HEARING AND SPEECH SCIENCES

# ASSESSMENT OF A MEASURE OF RESPONSE CONFIDENCE FOR A SPEECH RECOGNITION TASK IN NOISE

JOHN ANDREW DUNDAS

Dissertation under the direction of Gary P Jacobson, Ph.D.

The development of a measure of response confidence for a speech understanding task is presented in this dissertation. Normal hearing participants completed speech understanding tasks in background noise and rated confidence in their responses. In experiment 1, the relationships between measured performance, perceived performance and confidence in the correctness of responses were investigated. In experiment 2, the effect of sentence context on response confidence was investigated. The main findings of this study are; 1. Confidence ratings of speech intelligibility performance can be consistently measured using simple tools, 2. Confidence ratings are strongly correlated with measured performance, 3. Confidence ratings are highly repeatable, 4. Sentence based test materials with a high degree of context result in the most accurate calibration to measured performance, and, 5. Low context sentences result in a faster growth of confidence (i.e., overconfidence). These findings suggest that confidence ratings could be a useful outcome measure in the evaluation of treatment efficacy in the hearing impaired population.

Approved: Gary P. Jacobson, Ph.D. Date: 12/08/2009

# ASSESSMENT OF A MEASURE OF RESPONSE CONFIDENCE FOR A SPEECH RECOGNITION TASK IN NOISE

By

John Andrew Dundas

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Hearing and Speech Sciences

December, 2009

Nashville, Tennessee

| Approved: | Date: |
|---|---|
| Gary P. Jacobson, Ph.D. | 12/08/2009 |
| Troy A. Hackett, Ph.D. | 12/08/2009 |
| Benjamin W.Y. Hornsby, Ph.D. | 12/08/2009 |
| Todd A. Ricketts, Ph.D. | 12/08/2009 |

# DEDICATION

*“Ask not alone for victory; ask for courage, for if you can endure you bring honor to yourself. Even more, you bring honor to us all”* - Unknown*.*

For Laura and Fletcher. A better life lies ahead.

# ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the Vanderbilt Department of Hearing and Speech Sciences, the Vanderbilt Bill Wilkerson Center for Otolaryngology and Communication Disorders, and Professor Todd A Ricketts.

I am grateful to the faculty members and fellow graduate students who have provided invaluable assistance and criticism during this and related projects. Each of the members of my Dissertation Committee has provided extensive professional and scientific guidance both within the scope of the project and in my education and life in general. I would especially like to thank Dr. Gary Jacobson, as the chair of my committee. As teacher, mentor and friend, he has challenged and supported me to an extent far greater than I could ever hope to describe here. He has taught me through example what a good scientist, clinician and outstanding professional should be.

Nothing has been more important in the pursuit of this degree than the support of my family. I would like to thank my parents, Richard and Lynda, who have supported the decision to return to, and stay in, school despite the financial and geographic hardships the decision has imposed. Most importantly I thank my incredible wife, Dr. Laura Dundas. Without her support and belief in my abilities, this project would not have been possible. Finally, I must thank my son Fletcher, who reminded me of why I embarked on this quest in the first place. He has served as a tremendous source of motivation and inspiration since the news of his impending arrival was delivered on that wonderful July morning.

## TABLE OF CONTENTS

Appendix

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## Effects of Hearing Loss on Quality of Life

It has been well established that hearing loss can impose marked, negative psychosocial effects on individuals due to both direct and indirect communication based failures. These effects have been investigated in both clinical (Jones, Victor & Vetter, 1984; Mulrow, Aguilar, Endicott, Velez, Tuley, Charlip & Hill, 1990) and non-clinical population-based studies (Tambs, 2004), and have suggested links between self-reported hearing loss and depression, communication and cognition. Authors have also noted a link between untreated hearing loss and reductions in reported quality of life (Knutson & Lansing, 1990; Mulrow, Aguilar, Endicott, Tuley, Velez, Charlipp, Rhodes, Hill & DeNino, 1990; Tesch-Romer, 1997). Other changes in sensory function that are commonly associated with aging have been shown to have similar effects on psychosocial welfare and quality of life, including falls (Perry, Steen, Galloway, Kenny & Bond, 2001), cognitive decline and decreased vision (Kempen, van Heuvelen, van Sonderen, van den Brink, Kooijman & Ormel, 1999). One particularly interesting outcome related to psychosocial welfare is the finding that a loss of confidence and independence may result from a decline in function (Parry et al., 2001). This observation has obvious ramifications for the hearing impaired population, who may withdraw from, or fail to participate in communicative endeavors due to a lack of confidence in performance.

## Hearing Loss, Confidence and Communication

Individuals make decisions regarding interaction with the world based upon a combination of factors including knowledge (available information) and personal beliefs. The relative contribution of these factors varies with the quality and quantity of information available, prior experience, and the strength of belief of the individual. Strength of belief may vary between individuals to an extent that exceeds measurable disparity in performance. This disparity has been investigated extensively in a contemporary body of literature related to eyewitness testimony, but it is important to note that questions of the validity of personal beliefs date back to the time of the ancient Greek philosophers, who first posed questions regarding the justification of certainty. These questions evolved into the philosophical field of epistemology, the study of the relationship between knowledge and belief. As belief can exist in the absence of knowledge, the degree, or strength, of belief in knowledge or performance expressed by individuals has been referred to as confidence. Confidence is defined by Webster's new World Dictionary as *"the quality or state of being certain"* and *"a feeling or consciousness of one's powers"* (Neuman, 1998).

The decrease in function that results from a sensory deficit was, until recently, referred to as 'disability'. The World Health Organization (1980) defined disability as *"A restriction or lack of ability manifested in the performance of daily tasks"*, essentially a metric of the degree of difficulty that patients notice in their life. Handicap, on the other hand was defined as *“a social, economic or environmental disadvantage resulting from an impairment or disability”* (World Health Organization, 1980). Thus disability can be thought of as a loss of function, and handicap as the resultant impact of that loss of function on the life of the patient. The World Health Organization has since revised the manner in which health is defined in the International Classification of Functioning, Disability and Health (2001), to reflect changing attitudes about sensory and other health impairments. The current manner of considering disability and handicap is to include them under the umbrella of ‘health’ and consider disability as the *effect* of the decrement in health, rather than the *cause*. In essence, “disability” has now been renamed “health decrement”, while what was previously referred to as “handicap”, (the effect of the health decrement) is now considered “disability”. To avoid confusion, and to ease comparisons with older hearing aid outcome metrics, the 1980 definitions will be used throughout this document.

When experienced as a result of decreased function or disability, a loss of confidence can be considered to be a factor that contributes to the experience of handicap. A reduction in confidence suggests a psychosocial impact of disability that may adversely affect the likelihood of performing a task or entering a situation previously known to cause anxiety, result in failure or to constitute an unacceptable risk. This parallels the WHO (1980) definition of handicap, as discussed above. Thus, the terms confidence and handicap can be considered to be intimately related. While failure to communicate effectively and confidently appears, on the surface, to be of minimal importance when compared to reductions in safety caused by visual disturbances or physical injury due to, for example, elevated risk of falls, the reality is potentially much more significant. Failure to communicate confidently and effectively with ones' physician could result in drug toxicity or allergic reaction. Failure to communicate confidently and effectively with family and friends could result in a loss of intimacy and sense of disconnection from life. Failure to communicate confidently and effectively with members of the general public could result in social stigma and withdrawal from the demands of everyday events and activities in society. These and a multitude of other negative effects of hearing loss may adversely affect perceived quality of life (Jerger, Chmiel, Wilson & Luchi, 1995; Joore, Potjewijd, Timmerman & Anteunis, 2002, Arlinger, 2003; Dalton, Cruickshanks, Klein, Klein, Wiley & Nondahl, 2003).

## Assessment of Hearing Aid Outcome

Hearing aid outcome measures attempt to quantify outcome as a function of factors that include absolute performance, perceived disability, perceived handicap and satisfaction. Absolute performance can be assessed via measures of speech intelligibility performance that simulate various listening situations and compare performance between the unaided and aided conditions. This change in performance is a measure of hearing aid benefit. Recognizing that improving speech intelligibility does not guarantee a happy patient and successful hearing aid fitting, other authors have developed more subjective measures that ask the patient to report the degree of benefit that is perceived. Three important questions commonly raised in the subjective assessment of hearing aid outcome are: 1. *“How much difficulty do you have in a given situation when (not wearing / wearing) your hearing aids?”* 2. *“How much does any change in perceived performance affect your life?”* and, 3. *“How satisfied are you with your hearing aids?”*

The first two questions align along the dimensions of disability and handicap, respectively. These concepts must be distinguished from each other in order to consider the range of impact of hearing loss and amplification on the patient’s quality of life. In the third question, satisfaction is much more difficult to attribute to performance based factors as it may relate to cost/benefit ratio, physical comfort, expectations, ease of use or features, as easily as it might relate to performance improvements.

A largely unexplored dimension of outcome lies in the measurement of confidence in communication. Confidence would seem to be an overall measure of the value of the intervention related to performance, however, as of yet, the relationship between confidence and speech intelligibility performance has not been systematically investigated. While a high level of personal confidence is known to be correlated with increased quality of life (Whitney, Hudak & Marchetti, 1999; Whitney, Wrisley, Brown & Furman, 2004; Jowett & Ryan, 1985; Parry et al, 2001;Yardley & Smith, 2002; Dalton et al, 2003), it is not known whether speech intelligibility confidence measures could be used in the assessment of hearing aid outcome. Furthermore, it is not known whether communication confidence is driven by performance, or by other unknown variables.

It is well known amongst clinicians that patients may perceive that they understand more or less speech than they actually do. The Performance Perceptual Test (PPT) (Saunders and Cienkowski, 2002) was developed to compare perceived performance with measured performance. This is accomplished through a comparison of the SNR at which individuals believe they can just understand all of a speech perception test stimulus with the measured 50% performance level. The goal of the test is to identify disparities between performance and belief, so that patients can be counseled regarding their over- or under-confidence and hopefully recalibrate and discrepancy. This rating of intelligibility has been shown to be highly reliable in contrast to other intelligibility rating procedures as will be discussed below.

## Confidence versus Subjective Ratings of Performance

It is important to note that authors distinguish between confidence ratings and estimates of performance level. Adams and Adams (1961) note that confidence rates the degree of *certainty* of correctness in a binary factor(i.e., one that can have only one of two outcomes, such as correct or incorrect.) Estimates of performance, on the other hand are described as scalar measures that rate the *proportion* of correctness. Combining a large number of confidence ratings of binary events however, results in an overall estimate of performance without the cognitive demands of tracking continuous discourse or a list of stimuli. Confidence ratings may offer a new method of assessing outcome as this tool may help to explain the apparent contradiction that emerges in patients who perceive and report benefit even in cases where no objective benefit can be measured (and vice versa). These patients may perceive that they are more confident in their performance and that this increase in confidence is a desirable outcome as it reduces the overall stress of communication. It is appropriate at this point to consider the concepts of confidence and rated performance in some detail.

## Quantifying Confidence

Investigators in the field of psychology have sought to develop scales and criteria upon which degrees of belief can be quantified (e.g., Adams, 1957; Adams & Adams, 1961; Koriat, Lichtenstein & Fischhoff, 1980; Paap, Chun & Vonnahme, 1999). It has since been established that in the process of generating an estimate of performance, the individual first arrives at a confidence judgment based on internal cues or “feelings of doubt” (Adams & Adams, 1961). This outcome is then thought to be transformed into a quantitative estimate of the probability of accuracy. Thus, confidence is rooted in perceived competence with a particular task. Work by numerous authors has demonstrated that a positive, monotonic relationship exists between confidence and performance for perceptual tasks including high speed reading, memory and eyewitness testimony (e.g., Adams, 1957; Adams & Adams, 1961; Lichtenstein & Fischhoff, 1977; Koriat et al., 1980; Bjorkman, 1994; Perfect & Hollins, 1996; Paap et al., 1999; Stankov & Lee, 2008; Kroner & Biermann, 2007; Tenney, Spellman & MacCoun, 2008). These authors suggest that the relationship between confidence ratings and performance indicates that individuals are able to appropriately assess and report the proportional correctness of their responses. To date, no systematic investigation of performance and confidence on speech perception tasks has been published.

## Calibration and Resolution Measures for Assessment of Confidence

Early work in the area of self confidence assessed the validity of the relationship between confidence and performance by considering two aspects of the relationship, namely calibration and resolution.

### *Calibration*

The degree to which confidence in performance ($p$) reflects actual performance ($P$) is referred to as calibration (Lichtenstein & Fischoff, 1977). Calibration could also be thought of as an index of the skill of the rater at assigning probabilities (i.e., confidence ratings) to differing levels of performance. This measure is closely aligned with the concept of accuracy. Individuals who report response confidence that closely matches the measured level of performance are thus said to have better, higher, or more accurate calibration than individuals who over or underestimate their performance. Calibration is calculated in an approach similar to the sum of squares terms of the analysis of variance, and reflects the expected binomial distribution of confidence ratings for correct vs. incorrect performance results. That is, for each performance level, the expressed confidence rating (i.e., the rated probability that the response is correct), will potentially differ from the measured performance value. Calibration then represents the observed deviation of the confidence rating from the expected (correct) value. Calibration can be calculated mathematically from:

Equation 1: Calibration $$Calibration = \frac{1}{N}\sum_{t=1}^{T} n_t (r_t - c_t)^2$$

This allows us to quantify the relative magnitude of the deviation of confidence from actual performance. The difference between rated confidence ($r_t$) and measured performance ($c_t$ ) is calculated for each confidence rating level. This deviation is squared to remove the effects of positive and negative differences. To account for the frequency of use of different confidence intervals (*T*), the squared deviations are weighted by multiplication by the number of times the response interval is used ($n_t$). These deviations are then averaged by dividing the summed weighted difference values by the total number of observations (*N*). Calibration scores are scaled between an optimum value of 0 and an extremely poor score of 1, which could only result if a participant rated their confidence in complete opposition to their actual performance. Thus a minimized calibration score would suggest excellent agreement between rated confidence and performance.

*Resolution*

The ability of the individual to assign feelings of confidence to rating categories with changes in performance is referred to as resolution (Baranski &

Petrusic, 1994). Resolution is similar in concept to precision and represents the smallest detectable change in performance. Individuals who are able to report appropriate scalar changes in confidence for small changes in performance are said to have higher resolution. The measure assesses the ability of the participant to use the selected confidence categories to distinguish when an event occurs versus when it does not (Baranski & Petrusic, 1994).

Equation 2: Resolution

$$resolution = \frac{1}{N}\sum_{t=1}^{T} n_t (c_t - c)^2$$

Similar to the calibration equation, resolution is calculated by computing the weighted mean squared difference between measured performance ($c_t$) and overall mean performance ($c$) within a response category (T). These differences are squared to eliminate positive and negative differences and weighted through multiplication by the frequency of use of the rating category ($n_t$). The mean change in performance for a one unit change in rating is then computed by dividing the summed values for each response category by the total number of responses. Similar to the calibration equation described above, the squaring of difference scores ensures that the positive and negative deviations of ratings relative to the measured performance level do not cancel one another out. Instead, a measure of the mean dispersion, or error in measurement is generated. In equations 1 and 2, $N$ is the total number of responses, $n_t$ is the number of times the response confidence level $r_t$ was used, $c_t$ is the proportion correct for items rated confidence level $r_t$, and $T$ is the total number of response categories used. In equations 2, *c* is the overall proportion of correct responses. A resolution score can be thought of as a slope term. In the unique case of a binary event (i.e., the answer can only be 100% correct or 0% correct) a resolution score that approaches a value of 0.25 suggests optimal resolution. A resolution value approaching this optimum would suggest that the individual is better able to sort the probability of their responses being correct into the various categories allowed. A score that approached zero would suggest that the individual was completely unable to perform this task. According to Baranski and Petrusic (1994) however, resolution scores greater than 0.1 are rarely encountered.

