35, 40 and 45 words. It was also found that using more than 45 words, but less than 50 words, only resulted in a gain of about 1%. Therefore the added benefit of increasing the number of words to 45 or more was not worth the additional cost.

Based on the results of this study it can be concluded that using an rMRT of 30 words would achieve the same results as a full 50-word MRT 98% of the time

and allow the USAF to save approximately 39% in overall test costs (Cole, 2007) with a relatively low risk of accepting a Type II error.

A split-half reliability analysis was also run on the same data to determine the internal reliability of the MRT method. For every participant each set of 50 responses was randomly split into two data sets and was analyzed for score. This resulted in 38 x 2 (76) sets of data. In all instances the two scores were highly correlated ($p < .02$) which indicated that the data and the MRT method had good internal reliability (see Figure 6).

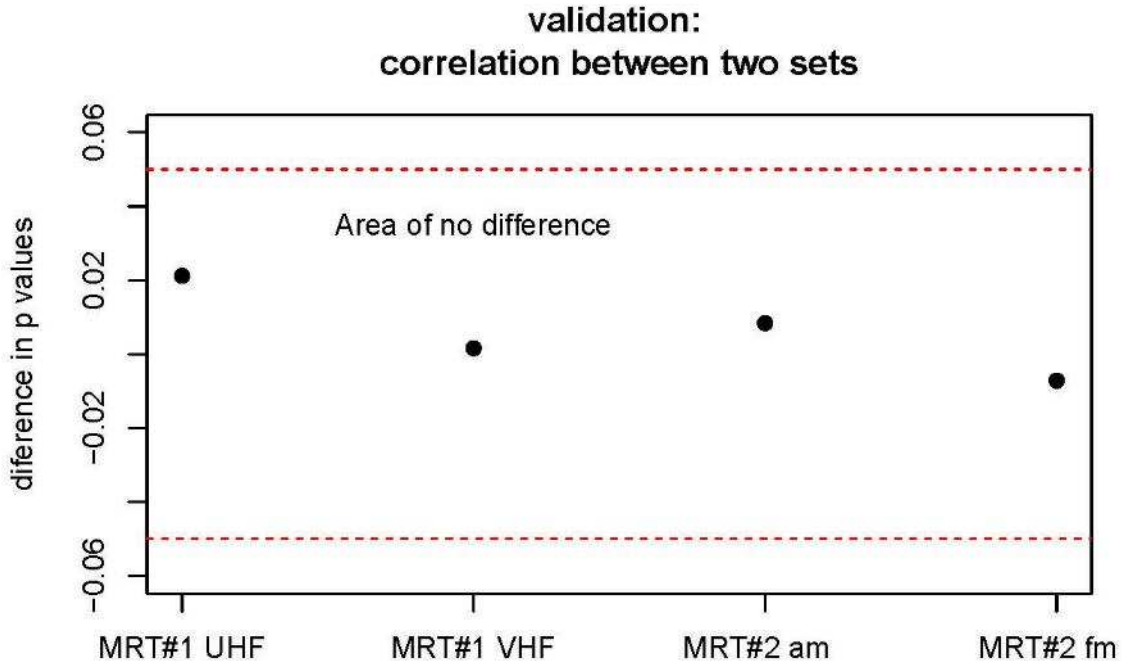

Figure 6: MRT 2 – Split-Half Reliability

Discussion

Many of the early research studies conducted on SI used pre-recorded word lists. Pre-recorded word lists could have been used for the GH MRT; however, studies have shown that the most significant source of variability is in the speaker's voices (House, et al., 1965; Erickson & McKinley, 1997). Therefore, for optimum results, it would be preferable to have recordings with several different speakers, both male and female. Also, using electronic equipment to play back the recordings would insert an additional component into the tests which would increase the complexity of test administration and require additional equipment. Additionally, the electrical characteristics of the recording and playback equipment could introduce distortion into the speech signal which may be heard by the listeners. For these reasons, and to be as operationally representative as possible, the MRTs conducted by the USAF do not use pre-recorded tapes of the word lists.

Another requirement when testing military systems and subsystems is that there is no degradation or loss of capability when fielding new versions of a system or subsystem. Therefore, regardless of what the required SI score is, if a voice communications system has an SI score that is greater than the requirement, then a newer or replacement voice communications system for that platform must meet or exceed the previous score to be acceptable. For example, for the 2008 MRT that was accomplished on the RQ-4 Global Hawk, the required SI score was 80%. But the 2008 MRT resulted in an overall SI score of 84%. That score of 84% is now considered the baseline score and any future tests of a voice communications system on that platform will be required to meet or exceed 84%.