### *Personal Calibration*

The issue of personal calibration is an important one when considering speech communication ability. If an individual is consistently overconfident in their level of performance, they may make embarrassing errors in communication that could have a negative impact on personal interactions. Similar tendencies could lead to poor decision making that could affect personal health, safety or security. Conversely, individuals who are consistently under confident in their performance may withdraw from society and interpersonal interactions, expecting failure. This too may contribute to reductions in health and well being, affecting quality of life. Along these lines, Lichtenstein and Fischhoff (1977) argue that confidence calibration quality is a limiting factor of the quality of individual performance in uncertain environments.  It is therefore clear that understanding the abilities of individuals to assess the quality of their own function has potentially important theoretical and practical implications related to hearing loss and speech understanding. In fact, Stankov and Lee (2008) argued that the development of appropriately calibrated confidence in performance was of greater importance in decision making than was the actual performance level. If this assertion is to be believed, we are forced to consider the possible negative effects of over or under confidence in performance on health related quality of life.

In the 1970s, researchers posed the question of whether individuals who know more, also know more about *how much* they know (Lichtenstein & Fischhoff, 1977). The authors reported that individuals with higher levels of knowledge (and therefore performance) exhibited superior calibration to less knowledgeable controls up to approximately 80% correct performance, beyond which, the more knowledgeable subjects tended to underestimate their performance. Stankow and Lee (2008) reported that less knowledgeable participants tended to overestimate their performance. The implication is that when individuals are equipped with better information or experience on which to base their decision making process, they are better able to generate appropriate estimates of confidence. It is possible that improving access to speech stimuli may have this effect on the individual. Thus, amplification may provide sufficiently improved speech information to hearing impaired individuals that an improvement in calibration would result. Further, it appears that calibration can also be improved through training. Koriat, Lichtenstein and Fischhoff (1980) investigated the ability of participants to improve their calibration by providing feedback on the accuracy of confidence judgments during practice sessions. The authors demonstrated that providing feedback greatly decreased under- and overconfidence in performance and thereby improved calibration. The authors also reported that individual differences in calibration decreased with practice, and that resolution improved. It appears to be important then that individuals receive feedback on their performance level so that an appropriate calibration can be achieved between confidence and performance. Other factors have also been shown to influence the confidence/performance calibration. For example, Lichtenstein and Fischhoff (1977) demonstrated that when an experimental group was trained on a visual recognition task prior to testing, that their performance, calibration and resolution were significantly better than those produced by untrained controls.

Confidence and performance function research relative to perceptual tasks has been conducted primarily for visual and written stimuli. As of yet, no evidence has come to light of the systematic study of confidence ratings and speech intelligibility performance. Thus it is unclear whether performance affects absolute confidence, calibration or resolution. Further, it is unclear whether hearing loss has an effect on these factors. This oversight appears significant when considering the possible ramifications of communication confidence in the hearing impaired population. It is also possible that communication confidence may help to develop the understanding of hearing aid outcome, currently unpredictable due to the documented shortcomings of more common measures of hearing aid outcome.

## Measurement Scale Effects on Resolution

Early research on confidence ratings was based in intellectual knowledge and utilized categorical scales which limited the response options available to the respondent. J.K. Adams (1957) sought to improve perceptual confidence rating tools by providing a scale which allowed a rating of confidence as a percentage. This scale resulted in the first assessment of calibration in confidence versus performance, entitled ‘realism of confidence judgments’ (Adams & Adams, 1961). As the number of possible responses increases, opportunities for the individual to apportion ratings of confidence into appropriate categories are increased. As such, resolution should be improved when the individual is not restricted by the available responses. Theoretically, a scale with no restrictions should then result in improved sensitivity to changes in perceived performance.

## Communication Confidence

One conclusion to be drawn from the literature is that if hearing aids improve access to speech information, then they should also reduce the cognitive demands on the listener, as less compensatory auditory decoding is required to process the improved auditory speech signal. This effect was demonstrated by Downs (1982) in that reaction times to competing task stimuli were decreased when wearing appropriately fit hearing aids, as compared to the unaided condition. Similarly, Rakerd et al. (1996) reported that a lower proportion of cognitive resources were required by aided listeners than when completing the same listening task while unaided. It is posited that if less effort is required to successfully complete the task, not only should long-term performance improve, but also that the confidence in the response should improve. If actual performance and confidence both improve, the overall stress of communication should decrease, leading to lower ratings of handicap, higher satisfaction and increased ratings of quality of life. It may yet be determined that participants with higher perceived performance experience greater confidence in communication, and that communication confidence may be correlated with the outcome of the hearing aid experience. However, as previously noted, traditional measures of outcome have been conducted as comparisons of unaided to aided performance on various speech test measures, or via subjective ratings of perceived unaided versus aided difficulty, disability, handicap and/or satisfaction. These approaches do not assess processing effort, perceived ease of communication or communication confidence. A review of the outcomes measures literature has not brought to light any other evidence of the use of subjective ratings of confidence in speech material intelligibility performance. This study suggests the development of a new measure of the impact of amplification on quality of life, which we will call ‘communication confidence’.

As discussed above, confidence is believed to be rooted in perceived competence with a particular task (Adams & Adams, 1961). However, competence does not guarantee confidence, nor is the reverse true. In individuals with hearing loss, confidence could be described as a perceptual correlate of communication success. It may be that patients experience greater communication confidence because the effort involved in communication is reduced in particular situations with the use of hearing aids. For example, Kodman (1961) reported that binaural amplification did not improve measured word recognition, but that subjectively, patients reported reduced listening effort. McCoy and colleagues (2005) demonstrated that for hearing impaired listeners to perform at the same level as normal hearing controls on a speech recognition task, a greater proportion of processing capacity was required. Noble (2006) reported that patients fit bilaterally with hearing aids reported lower listening effort and better spatial hearing performance than did patients fit unilaterally. Perhaps then, communication confidence is not tied to actual word recognition performance, but to the reduction in expended cognitive resources required for the enjoyment of, and participation in, daily communication activities.

## Performance Ratings of Intelligibility

While it is posited that communication confidence is not solely dependent on measured speech intelligibility performance, one potential correlate of confidence may lie in ratings of perceived performance, commonly referred to as ‘speech intelligibility ratings’. The two concepts bear certain superficial resemblances to each other, wherein the individual must make an internal judgment of the quality of the response to the stimulus based on perceptual cues known only to the individual.

Early work in this area arose from studies of the effects of masking on speech understanding. Hawkins and Stevens (1950) presented sentences in a background of white noise. Participants were required to track to the “threshold of intelligibility”, (TI) and the “threshold of detectability” (TD). The TI is defined as the level at which the listener can just understand the meaning of almost every sentence. This measure is interesting due to two components of the definition, specifically the requirements to ‘understand the meaning’ and ‘almost every sentence’. It would appear that this makes the rating extremely subject to interpretation of test instructions. The TD is defined as “the level at which the listener can just detect the presence of speech, about half the time.” This definition faces a similar problem to that of the TD in that the listener is required to make a judgment of ‘about half the time’ in addition to detecting speech. Similar work was published by Falconer and Davis (1947) who reported on the threshold of intelligibility for connected discourse in dB. In this early work, intelligibility was considered to be a decision by the listener of how well the message was understood (Speaks, Parker, Harris & Kuhl, 1972).

Subjective ratings of intelligibility have been investigated by numerous authors in an effort to better understand the relationship between perceived performance and measured performance by attempting to scale ratings of intelligibility on ordinal and integer graphic scales (Cox & McDaniel, 1984; Cox, Alexander & Rivera, 1991; Preminger & van Tasell, 1995; Saunders & Cienkowski, 2002). As discussed above relative to confidence ratings, when rating intelligibility, coarser scales appear to encourage participants to choose an ordinal value for their rating of intelligibility, a problem in that this to some extent preordains the outcome of the estimates by reducing the resolution of the responses. Finer scales appear to allow the listener more latitude in their ratings of intelligibility. For example, Speaks and colleagues (1972) compared intelligibility ratings obtained using a restricted (0%, 25%, 50%, 75%, 100%) scale to those obtained with an unrestricted scale. The authors reported that while both methods allowed fairly accurate predictions of measured intelligibility, the unrestricted scale resulted in a stronger correlation with measured performance (0.93) than when using the restricted scale (0.84).

Preminger & Van Tasell (1995) investigated the relationship between speech quality and speech intelligibility by investigating several areas of speech production thought to be important to speech perception, namely intelligibility, pleasantness, loudness, effort (of listening), and the total impression. The authors reported that intersubject reliability was high, that the various speech rating dimensions were indistinguishable from one another, and that each dimension was strongly correlated with intelligibility, with the exception of overall impression.

Rankovic and Levy (1997) argued for the use of nonsense consonant-vowel-consonant materials due to the fact that small differences in performance are more detectable with these materials than with sentence or passage materials as context and familiarity is removed. When SNR was varied in speech weighted white noise, participants estimated performance as the percentage of the target stimuli repeated correctly, using a large integer scale in the form of a horizontal bar ranging from 0% to 100% correct. The authors reported that throughout the range of performance, the ratings of intelligibility overlapped the range of performance, suggesting that listeners are able to accurately estimate their performance level for speech-like test materials. Conversely, Preminger and colleagues (2000) described a study wherein subjects were allowed to alter hearing aid gain characteristics to maximize perceived speech understanding. Results suggested that significantly improved rated intelligibility was not correlated with performance on the CUNY Nonsense Syllables Test.

Saunders & Cienkowski (2002) reported the development of a test designed to identify disparities between rated and measured intelligibility using Hearing in Noise Test (HINT) sentences. The authors argue that estimates of intelligibility can be unrealistic, and that the discrepancy between rated and measured performance can affect outcomes with amplification.

Interestingly, authors have determined that depending upon the test protocol, rated intelligibility can be highly correlated (e.g., Cox et al., 1991) or poorly correlated (e.g., Preminger, Neuman, Bakke et al., 2000) with measured intelligibility. It is possible that the disparity in results could be attributed to differences in instruction sets between studies. Speaks and colleagues (1972) remarked that a major problem exists with these types of measures, in that the observer can never truly know what the participant means when they report that they 'just understand' the test materials. That is, individuals may interpret these directions as asking that they indicate the point at which they are receiving enough of the message that they understand all of the meaning, each of the individual words, that they understand the gist of the passage or some other gradation of perceived performance.

In summary, while confidence and rated intelligibility appear on the surface to be related, psychology research suggests that response confidence is a phenomenon independent of self perceived performance (Adams & Adams, 1961). However, no systematic studies of confidence have as yet been conducted using a speech intelligibility task. Thus, it has not yet been determined whether rated intelligibility is analogous to or correlated with communication confidence, or whether these factors represent unique constructs of listening and intelligibility.

## Proof of Concept

A pilot experiment was conducted to investigate the feasibility of measuring communication confidence with a visual analog scale rating tool. The effect of listening difficulty encountered under various listening conditions was explored to determine whether communication confidence ratings would vary independently of performance. Results indicated that confidence ratings were significantly correlated with performance, and the ‘difficulty’ of the listening situation affected the communication confidence ratings of the participants. For a detailed review of the pilot study, please see Appendix C.

## Goals of the Present Study

Confidence has been tied to both handicap and quality of life in a diverse range of studies investigating topics including vestibular function (Whitney, Hudak & Marchetti, 1999, Whitney, Wrisley, Brown & Furman, 2004), skin disease (Jowett & Ryan, 1985), urinary incontinence (Parry et al, 2001), falls (Yardley & Smith, 2002), and hearing loss (Dalton et al, 2003). To date, however, confidence is a dimension of speech communication that has not been described. Furthermore, the effect of audiologic rehabilitation on confidence has not as yet been investigated. It is possible that audiologic rehabilitation in the form of appropriately fit amplification could serve to increase confidence in communication, and that this improvement could be utilized in assessing changes in health related quality of life.

A long term goal is to determine whether a communication confidence rating tool could be used in the assessment of hearing aid outcome following research described in this study. The hope is that this tool may help clinicians to draw distinctions between measured performance and reported performance changes (benefit), and their contributions to patient perceptions of the hearing aid experience (e.g., satisfaction).

An important first step in this process, and a primary purpose of the current study, is to explore the nature of the relationship between measured performance and communication confidence for normal hearing adults by obtaining confidence ratings at varying performance levels.

Second, the test-retest reliability of communication confidence ratings will be investigated, to determine whether confidence ratings could be compared within the individual from visit to visit.

Third, the relationship between communication confidence and rated intelligibility will be investigated in an effort to determine whether these measures can be experimentally differentiated from each other.

Fourth, the effect of stimulus context will be investigated by comparing communication confidence ratings for high predictability/context connected speech and low predictability/context connected speech. Two experiments were designed and executed to achieve these goals.

# CHAPTER II

# METHOD

## Experiment 1

Experiment 1 was conducted with three purposes in mind. The first was to examine the test-retest reliability of communication confidence ratings. The second was to explore the nature of the relationship between measured performance and communication confidence for normal hearing adults by obtaining confidence ratings at varying performance levels. The third was examine the relationship between communication confidence and rated intelligibility.

## Research Questions

In Experiment 1, the following research questions were addressed:

1.1 When monosyllabic words are presented in multi-talker babble, what is the test-retest reliability of communication confidence ratings as a function of performance level?

1.2 When monosyllabic words are presented in multi-talker babble, what is the relationship between communication confidence ratings and performance?

1.3 What is the relationship between communication confidence ratings and ratings of intelligibility?

1.4 Do calibration and/or resolution vary between ordinal and visual analog scales of confidence?

## Hypotheses

1.1 Test retest reliability as measured via Pearson product moment correlation coefficient will be high (i.e., indicating a strong relationship between responses obtained during the two experimental sessions.)

1.2 Communication confidence ratings and performance will be strongly and positively related. (e.g., participants will rate confidence higher when performance is higher.)

1.3 Communication confidence ratings will differ significantly from intelligibility ratings.

1.4 Calibration will not differ between the ordinal and visual analog scales. However, resolution will be improved with the visual analog scale.

# METHOD

### *Participants*

Twenty-two adult participants between the ages of 23 and 43 years of age (mean, 27.9 years, SD, 5.44 years) took part in the experiments. Four were male, eighteen were female. Experiment 1 data from one female participant was corrupted and unusable. This participant’s experiment 2 data was utilized. All participants had normal hearing sensitivity at the time of testing. Left and right hearing threshold levels were averaged between ears as each participant exhibited symmetrical hearing sensitivity. All participants were recruited to participate in the study via word of mouth and poster advertisements, and visited the laboratory for the sole purpose of participating in the study. Participants were compensated for their time and efforts at the conclusion of the study, at the rate of ten dollars per hour. All study procedures were approved by and conducted in accordance with the guidelines of the Vanderbilt University Institutional Review Board.

## Test Measures and Procedures

### *Informed Consent*

All participants provided informed consent for participation in the study. The study purpose, goals and procedures were explained orally, and each participant was given the opportunity to read a copy of the informed consent document. Each participant was asked to summarize the study procedures prior to signing the informed consent document. All participants were consented by the primary investigator.

### *Otoscopy, pure tone audiometry and immittance*

Otoscopy was conducted on each participant to verify that ear canals were clear of occluding cerumen or foreign debris. Air conduction and bone conduction thresholds were measured at 500, 1000, 2000, 4000 and 8000 Hz using a standard clinical protocol (i.e., down 10 dB, up 5 dB). Normal hearing was defined as the presence of air conduction thresholds of better than or equal to 20 dB HL at all test frequencies. Symmetry was defined as no more than 15 dB difference between ears at any one frequency, and no more than 10dB difference between ears at any two adjacent frequencies. Tympanograms and screening acoustic reflexes were obtained in each ear to verify normal middle ear function. Participants with conductive hearing loss components (air/bone gaps of ≥ 10dB), or with absent acoustic reflexes were excluded from the study.

### *Mental Status Screening*

The Short Portable Mental Status Questionnaire (SPMSQ) (Pfeiffer, 1975) was administered to each participant in an interview format. All participants accumulated 1 or fewer errors, suggesting intact cognitive status.

### *Rating Scales*

Three rating scales were utilized to investigate ratings of confidence and performance. Cox and McDaniel (1989) developed a performance estimation tool which combined an integer scale with an ordinal scale (Figure 1). This scale was utilized in the speech intelligibility rating component of the study.