Another area of concern when conducting lengthy MRTs is the stamina of the participants. Research has shown that long, boring events can tax the vigilance and focus of individuals and can even contribute to a higher number of mistakes and accidents (Krueger, 1989; Lerman, et al., 2012). Additionally, according to House, et al. (1965) "the number 50 was used in deference to traditional usage and may not represent an essential characteristic of the materials" (p.159). Therefore there does not appear to be any strong evidence that 50 words is the optimal number of words to use for an SI evaluation. Reducing the number of word ensembles would help alleviate concerns that the participants make more mistakes as they became fatigued. Based on the results of this analysis it is recommended that further analysis be conducted to determine if the number of errors made by the listeners is greater in the final 10- to 20-word ensembles of each speaker and to determine if errors increase as the test progresses and the participants become more fatigued and possibly less vigilant.

Intelligible communication is an important condition in many aviation-related tasks. Commercial pilots must have the means to successfully communicate with a variety of other persons, some of whom are on the ground, others who are in the same aircraft, and still others who are in different aircraft flying in the same airspace. Military pilots have an even more critical need for intelligible communication. They not only need to be able to communicate clearly with air traffic control on the ground and their own aircrew on their aircraft, they must also be able to communicate with the aircrew of other planes flying in close military formations, command and control (C2) personnel at ground-based command centers, and ground-based military

personnel near the site of air strike targets (Ericson, Brungart, & Simpson, 2005). The ability to clearly hear and understand speech can result in a safe and successful mission just as the inability to clearly hear and understand can result in a failed mission. And failed missions can have catastrophic consequences such as the loss of the crew, the loss of expensive aircraft, or destroying the wrong target and seriously injuring or even killing civilian bystanders, a situation commonly known as collateral damage.

Based on the results of this study a 30-word rMRT can be used in place of a full MRT and yield statistically equivalent results. This can result in an approximately 39% savings in time to administer the SI tests (Cole, 2007). This translates into a significant savings in terms of flight time and fuel, personnel time and costs, and reduces the burden on the participants of communications tests. In today's fiscally strapped world, saving money is an important goal for everyone, even the USAF.

As long as stringent measures are maintained in the selection, training, and administration of communication tests, this study indicated that an rMRT of 30 words could be used in future operational tests of RQ-4 Global Hawk drones to accurately determine the SI of new or improved communications systems.

However, it remains to be determined if an rMRT would also yield statistically equivalent results to a full 50-word MRT on all military platforms. The required SI actually varies with the platform. Different aircraft platforms move at different speeds and have different missions. Larger military cargo aircraft, like C-130s, are relatively slow moving. Additionally, the nature of cargo aircraft missions are generally less risky and time dependent than faster moving aircraft, like fighters or

bombers. Therefore the SI for a faster aircraft is generally higher than for the slower aircraft like the RQ-4 Global Hawk drone. Bombers like the B-2 travel at a much higher rate of speed and have inherently risky missions that require both accuracy and a critical time component. The risk of a failed mission and possibly grave consequences is high with these platforms so the SI requirement is generally higher. (R. L. McKinley, personal communication, April 19, 2001)

Therefore, it is recommended that additional analyses be conducted on the different military platforms to see if 30-word rMRTs would also result in an SI score that is equal to the SI score of a full 50-word MRT for the different military platforms (i.e., bombers, fighters, and cargo aircraft).

References

American National Standards Institute (ANSI) (1989). ANSI S3.2-1989 *Method for measuring the intelligibility of speech over communications systems.* New York: American National Standards Institute.

ASC/EN Airworthiness Certification Criteria Expanded Version of MIL-HSBK-516B, Aeronautical Systems Center, Wright-Patterson AFB, Ohio, 26 September 2005.

Bell Telephone Magazine. (1947). *Two Men and a Piece of Wire – and Faith*, Volume XXVI, 52.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25,* 975-979.

Cole, D. S. (2007). In Search of Alternative Modified Rhyme Tests to Reduce Test Duration. Master's Thesis, California State University, Northridge.