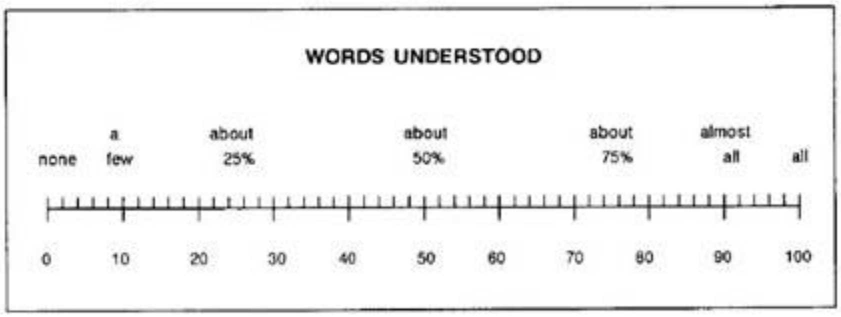

Figure 1. Performance rating scale from Cox and McDaniel (1989).

A confidence rating scale was developed by the investigator that was adapted from the Cox and McDaniel (1989) scale described above. This scale modified the descriptive markers to reflect perceived confidence rather than performance (Figure 2). Specifically, ‘a few’ and ‘almost all’ were changed to ‘a bit’ and ‘almost completely’, respectively. This scale will be referred to as the “ordinal confidence scale.”

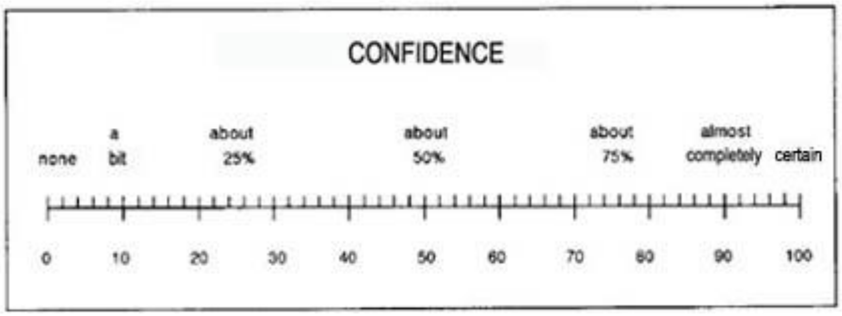

Figure 2. Confidence rating scale derived from Cox and McDaniel (1989).

Finally, a second confidence rating tool was designed as an unrestricted visual analog scale by removing numeric markers altogether. Rather, participants placed a vertical mark on a scale that ranged from ‘very low’ to ‘very high’ to indicate the level of their confidence in their response (Figure 3).

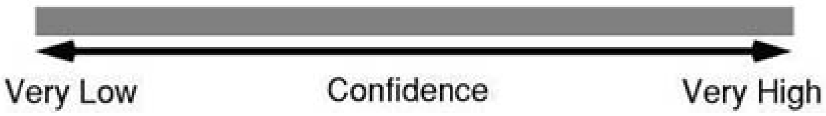

Figure 3. Visual analog confidence rating scale.

Each rating tool was scaled to a 10cm length so that the physical placement of the mark could be compared between scales.

## Test Stimuli

### *Randomly Ordered NU-6 in Speech Babble Noise*

Male talker Northwestern University Number 6 (NU-6 ordered by difficulty, Auditec of St. Louis) sentences were randomly ordered into ten-sentence blocks using a MATLAB program (Matlab V. 7.0.4, The MathWorks, Inc.). The NU-6 sentences consist of a carrier phrase and a target word (i.e., “Say the word (target word)”). A five second pause was placed between each of the sentences in the block. A random segment of cafeteria noise shaped to match the average spectrum of the NU-6 keywords was selected by the Matlab program and merged into the left channel of a stereo audio track. Noise played continuously during each experimental block. The ten-sentence block comprised the right channel. Twenty blocks of ten sentences were generated from the NU-6 word lists for each participant, and written to compact disk using Adobe Audition (V 1.5, Adobe Systems Inc.). This process was repeated as required so that each participant performed the experiment with a unique randomization of the NU-6 test words.

## General Procedures

### *Stimulus presentation and calibration*

Following audiometric assessment and mental status screening, subjects were seated in a sound-treated booth. Target words and noise were played back from a compact disk player (Technics SL-PG450) routed through the external A and B inputs of a Grason Stadler audiometer (GSI-16). Stimuli were presented from a single loudspeaker (Tannoy System 600A) located at 0° azimuth at a distance of 1.5 m from the center of the listener's head. Target sentences were presented at 70dB SPL, as measured using a Type I sound level meter (Larson Davis model 824). The sound level meter was set to measure with flat weighting, 60-90 dB range, slow averaging. The presentation level was selected to simulate typical sound levels in noisy conversational situations (Wilson, 2003). Noise was routed through channel 1 of the audiometer, and speech stimuli through channel 2. SNR was varied by adjusting the channel 1 (i.e., noise) attenuator dial.

### *Initial SNR estimation*

Participants were instructed to repeat the last word of each sentence. If the keyword was not heard, the participants were instructed to indicate that they had not heard the word. Using the first two lists of ten words, the 50% correct signal to noise ratio (SNR) was determined using a modified version of the HINT adaptive procedure (Nilsson, Soli & Sullivan, 1994). Specifically, the noise level was increased when a correct response was obtained, and decreased when an incorrect response was obtained. Two, ten sentence blocks (twenty keywords) were used for this task. Noise level was adjusted in 5dB steps for the first five sentences, then in 2dB steps for the remaining 15 sentences. The mean of the last 16 noise levels was calculated to determine the 50% correct SNR ($SNR_{50\%}$). Six SNRs were then computed, at $SNR_{50\%,}$ $SNR_{50\%}$ +2 dB, $SNR_{50\%}$ +4 dB, $SNR_{50\%}$-2 dB, $SNR_{50\%}$-4 dB and $SNR_{50\%}$-6 dB. Each computed SNR was then randomly assigned to three of the ten word blocks. During visit 1, participants completed the speech recognition task, the confidence rating procedure and the intelligibility rating procedure. During visit 2, participants repeated the confidence rating task using the scale illustrated in either Figure 2 or Figure 3.

### *Speech in noise testing*

Participants completed testing during two visits of 90 to 120 minutes each. Visits were separated by a minimum of one week. Prior to the commencement of each session, participants were instructed according to the directions listed in Appendix A.

### *Speech recognition task*

Each participant was instructed to listen for and repeat the last word of sentences presented in a background of competing noise. Participant responses were recorded as text in a spreadsheet (Microsoft Excel XP) to allow automated scoring of responses.

### *Confidence rating task*

The confidence task was completed concurrently with the speech recognition task. Participants rated their confidence in their response following identification and repetition of the keyword using the scale in Figure 2 or Figure 3. Scale use order was randomized amongst participants. Participants repeated the confidence rating task using the remaining confidence rating scale (Figure 2 or 3). The scale used first was used again during the second visit in order to evaluate test-retest reliability.

### *Rated intelligibility task*

In the rated intelligibility task, participants listened to and repeated all ten sentence keywords prior to making an estimate of their performance using the scale illustrated in Figure 1.

### *Data collection*

Eighteen blocks of ten sentences each were administered for each of the tasks described above; three blocks at each of six SNRs. Verbal responses were tabulated in a Microsoft Excel File (Microsoft Excel 2002) to track the number of correct and incorrect responses. Percent correct scores, mean confidence ratings and mean rated intelligibility judgments were then calculated for each SNR based on the number of correct responses in the three blocks block of ten words.

## Experiment 2

Experiment two was conducted to investigate the effects of sentence context on confidence ratings of performance. Participants, procedures and instrumentation were identical to those described above in experiment one, with the exception of test stimulus. Revised Speech In Noise Test (R-SPIN) sentences (Kalikow, Steven & Elliott, 1977) were used as the speech stimulus. The SPIN test is a speech recognition test utilizing sentences that have either high or low predictability target words. The target word of each sentence is the last word of the sentence. The high predictability keywords are made predictable by the context of the carrier phrase. Conversely, target words with low predictability are found in sentences where the context of the carrier produces ambiguity. For example, “The sailboat broke its MAST” would have greater predictability than “They are considering the MAST” because of the contextual influence of the cue words “sailboat” and “broke” on the listener. Low-predictability sentences, on the other hand, have a final word that cannot be predicted from the context of the sentence.

## Research Questions

2.1 When sentences are presented in multi-talker babble, what is the relationship between communication confidence ratings for sentence keywords and performance?

2.2 When sentences are presented in multi-talker babble over a range of signal-noise-ratios, does communication confidence vary as a result of high context or low context conditions?

2.3 Is calibration improved in a high context test condition relative to a low context condition?

## Hypotheses

2.1 Communication confidence ratings and performance will be strongly and positively related. (e.g., participants will rate confidence higher when performance is higher.)

2.2 Confidence ratings will be significantly higher across performance levels when participants are tested using high context test materials than when listening to sentences with low context. (e.g., participants will rate confidence as higher when listening to sentences with high context.)

2.3 Calibration will be improved in the high context test condition relative to calibration in the low context condition.

## METHOD

### *Participants*

Participants from experiment 1 completed experiment 2 sequentially. Please see the participants section from experiment one for details.

### *Rating Scale*

The ordinal confidence rating scale (Figure 2) was utilized for all estimates of confidence during experiment two.

## General Procedures

### *Stimulus presentation and calibration*

Three randomly selected fifty word SPIN sentence lists were presented to each listener. Presentation level was calibrated in a similar fashion to the procedure detailed in experiment one. Test stimuli were presented in a background of R-SPIN noise (i.e., multi-talker babble) at 70 dB SPL at three signal to noise ratios based on the 50% SNR established during experiment one. These signal to noise ratios were designated $SNR_{50\%,}$ $SNR_{50\%}$ +3 dB and $SNR_{50\%}$ -3 dB$_{.}$

### *Contextually Influenced Sentences in Noise*

Three, fifty-sentence lists were randomly selected from the eight equivalent lists of the Speech In Noise (SPIN) test. Sentences were presented in a background of 12 talker babble noise, the level of which can be varied independently of the target phrases. The first list of 50 words was presented at the 50% correct SNR determined in the NU-6 speech in noise task of experiment one. The second and third lists were presented at $SNR_{50\%}$ + 3dB and -3dB, respectively. At the conclusion of each sentence, the participant repeated the target (last) word of the sentence. The response was recorded in a spreadsheet

(Microsoft Excel XP). Percent correct scores were calculated for each SNR for both low and high predictability sentences.

### *Confidence ratings*

Following the attempt to repeat the sentence target word, participants rated confidence in the correctness of their response by making a vertical mark in an appropriate region of the ordinal confidence scale (Figure 2). Mean confidence ratings were tabulated for each signal noise ratio for both high and low predictability sentences.

## Data Analysis

The central questions of this study concern the relationships between confidence, performance and rated intelligibility. Accordingly, it was established that the data collected would be examined using regression and correlation approaches. Regression statistics were computed using a statistical analysis software package (SYSTAT v 10.0) on a Dell Inspiron B130 personal computer. In order to examine test-retest reliability, mean performance and mean confidence values were calculated for each presentation SNR to minimize differences in performance.  Pearson's R correlation coefficients were calculated as a measure of test-retest reliability.

# CHAPTER III

# RESULTS

## Experiment 1

### *Audiometric Data*

Figure 4 illustrates the mean pure tone thresholds of the normal hearing participants with error bars showing +/- one standard deviation.

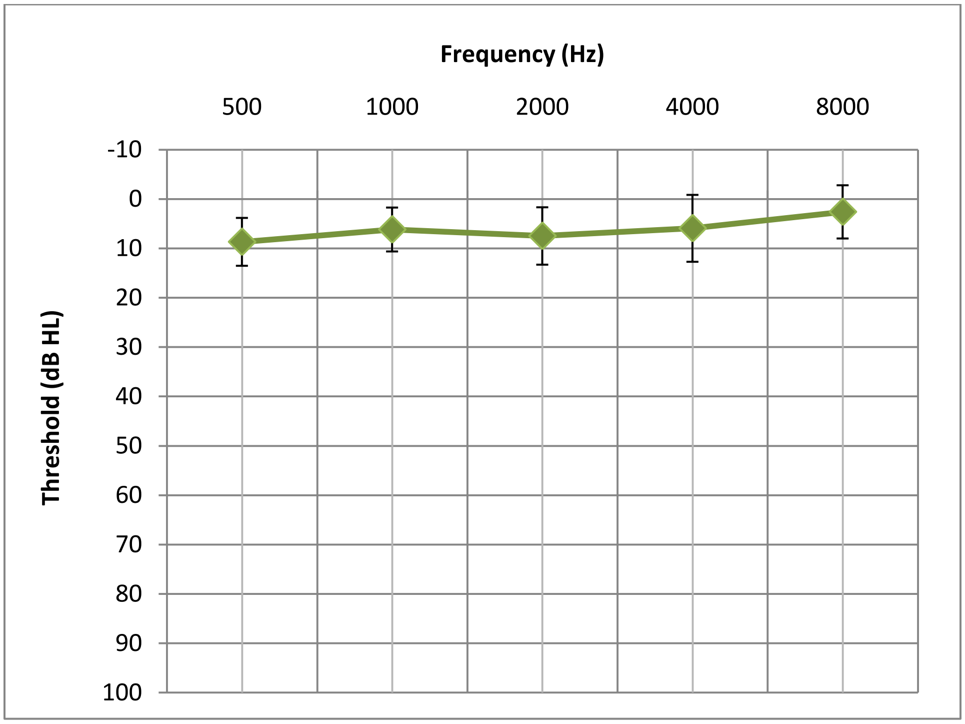

Figure 4. Mean audiometric test data for normal hearing participants.

## Confidence Ratings, Rated Intelligibility and Performance Data

*Test-Retest Reliability*

Thirty-word performance scores at each SNR were compared between runs one and two for all participants. This resulted in 120 data pairs for each measure for each of the two confidence scales. The correlation between performance results for runs one and two for the ordinal scale was significant (*r=0.942, p<.001*). Similarly, the visual analog scale results revealed a positive test-retest correlation of performance (*r=0.896, p<.001*) for the visual analog scale.

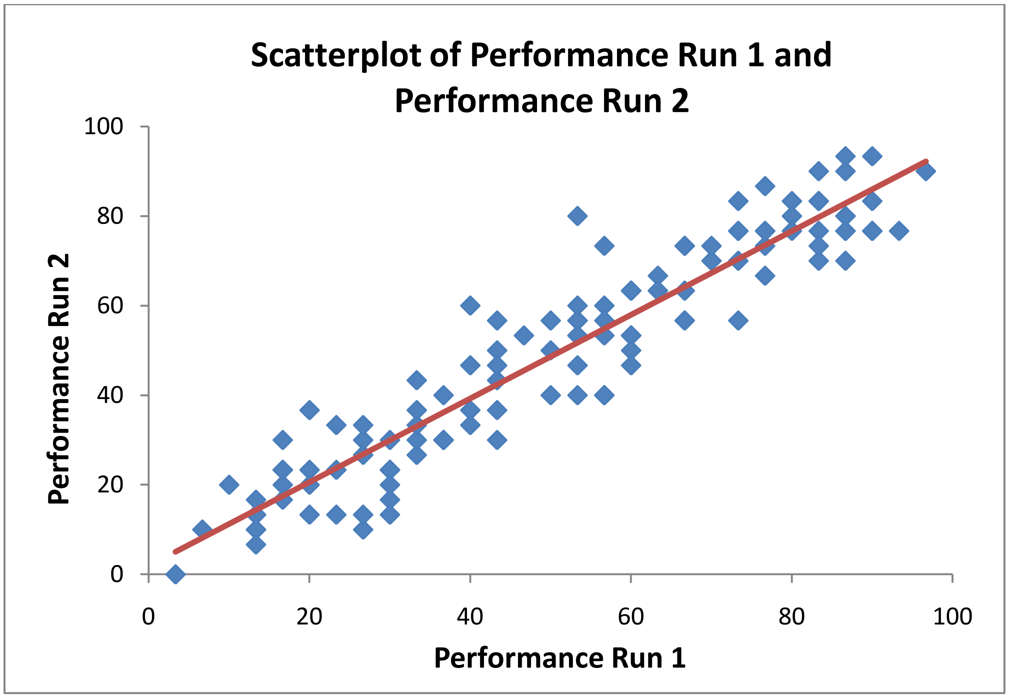

Figure 5. Test re-test reliability of performance for 21 normal hearing participants, ordinal scale data.