Dunavold, P. A., & Herrera, C. (2009). AFFTC-TR-09-02, RQ-4 Global Hawk Block 20/30 Speech Intelligibility Test and Evaluation. Air Force Flight Test Center, Edwards Air Force Base, CA, May 2009.

Dunavold, P. A., Cole, D. S., and Thomas, D. C., *Technical Information Memorandum, Speech Intelligibility Rhyme Evaluation in Aircraft Test and Evaluation: The Modified Rhyme Test,* Air Force Flight Test Center, Edwards AFB, California, unpublished

ICAO (2007). *Manual on Radiotelephony*, Document 9432-AN/925, International Civil Aviation Organization, Montreal, Canada.

Ericson, M. A., Brungart, D. S., & Simpson, B. D. (2005). Factors that influence intelligibility in multitalker speech displays. *The International Journal of Aviation Psychology*, 14(3), 313-334.

Ericson, M. A., & McKinley, R. L. (1997). The intelligibility of multiple talkers separated spatially in noise. *Binaural and Spatial Hearing in Real and Virtual Environments* (pp. 701-724). Mahwah, NJ, Lawrence Erlbaum Associates.

Fairbanks, G. (1958). Test of phonemic differentiation: the rhyme test. *The Journal of the Acoustical Society of America, 30(7),* 596-600.

Griffiths, J. D. (1967). Rhyming minimal contrasts: a simplified diagnostic articulation test. *The Journal of the Acoustical Society of America, 42(1),* 236-241.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning Data Mining, Inference, and Prediction. (2nd Edition) Springer.

House, A. S., & Stevens, K. N. (1955). Auditory testing of a simplified description of vowel articulation. *The Journal of the Acoustical Society of America, 27(5),* 882-887.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America, 37(1),* 158-166.

Kreul, E. J., Nixon, J. C., Kryter, K. D., Bell, D. W., Lang, J. S.,& Schubert, E. D. (1968). A proposed clinical test of speech discrimination. *Journal of Speech and Hearing Research, 11,* 536-552.

Krueger, G. P., (1989). Sustained Work, Fatigue, Sleep Loss and Performance: A Review of the Issues. USAARL Report No. 89-22, U.S. Army Aeromedical Research Laboratory, Ft Rucker AL.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation Index. *The Journal of the Acoustical Society of America, 34(11),* 1689-1697.

Kryter, K. D., Licklider, J. C. R., & Stevens, S. S. (1947). Premodulation clipping in AM voice communication. *The Journal of the Acoustical Society of America, 19(1),* 125-131.

Kryter, K. D., & Whitman, E. C. (1965). Some comparisons between rhyme and pb-word intelligibility tests. *The Journal of the Acoustical Society of America, 37(6),* 1146.

Lerman, S. E., Eskin, E., Flower, D. J., George, E. C., Gerson, B. Hartenbaum, N., Hursh, S.R., Moore-Ede, M. (2012). *The Journal of Occupational and Environmental Medicine*, 54(2), 231-258.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin, 44(2),* 105-129.

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, 3*29-335.

Miller, G. A., & Mitchell, S. (1947). Effects of distortion on the intelligibility of speech at high altitudes. *The Journal of the Acoustical Society of America, 19(1),* 120-125.

Miller, G. A., & Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27(2),* 338-352.

Neary, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America, 85(5),* 2088-2113.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24,* 175-184.

Rosenzweig, M. R., & Postman, L. (1958). Frequency of usage and the perception of words. *Science, 127(3293),* 263-266.

Stevens, K. N., & House, A. S. (1955) Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America, 27(3),* 484-493.

Steeneken, H. J. M. (2003). *The measurement of speech intelligibility.* Soesterberg, Netherlands: TNO Human Factors.

Thompson, M. (2013). Costly Flight Hours. Time. Electronic version retrieved March 17, 2016 from: http://nation.time.com/2013/04/02/costly-flight-hours/

Williams, C. E., Mosko, J. D., & Green, J. W (1976). Evaluating the ability of aircrew personnel to hear speech in their operational environments. *Aviation, Space, and Environmental Medicine*, 47(7), 239-244.