Next, confidence ratings obtained during visit one were compared to ratings obtained during visit two as illustrated in Figure 6. Thirty word mean confidence data for the ordinal scale was observed to be highly reliable (*r=0.901, p<.001*). Ordinal scale results are plotted in Figure 6. A positive test-retest correlation was also observed for the visual analog scale (*r=.833, p<.001*). These data suggest that for a given performance level, confidence ratings were highly repeatable between test sessions.

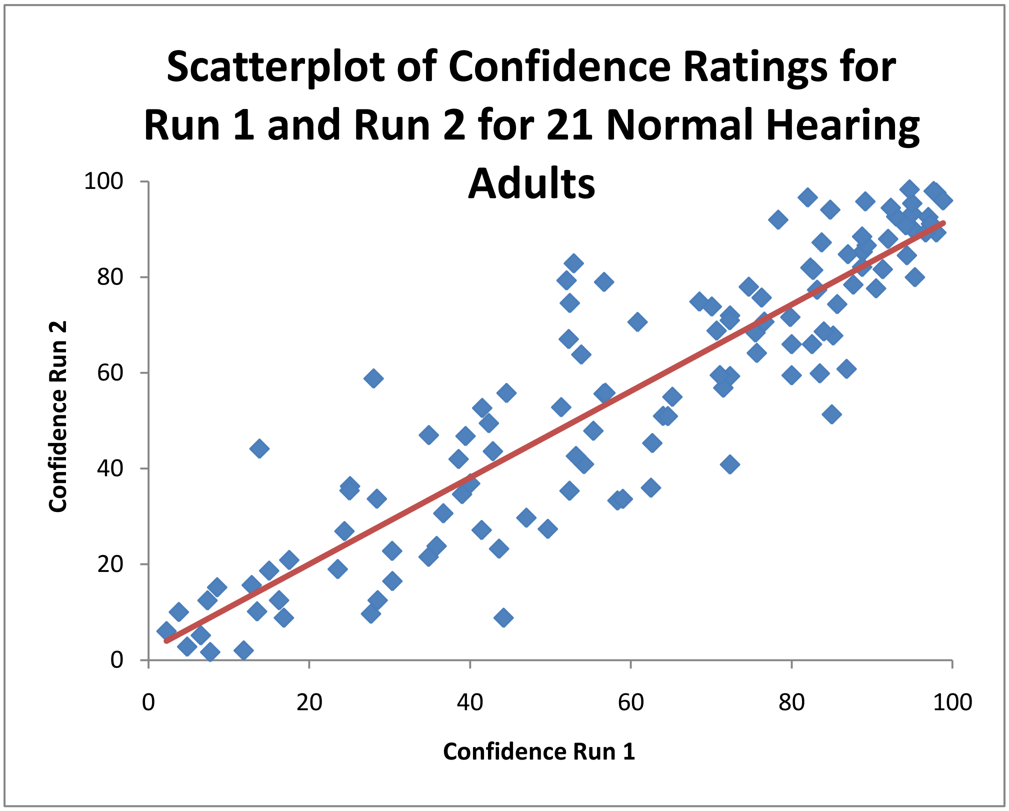

Figure 6. Scatterplot of 30 word mean confidence ratings obtained from run 1 versus run 2, ordinal scale. Red Line = correlation trendline.

An alternate method of analyzing these data was next employed, as described by Adams and Adams (1961). Individual confidence ratings were sorted into ten point ranges (e.g, 0-9, 10-19…80-89, etc.). The confidence responses in each range were then compared with their corresponding speech intelligibility performance. For example, during run 1 using the ordinal scale, for the 299 confidence responses found in the 50-59 range, 128 speech responses were scored as correct by the examiner. This resulted in 42.8% correct performance for confidence responses in this range. This example data point is highlighted in Figure 7. Figures 7 and 8 illustrate the repeatability of performance and confidence ratings across the range of performance for both the ordinal and visual analog scales.

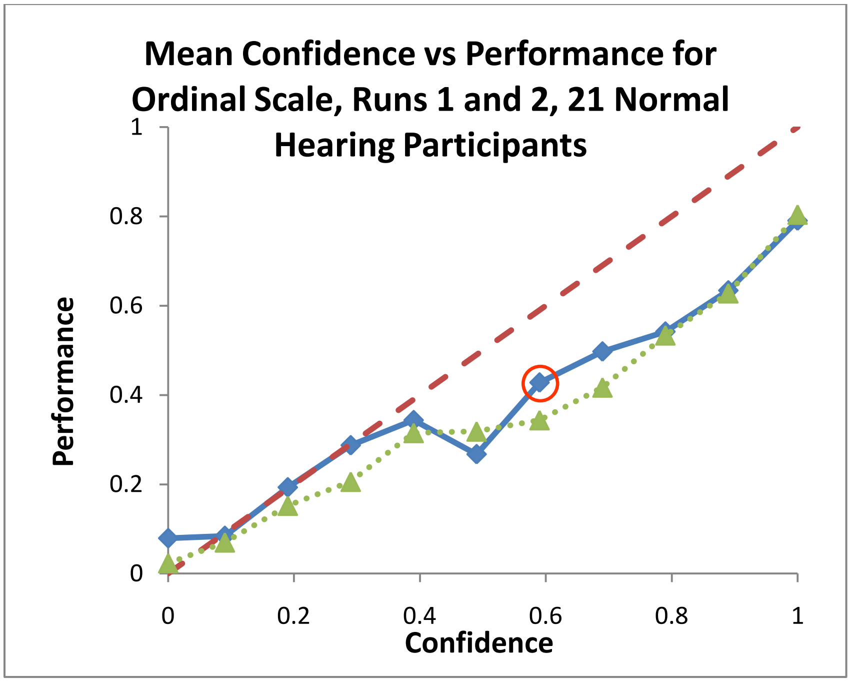

Figure 7. Test-Retest of Mean performance for each confidence rating range. Collected from 21 normal hearing adults using an ordinal scale. Legend: Diamonds = Visit 1, triangles = Visit 2, Dashed line = Ideal calibration, Red circle = referenced sample data point.

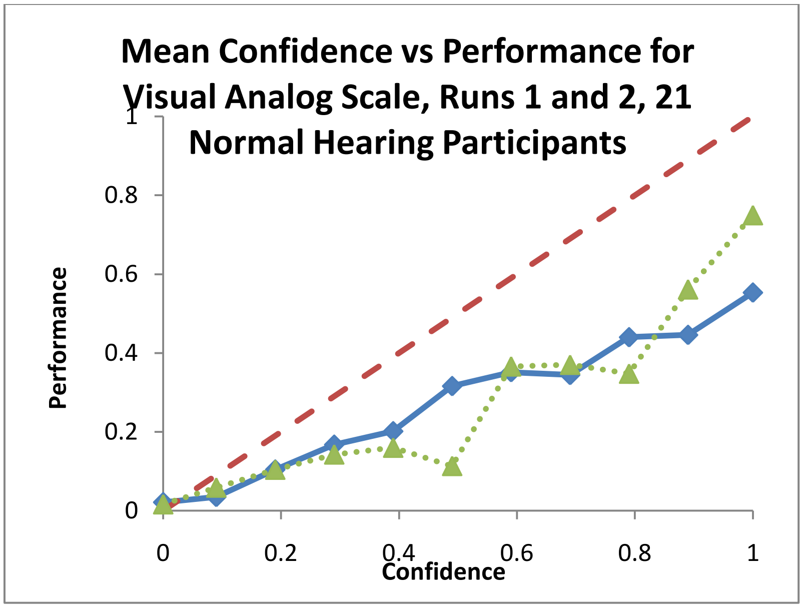

Figure 8. Test-Retest of Mean performance for each confidence rating range. Collected from 21 normal hearing adults using a visual analog scale. Legend: Diamonds – Visit 1, triangles – Visit 2, Dashed line - Ideal calibration.

*Confidence ratings and performance*

Absolute performance data was plotted as a function of SNR change (Figure 9). Performance increased with SNR.

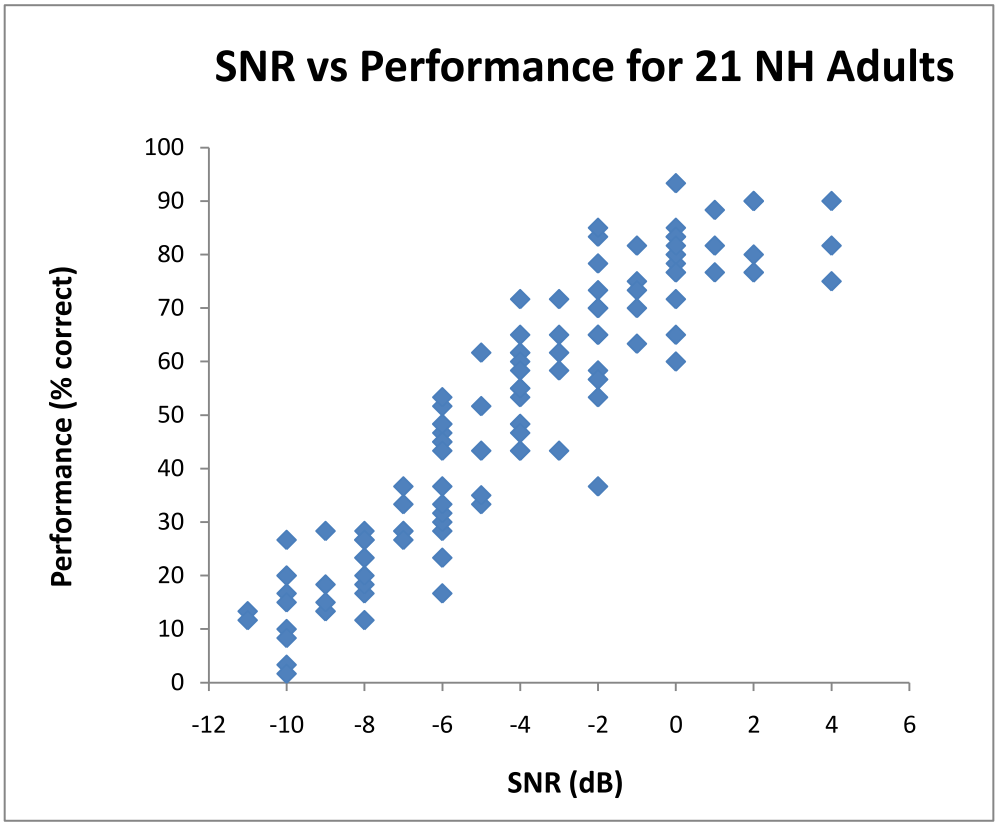

Figure 9. SNR vs performance data for 21 normal hearing adult participants

*Individual performance/confidence plots*

The first approach employed to examine the relationship between confidence ratings and performance utilized the mean confidence rating within a SNR block of thirty words compared to the proportion of words repeated correctly within the same block. Performance was observed to increase with improving SNR (Figure 9). Next, 30 word mean performance and confidence ratings were plotted for each participant. These data were plotted to visualize confidence/performance functions for 21 normal hearing participants as illustrated in Figure 9. In each case a positive relationship between confidence and performance is apparent.

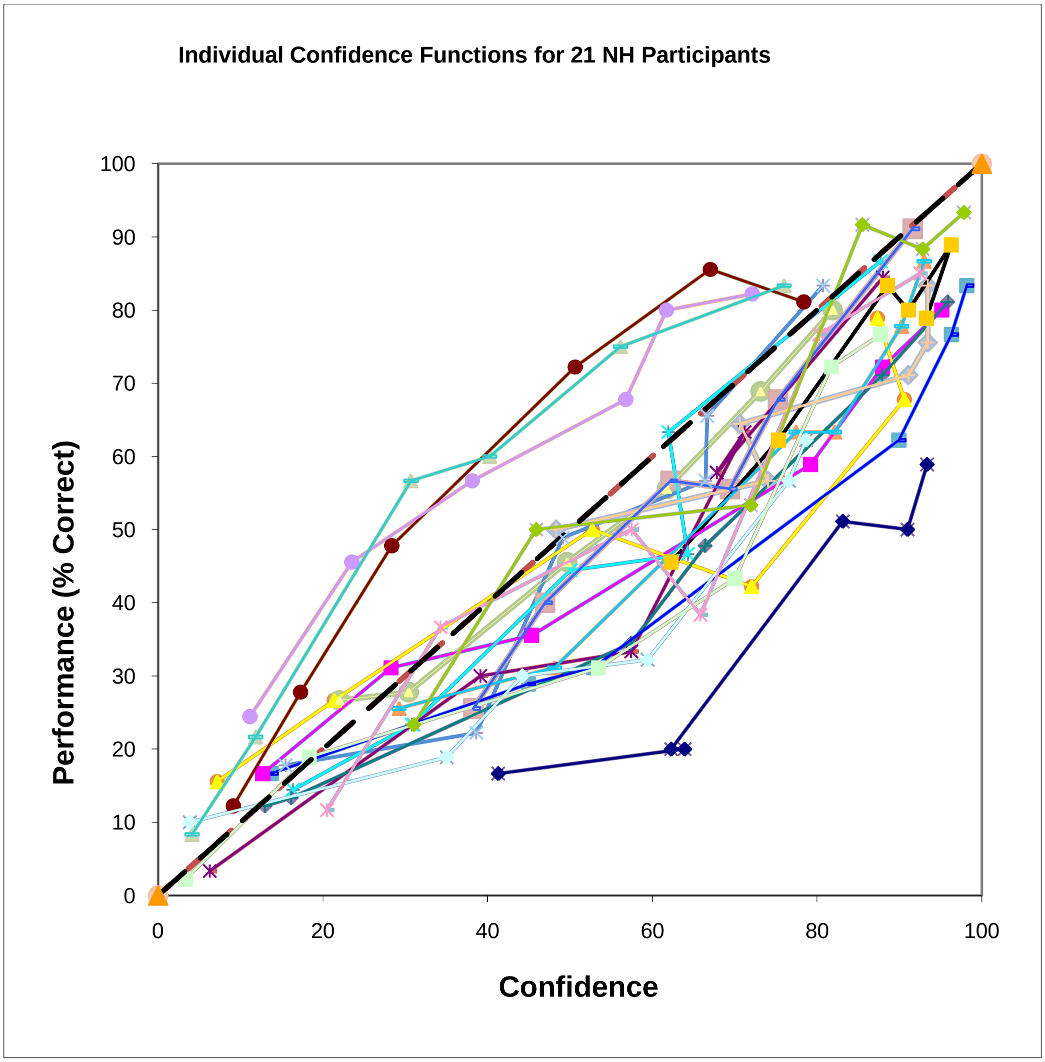

Figure 10. Mean confidence ratings vs. performance for 21 normal hearing participants, averaged across all 90 words presented at each SNR. Dashed line, ideal calibration.

Analysis of the raw data illustrated the large degree of variability in confidence observed between participants (Figure 10), Three distinct response patterns were noted. In order to better visualize these response patterns, overall calibration scores were calculated for each participant as the mean difference between confidence and performance. This value has been referred to as 'calibration in the large' (Baranski & Petrusic, 1994). The majority of participants were mildly overconfident in their responses. Scores that fell more than one standard deviation below the mean calibration in the large were labeled ‘under confident’ (n=3), those that fell within a standard deviation of the mean were labeled ‘realistic’ (n=15), and those that fell more than one standard deviation above the mean calibration score were considered to be ‘over confident’ (n=2). These groupings are illustrated in Figure 11.

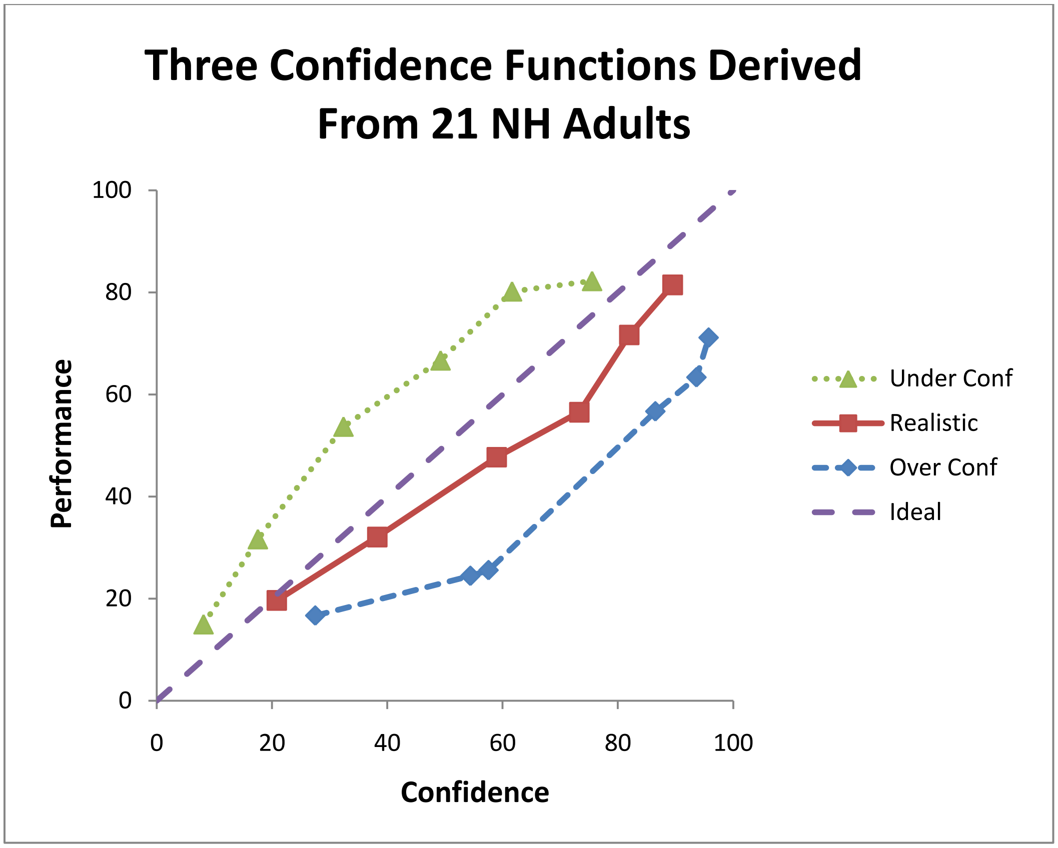

Figure 11. Three confidence functions derived from individual grand means (90-word performance and confidence averages) illustrating three unique response patterns on the confidence rating task.

*The performance confidence relationship*

The first approach employed to examine the relationship between confidence ratings and performance utilized the mean confidence rating within a block of 60 words (i.e., one SNR) compared to the proportion of responses within that block scored as correct. These data were plotted to compare confidence ratings to performance as illustrated in Figures 12 (ordinal scale) and 13 (visual analog scale).

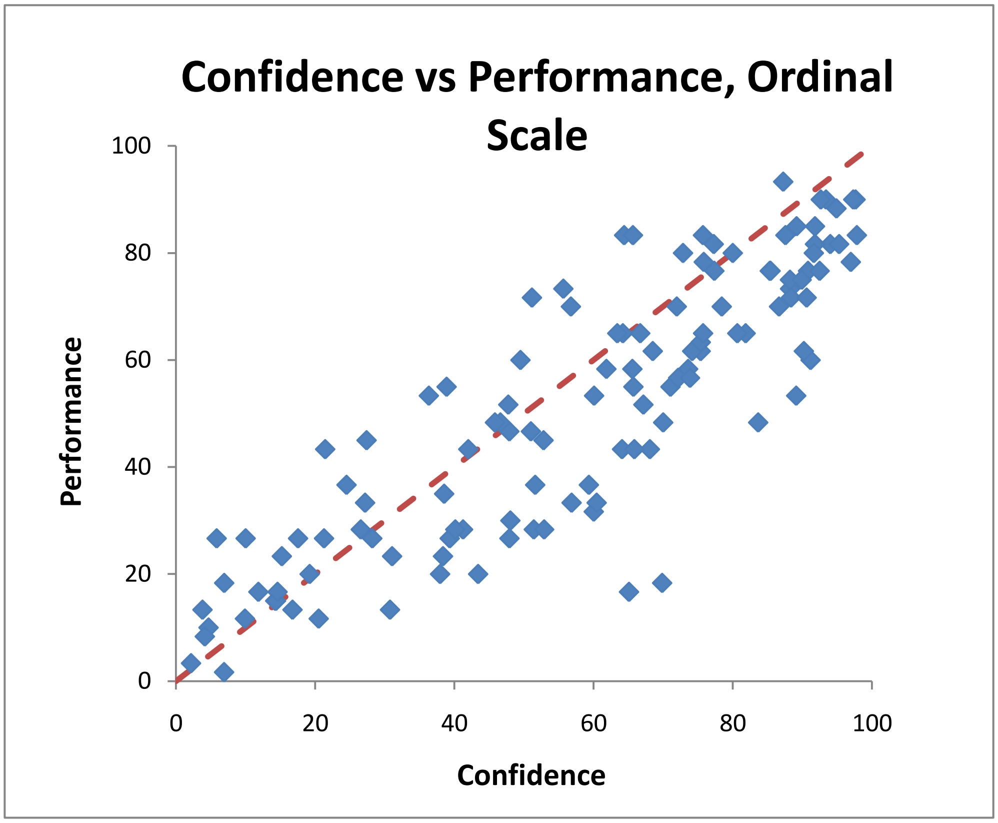

Figure 12. Confidence vs performance for 21 normal hearing adults using an ordinal scale. Diamonds, confidence/performance data. Dashed line, ideal relationship.

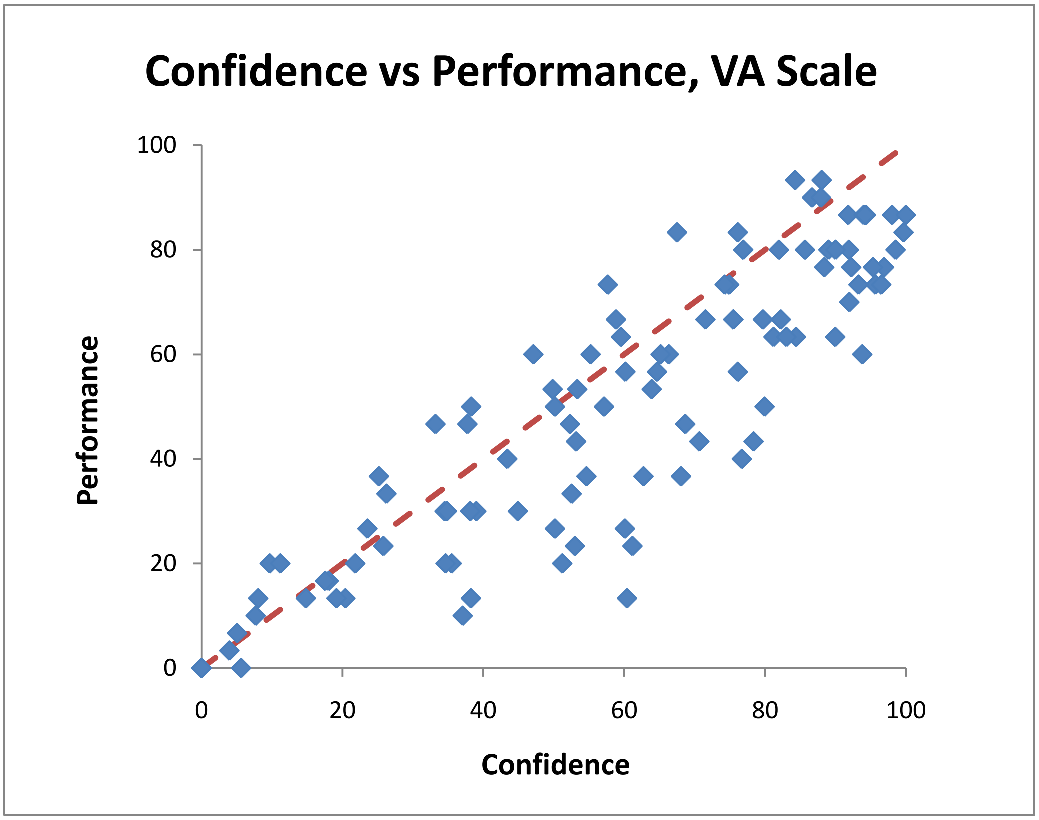

Figure 13. Confidence vs performance for 21 normal hearing adults, using a visual analog scale. Diamonds, confidence/performance data. Dashed line, ideal relationship.

Regression analyses were conducted on the ordinal and visual analog scale confidence rating data. Performance and confidence data was averaged within a signal to noise ratio such that each performance/confidence data point was based on 60 words. Initially, a simple linear regression was conducted to directly examine the performance/confidence relationship. Confidence rating was defined as the dependent variable, and performance as the independent (predictor) variable. These definitions were selected based on the central question of this study, namely, does performance predict confidence on a speech recognition task?

For the ordinal scale, data were analyzed via regression using performance as the predictor (i.e., independent) variable. The regression line was a good fit to the data ($R^2$*=0.75, p<.001*) and the overall relationship was significant ($F_{(1,118)}$*=349.05, p<.0001*).

For the visual analog scale, the regression line was also a good fit to the data ($R^2$*=0.85, p<.001*) and the overall relationship was significant. ($F_{(1,118)}$*=677.645, p<.0001*). These results suggest that performance accounts for approximately 75% of the variability in confidence when using the ordinal scale, and approximately 85% of the variability in confidence when using the visual analog scale.

In an effort to improve the prediction of confidence from performance data, multiple regression analyses were conducted, adding test SNR as a predictor variable to the parameters described in the linear regression above. This addition improved the fit of the regression line for the ordinal scale ($R^2_{adj}$*=0.785*), and remained significant ($F_{(2,117)}$*=218.768, p<.0001*). Similar results were observed for the visual analog data ($R^2_{adj}$*=.850,* $F_{(2,117)}$*=337.056, p<.0001*).

Interestingly, the regression results suggest differing contributions of SNR to the prediction of confidence for the two scales. In the case of the ordinal scale data, with performance held constant, confidence was positively related to SNR, increasing by 4.05 units for each dB of SNR improvement (*t=4.81, p>.001*). Conversely, the addition of SNR to the visual analog scale regression did not have a significant effect on the prediction of confidence (*t=-.573, p>.05*).

*Calibration and resolution measures of confidence ratings*

Confidence ratings were next considered from the perspective of accuracy of individual ratings. Individual Ratings were compared to performance by tabulating the proportion of responses scored as correct within a ten-point confidence range. That is, for each instance of a participant rating their confidence as 70-79%, what proportion of the time was the response scored as correct by the examiner? These data resulted in a plot of confidence versus performance for the ordinal and linear analog scales, as illustrated in Figure 14 .

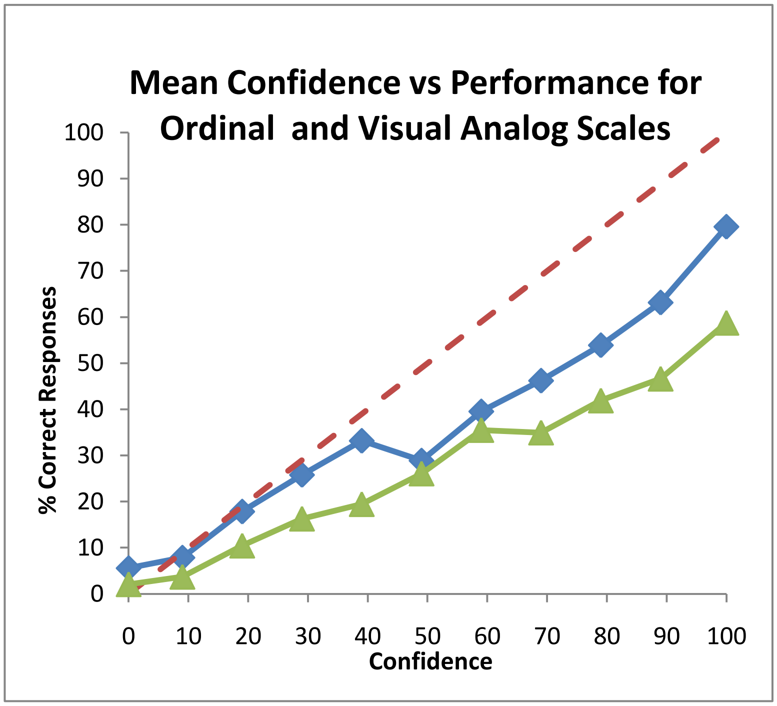

Figure 14 Mean performance for each confidence rating range for 21 normal hearing participants using ordinal (diamonds) and visual analog (triangles) scales. Dashed line; ideal confidence/performance relationship.

When rating confidence on the visual analog scale, the rating was substantially higher than the measured performance level. This trend was true of the ordinal scale only at performance levels greater than 30% correct. This result suggests substantial calculated over confidence in the correctness of responses for this scale. This result may have been influenced by the large number of 100% confident responses recorded when the keyword response was judged to be incorrect. This result was more common with the visual analog scale than with the ordinal scale.

*Calibration and resolution*

Calibration and resolution scores were calculated for ordinal and visual analog scales. Results calculated from the results of all participants are presented in Table 1.

Table 1. Calibration and resolution scores derived from aggregate confidence ratings, 21 normal hearing participants.

| | Ordinal Scale | | | Visual Analog Scale | | |
|---|---|---|---|---|---|---|
| | Run 1 | Run 2 | All | Run 1 | Run 2 | All |
| Calibration | 0.032 | 0.035 | 0.033 | 0.109 | 0.073 | 0.100 |
| Resolution | 0.072 | 0.085 | 0.077 | 0.040 | 0.075 | 0.044 |

As discussed above, calibration refers to the degree to which confidence reflects actual performance. An optimum value of 0 would suggest minimal dispersion of confidence ratings from measured performance. As can be seen in Table 1, for normal hearing participants, ordinal scale calibration values were superior to visual analog calibration scores when collapsed across performance conditions. Similarly, resolution values were superior for the ordinal scale.

*Confidence ratings and ratings of intelligibility*

To investigate the relationship between confidence ratings and ratings of intelligibility, data was sorted such that performance level could be matched between the confidence rating task and the intelligibility rating task. This allowed direct comparison of confidence and rated intelligibility for a given performance level. A repeated measures ANOVA was conducted comparing confidence responses to ratings of intelligibility for the normal hearing participants. The ANOVA was non-significant ($F_{(1,754)}=0.86, p>.05$) suggesting that for this sample, the ratings of confidence could not be distinguished from ratings of intelligibility. Correlation analyses suggested a significant positive relationship ($r=.85, p<.01$) between confidence and ratings of intelligibility for the normal hearing participants.