Appendix A: – ICAO Standards

The ICAO Standard Alphabet (ICAO Annex 10)

A	alfa	H	hotel	O	oscar	V	victor
B	bravo	I	india	P	papa	W	whiskey
C	charlie	J	Juliet	Q	quebec	X	x-ray
D	delta	K	kilo	R	romeo	Y	yankee
E	echo	L	lima	S	sierra	Z	zulu
F	fox trot	M	mike	T	tango		
G	golf	N	november	U	uniform		

ICAO Manual of Radiotelephony

Letter	Word	Pronunciation
A	Alpha	**AL** FAH
B	Bravo	**BRA**H **VOH**
C	Charlie	**CHAR** LEE or **SHAR** LEE
D	Delta	**DELL** TAH
E	Echo	**ECK** OH
F	Foxtrot	**FOKS** TROT
G	Golf	GOLF
H	Hotel	HOH **TELL**
I	India	**IN** DEE AH
J	Juliet	**JEW** LEE ETT
K	Kilo	**KEY** LOH
L	Lima	**LEE** MAH
M	Mike	MIKE
N	November	NO **VEM** BER
O	Oscar	**OSS** CAH
P	Papa	PAH **PAH**
Q	Quebec	KEH BECK
R	Romeo	**ROW** ME OH
S	Sierra	SEE **AIRRAH**
T	Tango	**TAN** GO
U	Uniform	**YOU** NEE FORM or **OO** NEE FORM
V	Victor	**VIK** TAH
W	Whiskey	**WISS** KEY
X	X-ray	**ECKS** RAY
Y	Yankee	**YANG** KEY
Z	Zulu	**ZOO** LOO

Appendix A: – ICAO Standards, cont.

Transmission of Numbers

Numeral or numeral elements	Pronunciation
0	ZE-RO
1	WUN
2	TOO
3	TREE
4	FOW-er
5	FIFE
6	SIX
7	SEV-en
8	AIT
9	NIN-ER
Decimal	DAY - SEE – MAL
Thousand	TOUSAND

Note: The syllables printed in capital letters in the above list are to be stressed: for example, the two syllables in ZE-RO are given equal emphasis, whereas the first syllable of FOW-er is given primary emphasis.

All numbers except whole thousands shall be transmitted by pronouncing each digit separately. Whole thousands shall be transmitted by pronouncing each digit in the number of thousands followed by the word THOUSAND.

Number	Transmitted as	Pronounced as
10	ONE ZERO	WUN ZE-RO
75	SEVEN FIVE	SEV-en FIVE
100	ONE ZERO ZERO	WUN ZE-RO ZE-RO
583	FIVE EIGHT THREE	FIFE AIT TREE
2 500	TWO FIVE ZERO ZERO	TOO FIFE ZE-RO ZE-RO
5 000	FIVE THOUSAND	FIFE TOUSAND
11 000	ONE ONE THOUSAND	WUN WUN TOUSAND
25 000	TWO FIVE THOUSAND	TOO FIFE TOUSAND
38 143	THREE EIGHT ONE FOUR THREE	TREE AIT WUN FOW-er TREE

Appendix A: – ICAO Standards, cont.

Numbers containing a decimal point shall be transmitted with the decimal point in appropriate sequence being indicated by the word DECIMAL.

Number	Transmitted as	Pronounced as
118.1	ONE ONE EIGHT DECIMAL ONE	WUN WUN AIT DAY SEE-MAL WUN
120.37	ONE TWO ZERO DECIMAL THREE SEVEN	WUN TOO ZE-RO DAY SEE-MAL TREE SEV-en

When transmitting time, only the minutes of the hour are normally required. However, the hour should be included if there is any possibility of confusion. Coordinated universal time (UTC) shall be used.

Appendix B: – Informed Consent

Informed Consent

We would like to use the data collected during three Global Hawk Modified Rhyme Tests in which you participated to complete a thesis which will be submitted as partial fulfillment of the requirements for a Master of Science degree in Psychology through the University of Idaho. The purpose of this study is to determine if the Speech Intelligibility (SI) scores of a Reduced Modified Rhyme Test (rMRT) are statistically equivalent to the SI scores of a full Modified Rhyme Test (MRT). This will be accomplished by re-analyzing the data collected from the original Global Hawk MRT done in October and November 2008. Your participation in this study is entirely voluntary. If you decide to participate in this study you will not be required to do anything other than sign this consent form. The data has already been collected.

There are no foreseeable risks or benefits to you from your participation in this study and your participation is completely voluntary. You will be free to refuse to allow your data to be included in this study.

All information provided by you during this study will be number coded and kept strictly confidential. Your identity will not be revealed without your written consent. The results of this study may be used at a later date for a journal article for a professional psychology journal. The purpose of the article will be to further scientific knowledge and share information with other interested persons in the field. Again, your identity will remain confidential and only general data will be used in the thesis and journal article.