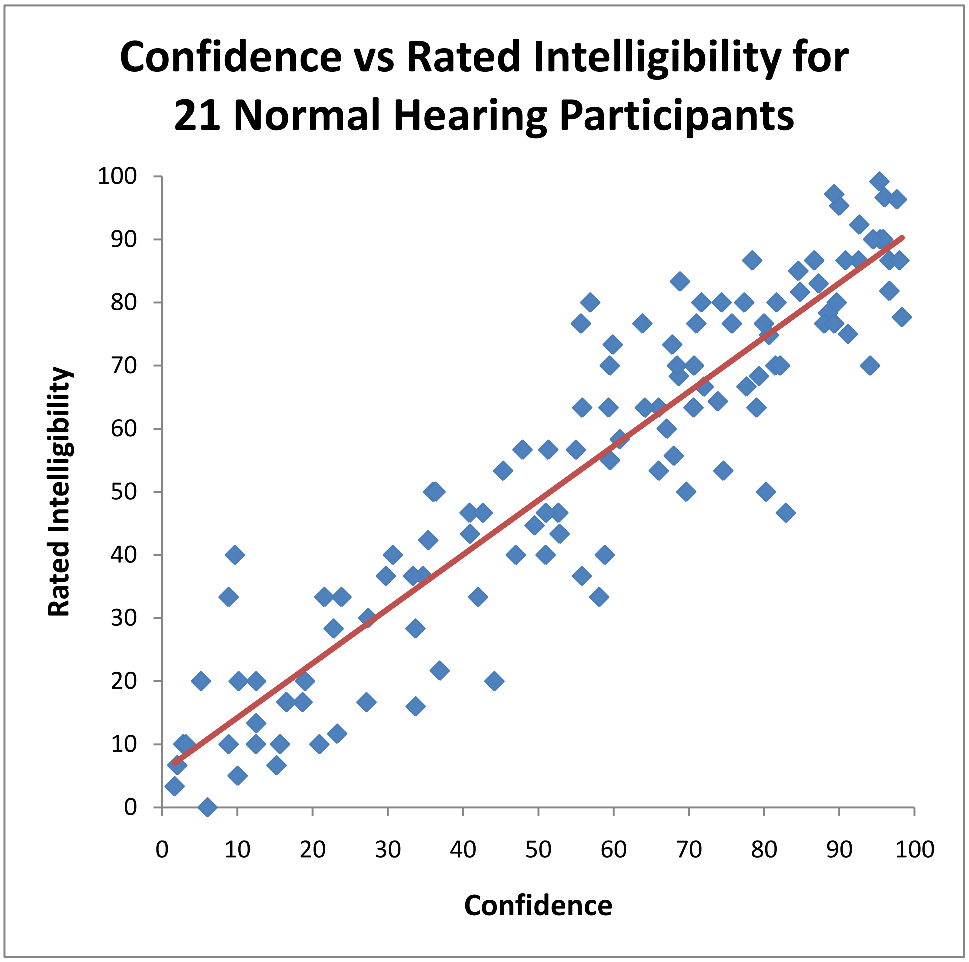

Figure 15. Scatterplot of rated confidence and rated intelligibility for normal hearing participants. Legend: Solid line = correlation trend line.

*Ratings of intelligibility and measured performance*

Ratings of intelligibility were collected for all participants and compared to measured performance. Rated intelligibility was found to be significantly correlated with performance (*r=.80, p<.001*). This relationship suggests that as performance increases, rated intelligibility should increase. A scatterplot of rated intelligibility versus performance for normal hearing participants is presented in Figure 16.

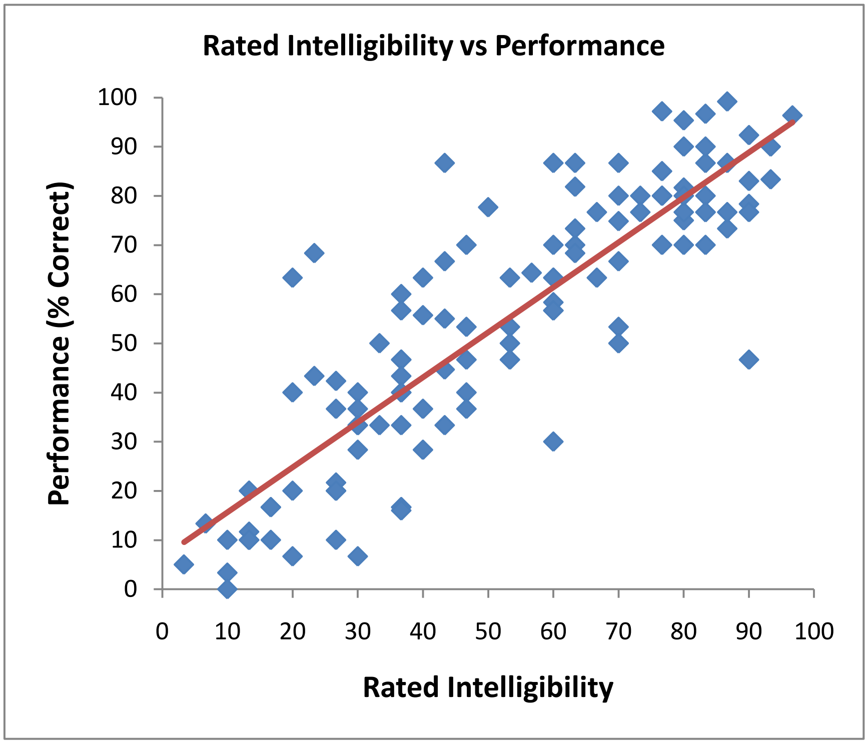

Figure 16. Scatterplot of rated intelligibility and performance for 21 normal hearing participants.

# Experiment 2

The purpose of experiment two was to investigate the effects of sentence context on confidence ratings of speech intelligibility. Contextual information eases speech perception in noisy situations due to the increased predictability of key words in the signal. It was hypothesized that increased context would result in improved calibration due to an increase in the predictability of the keywords. Towards this goal, sentences from the Revised Speech In Noise (R-SPIN) test (Kalikow, Steven & Elliot, 1977) were selected as the test stimulus. Each R-SPIN list contains fifty sentences, twenty-five of which are considered to have high predictability due to context. The remaining twenty-five have low predictability due to low context. Performance was varied across a wide range through manipulation of the signal to noise ratio. One list was presented at the 50% correct SNR established during experiment one, while two additional lists were presented at this SNR plus and minus 3 dB, respectively. The ordinal confidence scale previously described was used for participant ratings of confidence, as illustrated in Figure 2. Participants rated their confidence in the correctness of their repetition of the sentence keyword (i.e, the last word of each sentence).

# Results

### *Participants*

Twenty-two participants completed experiment two. Please see the participants section of experiment one for details.

### *Audiometric Data*

The audiometric thresholds of all participants were within the normal range. Pure tone thresholds were measured as previously described. See Figure 4 for mean pure tone thresholds of each group.

## Contextual Information and Confidence

### *SNR, performance and confidence*

For the twenty-two participants with normal hearing, 3300 ratings of confidence at SNRs between -7 and +2 dB were obtained. Mean confidence and performance (i.e., percent correct responses) were calculated for each participant in each test condition. These data were plotted versus SNR as illustrated in Figure 17. As expected, as SNR improved, performance and confidence increased for both normal hearing and hearing impaired participants. Participants achieved a wide range of performance as SNR was varied, and confidence ratings were observed to overlap the range of performance at each SNR.

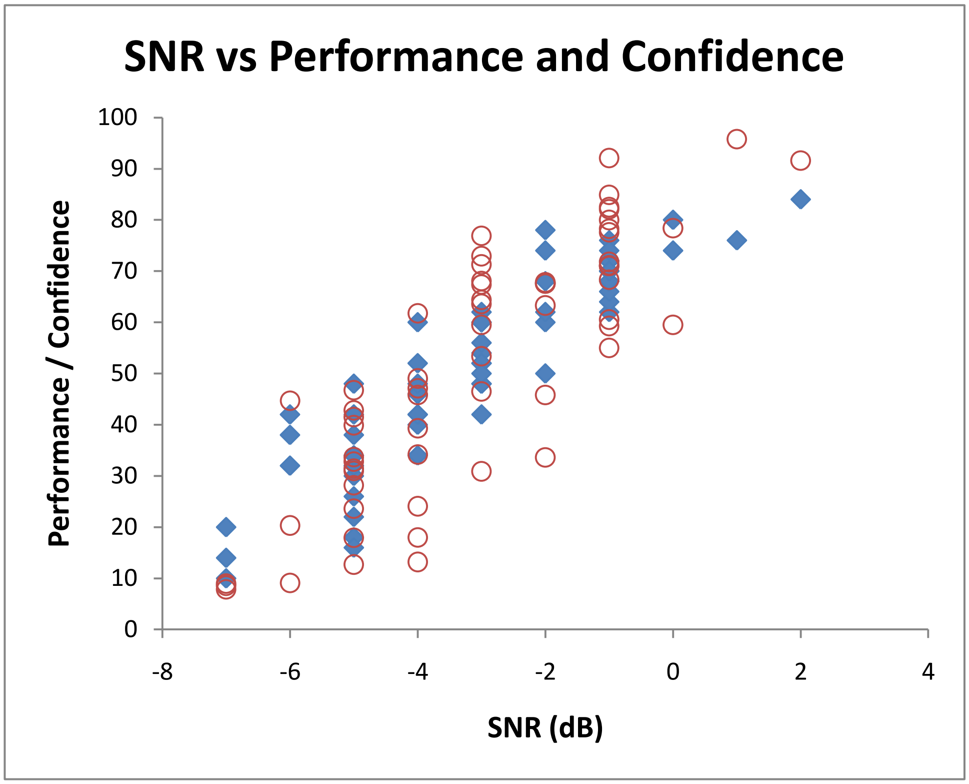

Figure 17. Mean confidence ratings and associated performance score for SNRs from -7 dB to +5 dB for normal hearing participants. Filled diamonds = performance, circles = confidence.

*Performance and confidence*

Next, the relationship between performance and confidence was investigated. A regression analysis was conducted on the R-SPIN performance confidence data. Performance was defined as the independent variable, and confidence as the dependent variable, again due to the experimental question as to whether performance predicts confidence. The linear regression of confidence rating on performance was significant for the normal hearing group ($F_{(1,61)}=125.248,\ p<0.0001$). The correlation between performance and confidence was significant ($R=0.820,\ p<.0001.$)

*Confidence and context*

Confidence and performance data were analyzed via multiple regression, using performance and SNR as predictor variables.

For the low context condition, the regression line was a fairly good fit ($R^2_{adj}=0.594, p<.001$), and the overall relationship was significant ($F_{(2,60)}=31.57, p<.0001$). With performance held constant, confidence was positively related to SNR, increasing by 8.5 units for every dB of SNR ($t=4.4, p<.001$). With SNR held constant, confidence was not significantly related to performance ($t=-0.01, p>.05$).

For the high context sentences, the regression line was an excellent fit ($R^2_{adj} =0.85, p<.001$), and the overall relationship was significant ($F_{(2,60)}=179.68, p<.0001$)*.* With performance held constant, confidence was not significantly related to SNR ($t=1.72, p>.05$). With SNR held constant, confidence was positively related to performance, increasing confidence by 0.88 units for each unit of increase of performance ($t=6.69, p<.001$).

Response frequency data revealed a larger proportion of high confidence ratings for the high context sentences than was observed in the low context sentence rating data. That is, a larger proportion of responses were considered to be ‘high confidence’ than was observed in the low context condition, regardless of actual performance level. Finally, regression coefficients were compared between high and low context conditions. Coefficients were observed to differ between the low and high context conditions ($t=4.65, p<.001$) suggesting differing growth of confidence functions between the two context conditions. The regression coefficients suggested that changes in confidence due to performance changes should be more dramatic with low predictability stimuli than when using higher predictability stimuli. These results were confirmed in the following analysis.

As described in experiment 1, confidence ratings were compared to performance by tabulating the proportion of responses scored as correct for a given confidence range. Briefly, the question of interest was, ‘for a 10% range of confidence ratings, what proportion of the time were the responses scored as correct by the examiner?’ These data resulted in a plot of confidence versus performance, as illustrated in Figure 18.

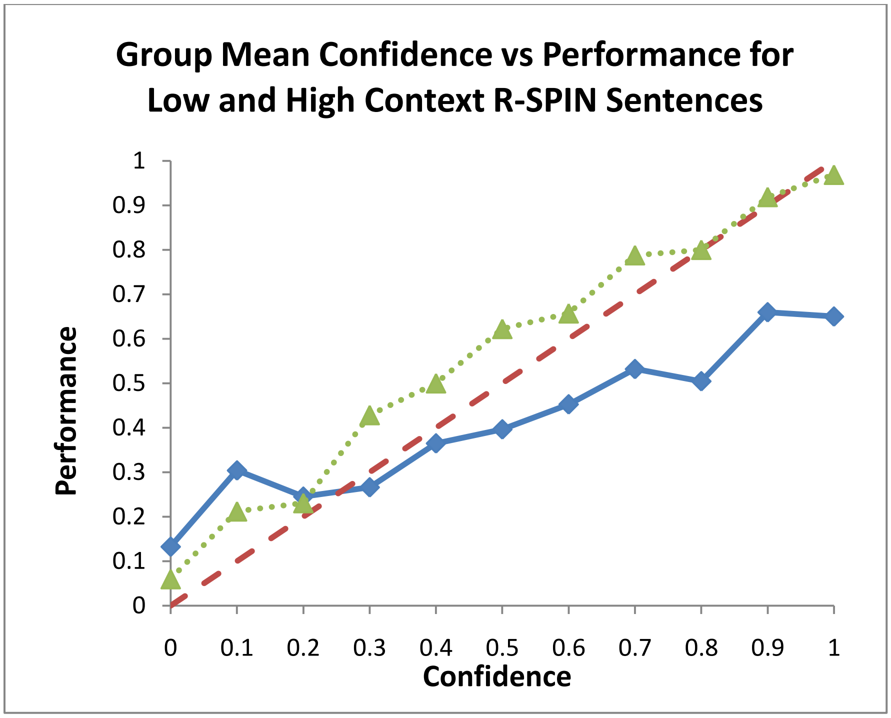

Figure 18. Confidence/Performance relationship for low and high predictability keywords presented in R-SPIN carrier phrases. Diamonds, low predictability keywords. Triangles, high predictability keywords. 22 Normal hearing participants.

Data were obtained that encompassed a large range of performance for both low and high context conditions. For low context sentences, the data suggest strong over confidence in response correctness, whereas for the high context sentences, participants were typically under confident in their responses. The high context data differed markedly from confidence ratings collected in experiment one using NU-6 monosyllabic words, where for most performance levels, participants were over confident in their responses. These results suggest that in low context sentences, participants were more likely to formulate a guess and assign a high confidence rating, as opposed to a more conservative confidence rating approach for sentences with context. Addition of context to the test sentences likely allows individuals to weigh the likelihood of a response being correct based on how it fits with better detected parts of the sentence. Alternatively, the addition of context may make it more apparent when the perceived keyword in correct. Either of these possibilities would reduce the number of guesses and preclude the appearance of overconfidence as seen in the low context sentence responses.

### *Calibration and resolution scores*

Calibration and resolution scores were calculated for low and high context R-SPIN sentence confidence ratings and performance. Results are presented in Table 2.

Table 2. Assessment parameters for R-SPIN sentence performance/confidence ratings, 21 Normal Hearing Participants.

| | **Low Context** | **High Context** | **All Conditions** |
|---|---|---|---|
| Performance | 0.3560 | 0.6832 | 0.5197 |
| Calibration | 0.0299 | 0.0029 | 0.0086 |
| Resolution | 0.0291 | 0.1352 | 0.1008 |

Calibration values were observed to be superior in the high context condition over the low context condition by a factor of ten. The observed calibration scores approach an optimum value of zero in the high context condition, suggesting that actual performance is predicted by confidence ratings. Similarly, resolution was superior in the high context condition. These results suggest that high context sentences resulted in the best calibrated responses to performance, and led to superior resolution scores. It is important to note however, that low predictability keywords elicited significantly faster growth of confidence with performance changes than did high predictability keywords. From this perspective, confidence ratings of low predictability sentences would appear to be a more sensitive metric of changes in performance than the other stimuli utilized in these studies.

# CHAPTER IV

# DISCUSSION

Initial analyses described the relationship between confidence and performance on a speech in noise task involving monosyllabic words presented in a background of multitalker babble. Confidence ratings were affected by performance level, measurement scale and stimulus context. As performance increased, confidence rating increased. However, the relationship between these factors varied when using the two different confidence rating scales. Participants were more likely to rate their confidence highly when using the visual analog scale, resulting in a large number of relatively high confidence ratings when performance was very low. Accordingly, the visual analog scale resulted in poorer test-retest reliability than the ordinal confidence scale. In contrast, the proportion of high confidence ratings was much lower when participants used the ordinal scale, leading to a more realistic judgment of performance, and therefore improved calibration. Stimulus type appeared to affect confidence ratings. NU-6 monosyllabic words resulted in higher confidence ratings for a given performance level than did R-SPIN sentences. Similarly, context appeared to play an important role in the calibration of confidence ratings. High context sentences resulted in better calibrated and therefore more realistic confidence ratings. While the exigent confidence literature would argue that improved calibration is a desirable outcome (i.e., confidence scores accurately predict performance), overconfidence and the attendant decline in calibration may allow detection of perceived changes in performance before actual changes in performance can be measured. For this reason, low predictability stimuli would appear to provide at least one advantage over stimuli that elicit better calibration to performance.