Do you have any questions? If you have any questions later, please feel free to contact us.

Patricia A. Dunavold
Graduate Student
University of Idaho
Phone: (661)275-3365

Brian Dyre, PhD
Professor of Psychology
University of Idaho
Phone: (818)677-2827

Please read the following paragraph, and, if you agree to participate, please sign below.

I understand that any personal information about me obtained from this research will be kept strictly confidential. I do understand that the results of the study will be published in a thesis and may be used for a professional journal article later.

Signature _____ Date: _____

Experimenter _____ Date: _____

Please place your initials here acknowledging receipt of a copy of this consent form.

_____

Appendix C: – List of Potential MRT/rMRT Words

1	sing	sit	sin	sip	sick	sill
2	book	took	shook	cook	hook	look
3	nest	vest	west	test	best	rest
4	kith	king	kid	kit	kiss	kill
5	pun	puff	pup	pug	putt	pub
6	fill	fig	fin	fizz	fib	fit
7	foil	coil	boil	oil	toil	soil
8	bust	just	rust	must	gust	dust
9	jig	wig	big	rig	pig	fig
10	sake	sale	save	sane	safe	same
11	kit	bit	fit	sit	wit	hit
12	came	cape	cane	cake	cave	case
13	sold	told	hold	fold	gold	cold
14	map	mat	math	man	mass	mad
15	gale	male	tale	bale	sale	pale
16	raw	paw	law	jaw	thaw	saw
17	dent	bent	went	tent	rent	sent
18	page	pane	pace	pay	pale	pave
19	fame	same	came	name	tame	game
20	duck	dud	dull	dub	dug	dun
21	rave	rake	race	rate	raze	ray
22	will	hill	kill	till	fill	bill
23	pass	pat	pang	pad	path	pan
24	peel	reel	feel	heel	keel	eel
25	bun	bus	but	buff	buck	bug
26	hear	heath	heal	heave	heat	heap
27	sad	sass	sag	sack	sap	sat
28	sun	nun	gun	fun	bun	run
29	kick	lick	sick	pick	wick	tick
30	cut	cub	cuff	cup	cud	cuss
31	peace	peas	peak	peal	peat	peach
32	way	may	say	gay	day	pay
33	ten	pen	den	hen	then	men
34	meat	feat	heat	seat	beat	neat
35	sip	rip	tip	dip	hip	lip
36	dig	dip	did	dim	dill	din
37	teach	tear	tease	teal	team	teak
38	sud	sum	sub	sun	sup	sung
39	pill	pick	pip	pig	pin	pit
40	led	shed	red	bed	fed	wed
41	top	hop	pop	cop	mop	shop
42	late	lake	lay	lace	lane	lame
43	bean	beach	beat	beam	bead	beak
44	rang	fang	gang	bang	sang	hang
45	seep	seen	seethe	seed	seem	seek
46	hark	dark	mark	lark	park	bark
47	pin	sin	tin	win	din	fin
48	tab	tan	tam	tang	tack	tap
49	bat	bad	back	bass	ban	bath
50	hot	got	not	pot	lot	tot

Appendix D: – Sample Talker Sheet

"This is speaker #1. I will begin speaking from MRT #1, word list #1 in 10 seconds."

"Number _____. Mark the word _____ now."

MRT #1 Talker #1 B

1	big	wig	rig	jig	pig	fig
2	sang	bang	gang	hang	rang	fang
3	lace	late	lake	lame	lane	lay
4	dim	dig	dip	did	dill	din
5	sin	fin	pin	tin	win	din
6	game	came	name	same	fame	tame
7	feat	beat	heat	seat	meat	neat
8	sung	sub	sun	sud	sup	sum
9	sad	sat	sap	sass	sack	sag
10	red	bed	shed	fed	led	wed
11	nun	bun	fun	gun	run	sun
12	bean	beam	beat	beak	beach	bead
13	pun	pup	pug	putt	pub	puff
14	pang	pat	path	pan	pass	pad
15	cut	cub	cuff	cuss	cud	cup
16	kit	bit	hit	sit	fit	wit
17	ten	then	den	hen	men	pen
18	wick	pick	lick	tick	sick	kick
19	thaw	paw	jaw	law	raw	saw
20	fit	fig	fizz	fill	fin	fib
21	peach	peace	peas	peak	peal	peat
22	cane	came	cave	case	cape	cake
23	pill	pip	pin	pick	pit	pig
24	sit	sip	sill	sin	sing	sick
25	gust	dust	bust	rust	must	just
26	bun	bus	bug	buff	but	buck
27	hot	not	tot	lot	pot	got
28	soil	toil	oil	foil	boil	coil
29	bat	bad	bath	bass	ban	back
30	rake	ray	raze	race	rate	rave
31	pay	way	day	say	may	gay
32	bale	sale	tale	male	pale	gale
33	tang	tam	tan	tack	tap	tab
34	peel	keel	reel	heel	eel	feel
35	seem	seen	seed	seethe	seep	seek
36	cook	book	look	hook	took	shook
37	rip	sip	tip	dip	hip	lip
38	sane	same	safe	sake	save	sale
39	dub	dun	duck	dud	dull	dug
40	bill	till	will	fill	kill	hill
41	map	mass	math	mad	mat	man
42	king	kiss	kit	kith	kid	kill
43	nest	vest	west	best	rest	test
44	gold	sold	fold	cold	hold	told
45	mop	cop	hop	shop	top	pop
46	went	bent	rent	dent	tent	sent
47	heath	hear	heat	heap	heave	heal
48	bark	park	lark	hark	mark	dark
49	pave	pay	pace	pale	pane	page
50	teal	teak	tear	team	teach	tease

Appendix E: – Sample Listener Answer Sheet

Name:_____ Date:_____ Location:_____

MRT #1

1	big	wig	rig	18	wick	pick	lick	35	seem	seen	seed
	jig	pig	fig		tick	sick	kick		seethe	seep	seek
2	sang	bang	gang	19	thaw	paw	jaw	36	cook	book	look
	hang	rang	fang		law	raw	saw		hook	took	shook
3	lace	late	lake	20	fit	fig	fizz	37	rip	sip	tip
	lame	lane	lay		fill	fin	fib		dip	hip	lip
4	dim	dig	dip	21	peach	peace	peas	38	sane	same	safe
	did	dill	din		peak	peal	peat		sake	save	sale
5	sin	fin	pin	22	cane	came	cave	39	dub	dun	duck
	tin	win	din		case	cape	cake		dud	dull	dug
6	game	came	name	23	pill	pip	pin	40	bill	till	will
	same	fame	tame		pick	pit	pig		fill	kill	hill
7	feat	beat	heat	24	sit	sip	sill	41	map	mass	math
	seat	meat	neat		sin	sing	sick		mad	mat	man
8	sung	sub	sun	25	gust	dust	bust	42	king	kiss	kit
	sud	sup	sum		rust	must	just		kith	kid	kill
9	sad	sat	sap	26	bun	bus	bug	43	nest	vest	west
	sass	sack	sag		buff	but	buck		best	rest	test
10	red	bed	shed	27	hot	not	tot	44	gold	sold	fold
	fed	led	wed		lot	pot	got		cold	hold	told
11	nun	bun	fun	28	soil	toil	oil	45	mop	cop	hop
	gun	run	sun		foil	boil	coil		shop	top	pop
12	bean	beam	beat	29	bat	bad	bath	46	went	bent	rent
	beak	beach	bead		bass	ban	back		dent	tent	sent
13	pun	pup	pug	30	rake	ray	raze	47	heath	hear	heat
	putt	pub	puff		race	rate	rave		heap	heave	heal
14	pang	pat	path	31	pay	way	day	48	bark	park	lark
	pan	pass	pad		say	may	gay		hark	mark	dark
15	cut	cub	cuff	32	bale	sale	tale	49	pave	pay	pace
	cuss	cud	cup		male	pale	gale		pale	pane	page
16	kit	bit	hit	33	tang	tam	tan	50	teal	teak	tear
	sit	fit	wit		tack	tap	tab		team	teach	tease
17	ten	then	den	34	peel	keel	reel				
	hen	men	pen		heel	eel	feel				

Talker #1

Appendix F: Code for rMRT Bootstrap Analysis

```
---
title: "Speech inteligibility"
output: pdf_document
---

## Data analysis for talker study, based on bootstrap.  First get the data

```{r loadData, engine='R'}
rm(list=ls())
require(boot)
require(graphics)