### *The Role of Performance*

Measured performance level influenced confidence ratings. As hypothesized, confidence and performance were shown to vary systematically with signal to noise ratio. However, when using monosyllabic words or low context sentences as stimuli, confidence rating typically increased at a rate much greater than would be expected based on measured performance, resulting in over confidence relative to measured performance. This result is in contrast to confidence data previously published regarding other perceptual tasks. For example, early studies in perceptual confidence dating to the late 1800s suggested that individuals are typically underconfident in their perceptual judgments (Baranski & Petrusic, 1994). In the case of the tasks utilized in this study, respondents were more commonly overconfident in their responses. This was particularly true of low context stimuli in conditions that elicited higher performance (i.e., ‘easier’ conditions). It was in these conditions that the performance/confidence disparity tended to be greatest. This result contrasts with previously published data that suggests that individuals tend to be overconfident in their responses under particularly difficult conditions (e.g., Adams & Adams, 1961; Bjorkmann, 1994; Baranski & Petrusic, 1994). An alternate interpretation however, is that the relative difficulty of the low context stimuli has a greater effect on confidence than does actual performance. If this interpretation is correct, it is interesting to consider that making the perceptual situation more difficult results in over confidence, a result that is completely contrary to the common-sense idea that increased difficulty would lead to decreased confidence.

### *The Role of Context*

The degree of context provided in the stimulus influenced confidence ratings. As hypothesized, increased context in the test stimulus sentence led to both higher performance and confidence ratings. From a performance perspective, this finding is in good agreement with previously published results (e.g., Pickett and Pollack, 1963; Kalikow et al., 1977). However, it was interesting to observe that the regression coefficients differed significantly between the two confidence/performance functions, suggesting that confidence varied at a different rate when contextual information in the test sentence was changed. Importantly, confidence ratings for low context stimuli show potential as an outcome assessment tool for users of hearing aids as they may help to reveal differences in perceived performance not previously detectable with traditional tests of speech intelligibility. That is, individuals may report an improvement in confidence without a detectable improvement in performance, particularly for low context stimuli. Similarly, regardless of performance level, confidence was rated significantly higher in the low context condition despite the fact that performance for this condition was consistently lower than the high context condition. This finding was mirrored in the differences observed in calculated calibration values between the low and high context conditions. Participants tended to be over-confident in the low context condition, and slightly underconfident in the high context condition. Ratings of confidence were therefore more realistic and better calibrated in the high context condition, where knowledge of language and grammar provided the greatest benefit. This observation suggests that when more information is available, individuals tend to respond with greater caution and deliberation than when little information is present for use. These results are in agreement with previous general knowledge based confidence studies (e.g., Bjorkmann, 1994; Kroner & Berman, 2007) The clinical manifestation of this effect may be observed with patients who are clearly experiencing difficulty understanding speech in novel or difficult listening situations, yet report little disability or handicap compared to when listening to television, radio or other somewhat predictable stimuli. As previously mentioned,  perceptual tasks with a higher degree of difficulty tend to produce overconfidence. It is clear that a similar effect was elicited with speech intelligibility stimuli in this study.

### *The Role of Measurement Scale*

Confidence ratings were shown to be significantly better calibrated (i.e., exhibited better agreement with measured performance) when ratings were performed on an ordinal scale than when using a visual analog scale. It had been hypothesized that the visual analog scale would result in more accurate, (i.e., better calibrated) responses due to the lack of constraints imposed by the numeric and written markers on the ordinal scale. Instead it appeared that participants were better able to assign a value to their feelings of confidence when provided with comparative markers and a numeric scale than when simply provided with endpoint markers.

## Rated Intelligibility vs. Confidence

Ratings of intelligibility were not significantly different from ratings of confidence. Both measures were positively correlated with performance, repeatable and sensitive to changes in the difficulty of the listening condition. It is possible that confidence ratings represent an alternate technique for the estimation of perceived intelligibility. However confidence ratings exhibit at least one advantage over ratings of intelligibility. Confidence ratings do not require that participants transfer data regarding ongoing performance to longer term memory to arrive at an estimate of performance. Instead, ratings are performed after each test item, allowing an immediate impression of performance to be recorded within seconds of experiencing the stimulus. The ‘scoring’ process is performed by the examiner at the end of each block of stimuli. Since averaging is performed by the examiner, memory requirements for the participant are minimal. Conversely, when performing a rating of intelligibility at the end of block of sentences or a test passage, the individual must access memory of their performance on individual items or throughout the test passage to generate an estimate of performance. This process may prove difficult for some, leading to over or underestimation of performance at the conclusion of the test block as reported by Preminger and colleagues, (2000). However, other authors have reported better test re-test reliability than was observed in the Preminger et al (2000) study. For example, Saunders and Cienkowski (2002) argued that the disparity in reliability of intelligibility ratings reported in previous studies was likely due to differences in instruction sets. In this study, ratings of confidence resulted in acceptable test-retest reliability for both ordinal and visual analog scales, similar to results observed with ratings of intelligibility by the aforementioned authors.

## Implications for Future Research and Practice

In the following section, the implications of this study for clinical practice and future research are discussed. These implications draw upon previous research and the present interpretations of the current study. First, potential extensions of the current study are described. Second, suggestions are made for the application of confidence ratings in the assessment of outcomes in patients fit with amplification as treatment for hearing loss.

### *Extensions of the Present Study*

The addition of a matched size group of hearing impaired participants would allow direct comparison of results to the normal hearing group. As it has now been demonstrated that confidence in performance can be measured, it follows that the test instruments developed in this study should be investigated with hearing impaired participants.

Second, the question must be raised of whether confidence response patterns produced by normal hearing participants would be reproduced in a hearing impaired group.

Third, comparison of confidence ratings obtained in aided and unaided conditions at fixed performance levels should be investigated. That is, does the provision of amplification in cases of hearing loss affect the ratings of confidence volunteered by participants? It would also be interesting to investigate the effects of degree and configuration of hearing loss on baseline confidence in quiet and in varying degrees of background noise. The present study varied SNR across a wide range, but for some participants, basement and ceiling performance was not achieved. Therefore, expanding the range of test conditions would provide a more complete picture of the performance/confidence relationship in both normal hearing and hearing impaired populations. Finally, gender effects cannot be effectively explored within the scope of this study due to the predominately female sample recruited for the study.

### *Applicability to Clinical Practice*

Most clinicians have encountered patients with hearing loss who are confident in their communication abilities despite the complaints and concerns of family members and friends. It is equally common to encounter patients with perceived gains in performance following the provision of amplification that cannot be duplicated in the test booth. Preferences for particular hearing aid settings have also reported that sometimes cannot be attributed to differences in audibility (e.g., Horwitz, Turner & Fabry, 1991). It is possible that the rapid growth of confidence in response to small improvements in performance exhibited by the participants of this study when using low predictability stimuli could be harnessed to help differentiate between the utility of features and settings of amplification. For example, patients may express greater confidence when fit with a particular set of fitting parameters despite published research suggesting that the features or settings do not provide measurable benefit in speech intelligibility. It is certainly possible that clinician scientists are attempting to measure a different parameter than the factor that the individual perceives a change in, or that the change is too small to be detected with current speech intelligibility tasks. Confidence ratings may offer some insight into these patient preferences.

Classic rated intelligibility tasks have been shown to vary in their correlation with measured speech intelligibility performance. It is certainly possible that a poor correlation between rated intelligibility and performance could be attributed to the inability of the individual to perform the complex averaging task of rating the proportion of keywords repeated correctly due to memory, cognitive and/or instructional problems. A possible solution to this problem was described by Saunders and Cienkowski (2002), who utilized a modified HINT procedure to measure both perceived and actual performance thresholds for the purpose of comparison in assessment of hearing aid outcome. The procedure utilized in the current study resulted in a new method of assessing perceived performance though the use of confidence ratings and a simple response scale. The comparison of measures of confidence in performance to actual performance on a binary task as described in this study may help in the identification of individuals with unrealistic appraisals of their own performance, and to distinguish them from individuals who withdraw from conversation not due to an inability to perform, but rather a fear of failure.

This measure shows potential in the assessment of outcomes for hearing aid interventions, features, styles and program settings. Correlation coefficients obtained using both confidence assessment scales suggest that participants are likely to report improvements in confidence at a rate greater than improvements in actual performance. This suggests that this measure may be sensitive to small advantages gained through signal processing or other hearing aid features. It is not clear at this juncture what difference in perception leads individuals to increase their ratings of confidence at a rate greater than that of the actual increase in performance, but similar trends have been reported in other sensory modalities. As noted in the results section above, improvements in performance led to an almost two-fold increase in confidence ratings when using low context sentences and monosyllabic words as test stimuli. Therefore, in order to detect small differences in performance, it would appear that minimizing knowledge in the test materials (e.g., removing any priming clues from the test stimuli) would result in the largest changes in confidence for a given change in performance. This measure would be most useful in assessing outcomes with amplification as it would theoretically allow the quantification of otherwise unexplainable preferences.  Conversely, if an optimally accurate assessment of the improvement in performance is desired more than a large change in confidence rating, it would be more appropriate to utilize a test stimulus with high context or predictability similar to the high context R-SPIN sentences described above.

As performance has been shown to account for approximately 75% of the variability in confidence in the participants of this study, it may be that confidence measures a component of perceived benefit that is overlooked in more traditional measures of outcome.

## Caveats

### *Participants*

The group of participants described in this study was not selected to be representative of the normal hearing population. Rather, normal hearing participants were recruited from a sample of young, highly educated, predominately female participants. Further, it cannot be ascertained to what degree motivation varied within the sample, but there was clearly variability in participant enthusiasm for the experimental tasks.

### *Methods*

The NU-6 based test stimuli used in the study were repeated several times throughout the course of the study. Learning effects were considered unlikely due to the duration of time between presentations of the stimuli. Statistical measures (two factor repeated measures ANOVA) suggest that there were no significant performance differences between runs 1 and 2 for the ordinal ($F_{(1,118)}=0.22, p>.05$) or visual analog ($F_{(1,118)}=0.16, p>.05$) scales. Nevertheless, in isolated cases it is possible that individuals with exceptional memories could learn and recognize words due to repeated exposure to words in a particular order. Presentation of different randomizations of the words, or utilizing a larger set of phonetically balanced words as test stimuli would minimize learning effects. However, this could also result in the loss of the ability to directly compare performance and confidence ratings obtained with different rating scales. If this comparison were not the goal of subsequent studies, unique test stimuli would ensure that no learning effects for stimuli were present in the test data.

In the high versus low context task, confidence ratings were obtained across a relatively wide range of performance levels utilizing only three SNRs. It would be a significant improvement to the study to utilize a wider range of performance levels by varying SNR to a larger degree than was accomplished in this study. This would allow the generation of a more complete performance/confidence function for high and low context stimuli.

## Summary

The purpose of this study was to conduct an initial exploration of the relationships between confidence, performance and perceived performance for monosyllabic words and sentences with either high or low context. Researchers in other perceptual areas have demonstrated strong correlations between response confidence and performance; however this relationship has not been explored to date using a speech perception task. The study also sought to differentiate between ratings of confidence and ratings of intelligibility, and to examine the effects of sentence context on confidence and performance.

In experiment 1, the relationship between measured speech understanding performance and communication confidence ratings for monosyllabic words was investigated. Monosyllabic words were selected as this stimulus most closely approximates the binary qualities of stimuli described in the literature on response confidence. That is, each response is rated individually, and the response is graded as either correct or incorrect. Thus, no averaging or complex processing of information need be conducted by the participant to arrive at an estimate of confidence, and no interpretation is required of the examiner. Performance was shown to account for approximately 75% of the variance in confidence. Second, test-retest reliability and consistency of communication confidence ratings was investigated, revealing strong positive correlations between test and re-test data. These areas of investigation were considered to be of interest due to the lack of experimental evidence that word recognition performance predicts response confidence, and due to the practical requirement that measures of outcome should be stable and reliable. Third, analyses were conducted to determine whether communication confidence ratings differ from ratings of intelligibility, a finding which would suggest that confidence ratings constitute a unique dimension of personal listening experience that is potentially useful in the assessment of the impact of amplification. Ratings of intelligibility were found to be indistinguishable from ratings of confidence in this experimental sample. Fourth, calibration and resolution scores were compared between ordinal and visual analog scales, revealing superior calibration and resolution for the ordinal scale.

In experiment 2, the relationships between stimulus sentence context, performance and confidence ratings were explored. It has been suggested by several authors that knowledge positively affects calibration (e.g., Adams & Adams, 1961; Lichtenstein & Fischoff, 1977; Koriat, Lichtensteian & Fischoff, 1980). Towards this end, this experiment made use of the Revised Speech Perception In Noise (R-SPIN) Test sentences (Kalikow, Steven & Elliot, 1977). As previously described, the R-SPIN test is composed of two types of sentences. The first contains contextual cues toward the keyword, while the second type does not. It was hypothesized that the addition of context to sentence-based speech recognition materials would positively affect calibration in a manner similar to that of knowledge. It was determined that high context stimuli resulted in improved calibration and resolution scores. Confidence ratings were compared between high and low context conditions, with the result that participants were found to be overconfident in their performance in the low context condition.

Disparities between confidence and performance in the real world lead to embarrassment, confusion and failures in communication. For these and other reasons, we would hope that individuals would exhibit the ability to judge when perception reflects reality, when a strong hunch is likely to be correct and when a wild guess is better left unsaid.

# CHAPTER V

# CONCLUSION

The findings of the present study present initial information regarding the relationship between confidence and performance on two speech recognition in noise tasks. Confidence was shown by the participants of this study to be a scalable concept that is strongly correlated with measured performance. This outcome was observed when measuring confidence on either ordinal or visual analog scales developed for the present study.

From the data collected in this study, it would appear that of the three stimulus types used, high context sentences led to the most accurate (i.e., best calibrated) relationship between performance and confidence ratings.

Conversely, low context sentences resulted in a greater change in confidence rating with performance, suggesting that this stimulus may help to uncover perceived differences in performance that are not easily detectable using more conventional methods of assessment. Regardless, performance was observed to account for only 70-75% of the variability in confidence, suggesting that unknown factors besides performance contribute to the perception of confidence. Confidence ratings may yet prove to be a valuable tool in the assessment of outcomes related to hearing aid intervention.