#dataLocation = "C:\\Users\\Dr. J.DrJ-HP\\Documents\\R\\Work\\dunavold\\"
dataLocation = "C:\\Users\\James\\Documents\\R\\Work\\dunavold\\"
'C:\\Users\\James\\Documents\\R\\Work\\dunavold\\'
MRT1U=read.csv(paste(dataLocation,"MRT#1 UHF AM BB S.csv", sep=''), header=TRUE)
MRT1V=read.csv(paste(dataLocation,"MRT#1 VHF AM NS.csv", sep=''), header = TRUE)
MRT2Uam=read.csv(paste(dataLocation,"MRT#2 UHF AM NS.csv",sep = ''),
header=TRUE)
MRT2Ufm=read.csv(paste(dataLocation,"MRT#2 UHF FM S.csv",sep = ''), header=TRUE)

dataSetList = list(MRT1U, MRT1V, MRT2Uam, MRT2Ufm)

titlePlot = c('MRT#1 UHF AM BB S','MRT#1 VHF AM NS', 'MRT#2 UHF AM
NS','MRT#2 UHF FM S')

```
set up the problem; fill the bucket

```{r  fillBucket , echo=FALSE}

bucket = array(dim=c(50,50,4) )

rowNum = dim(MRT1U)[1]  # rows, columns
colNum = dim(MRT1U)[2]

for (i in 1:rowNum)
  for(j in 1:colNum) bucket[i,j,1]=MRT1U[i,j]

rowNum = dim(MRT1V)[1]
colNum = dim(MRT1V)[2]

for (i in 1:rowNum)
  for(j in 1:colNum) bucket[i,j,2]=MRT1V[i,j]
```
```

```
rowNum = dim(MRT2Uam)[1]
colNum= dim(MRT2Uam)[2]

for (i in 1:rowNum)
  for(j in 1:colNum) bucket[i,j,3]=MRT2Uam[i,j]


rowNum = dim(MRT2Ufm)[1]
colNum = dim(MRT2Ufm)[2]

for (i in 1:rowNum)
  for(j in 1:colNum) bucket[i,j,4]=MRT2Ufm[i,j]

columnsCount = c(45,50,50,45)  # columns in data

```


```{r  runBootstrap}
nBoots = 100
nRuns=100
rowNum = 50

for(k in 1:4) {   # bucket number

  wordsNumber = 1 # 1 to 7
  plotSave=vector('numeric', 7)
  zUpper = vector('numeric', 7)
  zLower = vector('numeric', 7)

  colNum = columnsCount[k]

for(i in c(15, 20, 25, 30, 35, 40, 45)){
  print('                          ')
  print('----------------------------')
  print(paste('sample size = ',i))
  zValue = vector('numeric', 20)


  for(ii in 1:20) {   # repeats for confidence interval
    t=vector('numeric',nRuns*nBoots)
    kk = 1
    for(nn in 1:nRuns-1){
      a = bucket[, round(runif(1,1,colNum),0),k] # grab a column
      pA = sum(a)/50
```

```
    for (j in 1:nBoots) {
      aa = sample(a, replace=TRUE, size=i )
      pAA = sum(aa)/i
      pC = sum(c(aa,a))/(50+i)
      if (pA == pAA) {z = 0
      }else{  z = (pA-pAA)/sqrt(pC*(1-pC)/50 + pC*(1-pC)/i)}
      t[kk] = z
      kk = kk+1
    }
  }

#   hist(t,freq=FALSE, main=paste(titlePlot[k],'\nsample size = ',i), xlab='z-value')
  tSort = sort(t)
  zValue[ii]=tSort[0.98*nRuns*nBoots]
 }

  meanZvalue = mean(zValue)
  print(round(meanZvalue,2))
  print(round(pnorm(meanZvalue, lower.tail=FALSE),2))
  plotSave[wordsNumber]=round(pnorm(meanZvalue,lower.tail=FALSE),3)
  zSort = sort(zValue)
  zUpper[wordsNumber] = round(pnorm(zSort[19], lower.tail=FALSE),3)
  zLower[wordsNumber] = round(pnorm(zSort[2], lower.tail = FALSE),3)
  wordsNumber = wordsNumber+1
}

plot(plotSave,main=paste(titlePlot[k],'\n98% or higher match'), xaxt='n', xlab='number of
words', ylab='p value', ylim=c(0.02, 0.1),pch=19)
print(plotSave)
print(zLower)
print(zUpper)
segments(x0 = 1:7, y0 = zLower, y1=zUpper, col='red')
axis(1, 1:7, labels=c('15','20','25','30','35','40','45'))
abline(h=0.05, col='red', lty=2)
text(3, 0.04,'evidence that \nresults differ')
}
```

Validation
* break data from each each talker into two groups
* verify that the proportion of words correctly understood is the same across both sets

```{r validation}
diffSave = vector('numeric')
```

```
for(k in 1:4){  # bucket
  print(paste('case ',k))
  cCount = columnsCount[k]  # how many available
  t1=vector('numeric',nRuns*nBoots)
  t2=vector('numeric',nRuns*nBoots)
  PCsave1 = vector('numeric')
  PCsave2 = vector('numeric')


  kk = 1
  for(nn in 1:nRuns-1){
    a1=vector('numeric')
    a2 = vector('numeric')
    listenerNumber = round(runif(1,1,cCount),0)
    a1 = bucket[1:25, listenerNumber,k] # grab a column from first half
    a2 = bucket[26:50, listenerNumber,k] #grab a column from the second half
    pA1 = sum(a1)/25
    pA2 = sum(a2)/25
    for (j in 1:nBoots) {
      aa1 = sample(a1, replace=TRUE, size=25 )
      aa2 = sample(a2, replace=TRUE, size=25)
      pAA1 = sum(aa1)/25
      pAA2 = sum(aa2)/25
      pC1 = sum(c(aa1,a1))/50
      pC2 = sum(c(aa2, a2))/50
      if (pA1 == pAA1) {z = 0
       }else{  z = (pA1-pAA1)/sqrt(pC1*(1-pC1)/25 + pC1*(1-pC1)/25)}
      t1[kk] = z
      if(pA2 == pAA2) { z=0
       } else { z = (pA2-pAA2)/sqrt(pC2*(1-pC2)/25 + pC2*(1-pC2)/25)}}
      t2[kk] = z
      PCsave1[kk] = pA1-pAA1
      PCsave2[kk] = pA2-pAA2
      kk = kk+1
  }
      diff = PCsave1-PCsave2

      diffMean = mean(diff)
      print(paste('mean diff : ', diffMean))
      diffSave[k] = diffMean

}

plot(diffSave, main='validation: \ncorrelation between two sets',  ylab='diference in p values',
xaxt='n',ylim=c(-0.06, 0.06), pch=19, xlab='')
axis(1, labels=c('MRT#1 UHF', 'MRT#1 VHF', 'MRT#2 am  ','MRT#2 fm  '), at = c(1,2,3,4))
```

```
abline(h=c(-0.05, 0.05), col='red', lty=2)
text(2, 0.035,'Area of no difference')
```

```{r, eval=FALSE}

  par(mfrow=c(2,1))
  hist(t1,freq=FALSE, main=paste(titlePlot[k],'\nsample size = ',i), xlab='z-value')
  hist(t2, freq=FALSE, xlab='z-value')
  t1Sort = sort(t1)
  t2Sort = sort(t2)
  zValue1=t1Sort[0.98*nRuns*nBoots]
  PCsort1 = sort(PCsave1)
  pCinterval1 = PCsave1[0.98*nRuns*nBoots]
  zValue2=t2Sort[0.98*nRuns*nBoots]
  PCsort2 = sort(PCsave2)
  pCinterval2 = PCsave2[0.98*nRuns*nBoots]
  print(paste('pCinterval:\nset 1: ', pCinterval1,', set 2: ',pCinterval2))

  print(round(pnorm(zValue1, lower.tail=FALSE),2))
  plotSave1 = round(pnorm(zValue1,lower.tail=FALSE),3)
  print(round(pnorm(zValue2, lower.tail=FALSE),2))
  plotSave1 = round(pnorm(zValue2,lower.tail=FALSE),3)
```

```{r, eval=FALSE}

plotSave[wordsNumber]=round(pnorm(zValue,lower.tail=FALSE),3)
  print(paste('PC interval = ', pCinterval))
  wordsNumber = wordsNumber+1


plot(plotSave,main='validation: sample = 25 for each half', xaxt='n', xlab='number of words',
ylab='p value', ylim=c(0.02, 0.1),pch=19)
axis(1, 1:7, labels=c('15','20','25','30','35','40','45'))
abline(h=0.05, col='red', lty=2)
text(3, 0.04,'evidence that results differ')

```