## A: Test Instructions – Experiments 1 and 2

### Confidence Ratings

"*You will be hearing a man's voice reading sentences in a background of noise. The noise will consist of several people talking simultaneously. While this background noise may be very distracting, please try to ignore it as much as possible. Try to concentrate your listening on understanding each sentence that is read. At the end of each sentence, there will be a brief pause in the noise. Please repeat the last word of the sentence. Try to repeat back exactly what you believe you heard. Some sentences will be relatively easy to understand, others will be more difficult. This process will be completed several times. In each case, try to repeat back the last word of each sentence. After we have practiced with a few sentences, we will add a new task. At the end of each sentence, you will rate your confidence in your response by making a mark on a line. This mark will be placed such that it indicates your confidence in your response along a scale between "not confident at all" and "very confident". For example, if you feel that you have no idea what the last word of the sentence was, you might place your mark on the left end of the line, indicating that you were 'not confident at all.' Conversely, if you are absolutely sure that you repeated the target word correctly, you should make a mark on the far right hand end of the scale, indicating that you were 'very confident' in your response. If you are somewhat confident that your response was correct, but not absolutely sure in either way, you should make a mark somewhere between the two extremes as you feel is appropriate. We will also practice this task prior to beginning the trials."*

### Rated Intelligibility

"*You will be hearing a man's voice reading sentences in a background of noise. The noise will consist of several people talking simultaneously. While this background noise may be very distracting, please try to ignore it as much as possible. Try to concentrate your* *listening on understanding each sentence that is read. At the end of each sentence, there will be a brief pause in the noise. Please repeat the last word of the sentence. Try to repeat back exactly what you believe you heard. Some sentences will be relatively easy to understand, others will be more difficult. This process will be completed several times. In each case, try to repeat back the last word of each sentence. After we have practiced with a few sentences, we will add a new task. At the end of each group of sentences, you will rate the proportion of the sentence keywords that you believe you repeated correctly by making a mark on a line. This mark will be placed such that it indicates the proportion of keywords repeated correctly on a scale ranging from zero to 100. For example, if you feel that you repeated about ten percent of the words correctly, you would place the mark at the left side of the line near the scale marker '10'. Conversely, if you are absolutely sure that you repeated all of the target words correctly, you should make a mark on the far right hand end of the scale, indicating that you believe that you repeated 100% of the words correctly. If you believe that your performance was somewhere between those extremes, you should make a mark somewhere between the two extremes as you feel is appropriate. We will also practice this task prior to beginning the trials."*

## B. Hearing Impaired Group Confidence Pilot Experiment

Five hearing impaired participants (3 males, 2 females, mean age, 74.4 years SD, 9.29 years) were recruited to complete the test procedures outlined in the study above, These participants were recruited in an effort to determine whether the test procedures and scales developed for the above study could be effectively utilized by individuals with mild to moderate high frequency sensorineural hearing loss.

### *Audiometric Data*

Mean audiometric thresholds for the five hearing impaired participants are displayed in Figure A.

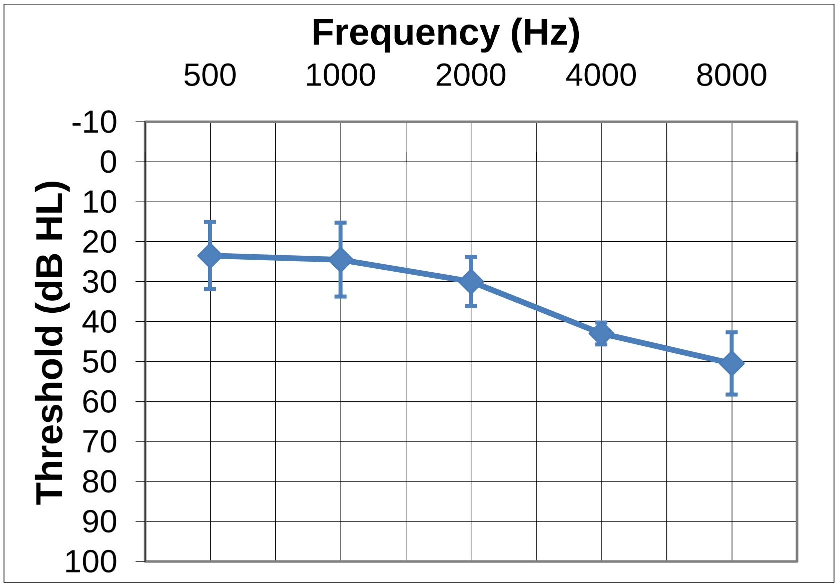

Figure B-1. Mean audiometric data and standard deviations for five hearing impaired participants.

### *Research Questions & Hypotheses*

See the research questions and hypotheses detailed in the main study above.

## RESULTS

## Experiment 1

### *Confidence and performance*

Linear regression analyses were conducted on the visual analog and ordinal scale confidence rating data obtained from the hearing impaired participants. Data was averaged across each SNR (i.e., 60 words) within participants, resulting in 30 performance/confidence data pairs per condition. Performance was once again defined as the independent variable, and confidence rating as the dependent variable. For the ordinal scale, performance was shown to account for 72% of the variance in confidence. ($R^2 = 0.72, p<.001$) and the regression of confidence on performance was significant ($F_{(1,28)}=73.75, p<.0001$). For the visual analog scale, performance accounted for 76% of the variability in confidence and the linear regression of confidence on performance was also found to be significant ($F_{(1,28)}=96.33, p<.0001$).

The hearing impaired participants exhibited conservative confidence ratings when performance was less than 50% correct. At higher performance levels the hearing impaired participants responded similarly to the normal hearing participants with over-confident ratings of performance. These results were observed for both ordinal and visual analog scales (Figure B-2).

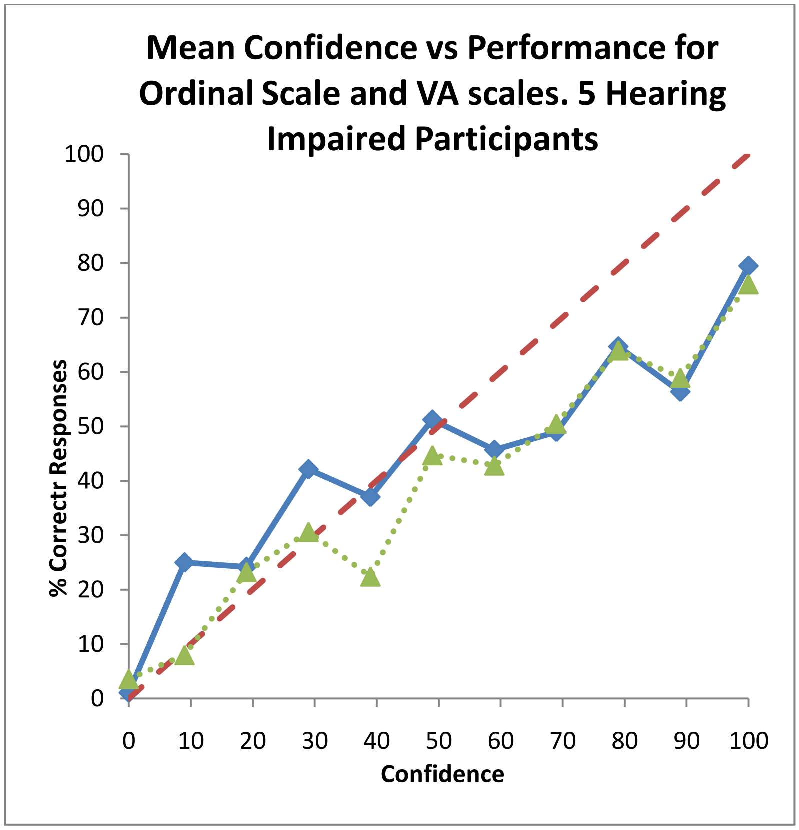

Figure B-2. Mean performance for each confidence rating range for 5 hearing impaired participants using ordinal (diamonds) and visual analog (triangles) scales. Dashed line; ideal confidence/performance relationship.

### *Confidence Test-Retest*

Confidence ratings on run 1 were observed to account for 60% of the variance in run 2. The regression was observed to be significant ($F_{(1,28)}=18.76, p<.001$).

### *Rated Intelligibility and Confidence*

Rated Intelligibility and confidence data were analyzed via multiple regression. Confidence ratings were observed to account for only 6% of the variance in rated intelligibility. The overall regression was observed to be significant ($F_{(2,87)}=3.31, p<.05$).

### *Calibration and Resolution*

Calibration and resolution scores were calculated for ordinal and visual analog scale data. Results suggested improved calibration and resolution for the ordinal scale, however given the small sample size of the hearing impaired group, statistical comparisons were not conducted.

| | **Ordinal Scale** | | | **Visual Analog Scale** | | |
|---|---|---|---|---|---|---|
| | Run 1 | Run 2 | All | Run 1 | Run 2 | All |
| Calibration | 0.02 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 |
| Resolution | 0.07 | 0.12 | 0.08 | 0.06 | 0.05 | 0.06 |

Figure B-3. Calibration and resolution scores for 5 hearing impaired participants.

## Experiment 2

### *Confidence and performance*

As described in experiment 1, confidence ratings were compared to performance by tabulating the proportion of responses scored as correct for a given confidence range. Again, the question of interest was, ‘for a 10% range of confidence ratings, what proportion of the time was the response scored as correct by the examiner?' These data resulted in a plot of confidence versus performance, as illustrated in Figure B-4 for the hearing impaired participants. Data were obtained that encompassed a large range of performance for both low and high context conditions.

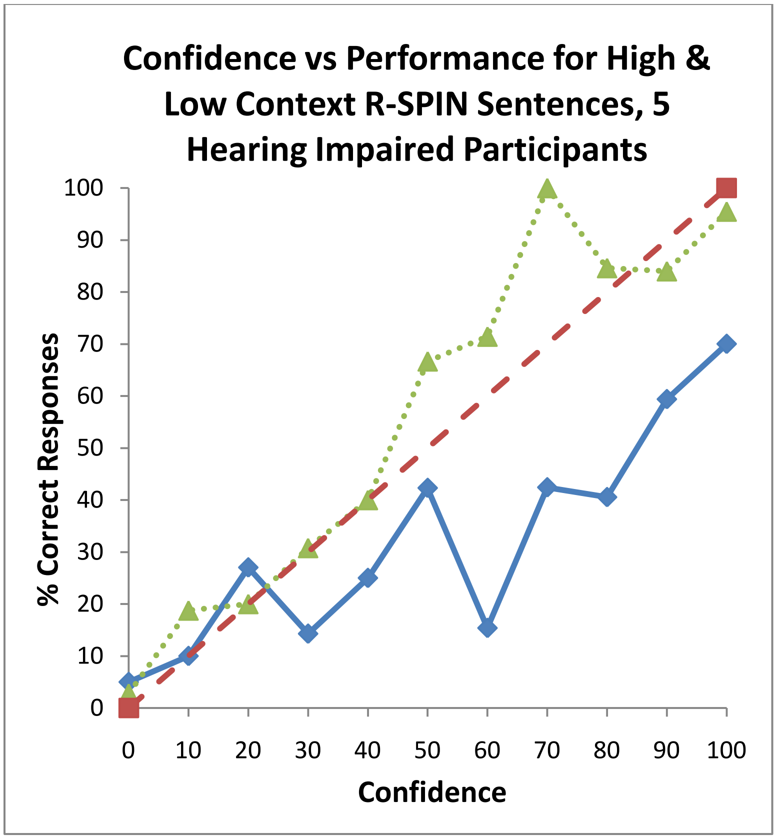

Figure B-4. Confidence vs. performance plots for high context (triangles) and low context (diamonds) R-SPIN sentences. 5 hearing impaired participants.

### *Context and confidence*

These data differed markedly from those collected previously for NU-6 monosyllabic words. For low context sentences, the data suggested strong over confidence in response correctness, whereas for the high context sentences, participants were typically well calibrated to their measured performance level.

### *Discussion*

Results of experiments 1 and 2 were grossly similar to those obtained from normal hearing participants. These experiments were conducted to ascertain whether hearing impaired individuals would be able to complete the confidence rating task using the test stimuli and presentation levels used for normal hearing participants. It appears that the five pilot participants in this study were able to make appropriate use of the confidence rating scales, and were able to perform the speech intelligibility tasks. Performance and confidence were shown to vary along with SNR, as expected.

These results suggest that further research should be conducted with hearing impaired participants in order to allow direct comparison of performance/confidence functions between normal hearing and hearing impaired groups.

### *Directions for Future Research*

Comparisons of confidence ratings should be obtained in both aided and unaided conditions. This would allow investigation of the effects of amplification on confidence ratings.

## C: Pilot Experiment

A pilot study was conducted in an effort to determine whether the difficulty encountered in various listening conditions could affect communication confidence ratings. In the experiments, signal-to-noise ratio (SNR) was varied, as were reverberation time, presentation level and signal bandwidth. Conditions of reduced signal to noise ratio were considered to be more difficult, whereas conditions of increased SNR were considered to be less difficult. Similarly, longer reverberation times were assumed to be more challenging than shorter reverberation times (Gordon-Salant & Fitzgibbons, 1993), low-pass filtered stimuli more challenging than wideband (ANSI, 1997), and lower listening level more challenging than higher listening level, based upon performance intensity function curves.

### *Participants*

Ten normal hearing young adults (6 females, 4 males, mean age =27.1 years, SD=3.38) were recruited to participate in the pilot study. Participants were informed that the study was designed to explore the concept of response confidence under a variety of listening situations. They were informed that they were free to withdraw from the study at any time without prejudice.

# Methods

## *Stimuli*

An adaptive speech intelligibility test procedure was performed using HINT sentences in the sound field in both a sound-treated audiometric test booth (considered to be non-reverberant) and in a reverberation test chamber (reverberant). Sentences and multi-talker babble were presented at a variable SNR from a Tannoy self-amplified studio monitor loudspeaker (model 800A). Signal and noise were presented at 0° azimuth under a variety of conditions as detailed below. For the low-pass filtered condition, HINT materials were digitized from the original compact disk test materials and saved to hard disk. The speech and noise files were filtered using Adobe Audition using a low-pass FFT 80dB per octave brick wall filter with a shoulder frequency of 1500 Hz. Stimulus level was adjusted to equalize RMS output between wideband and low-pass conditions. The resulting files were burned to recordable compact disk along with a calibration noise equivalent to the RMS level of the speech materials. The 1500 Hz cutoff was selected based on the Speech intelligibility index band importance function (ANSI, 1997). This function indicates that approximately 50% of speech cues are present in the frequencies below 1500Hz. It was theorized that in limiting speech information in this manner, that significantly greater signal to noise ratios would be required for participants to perform at a level of performance comparable to that achieved in the wideband condition.

HINT materials were played back from a compact disk player (Sony CDP-590) routed through a Grason Stadler audiometer (GSI-61). Speech was routed through channel 1, noise through channel 2. Presentation level was calibrated prior to each subject using the HINT calibration noise using a type I sound level meter (Larson Davis model 824). Signal to noise ratio was adjusted by varying the noise level using the attenuator dial for channel 2 of the audiometer.

*Test conditions:*

1. Non reverberant – Wideband 60 dBA (WB 60)
2. Non-reverberant – Low pass 1500 Hz 60 dBA (LP 60)
3. Non-reverberant – Wideband 45 dBA (WB 45)
4. Reverberant – Wideband 60 dBA (RV WB 60)
5. Reverberant – Low Pass 1500 Hz 60 dBA (RV LP 60)
6. Reverberant – Wideband 45 dBA (RV WB 45)

Presentation order was counterbalanced amongst participants to attempt to preclude learning and precedence effects.

*Instructions to participants*

Participants were provided with a written set of instructions. They were given the opportunity to familiarize themselves with the goal and procedure of the study, then asked to summarize the instructions. Participants were given the opportunity to ask questions. For a copy of test instructions, please see Appendix D.

## Procedure

Participants were trained in the communication confidence rating task through the administration of a practice session of the adaptive speech recognition task. All participants were able to perform the communication confidence rating task without difficulty. HINT passage order and condition order were randomized using a random number generator within Microsoft Excel (version 2002) for each participant. The adaptive HINT procedure was used to arrive at a 50% correct performance criterion under each test condition. Three additional HINT sentences were then presented at the newly realized 50% correct SNR. Participants were directed to rate response confidence for each of the three sentences by making a vertical mark on the visual analog rating scale provided. The visual analog scale consisted of a 10cm bar with anchors describing the rated response confidence. Confidence anchors were defined as “Very Low” and “Very High.” Noise level was then increased by 3dB (condition, +3dB) to decrease SNR and three responses were confidence rated. Noise level was then decreased by 3dB relative to the 50% correct level (condition, -3dB) to improve SNR, and three additional responses were confidence rated. These procedures were repeated for each of the test conditions, generating 54 data points for each participant.

## Analysis

Following completion of data collection, the response on each visual analog scale was converted to an integer value. Each millimeter along the length of the 100mm response bar was assigned a value of 1. In this fashion, a millimeter ruler was used to assign a numeric value to each response. For example, a response marked at 44mm from the leftmost anchor on the response bar was assigned a confidence rating of 44. As three sentences were rated in each test condition at each of 50% correct, 50% correct + 3dB Noise and 50% -3dB Noise, the mean of each set of three confidence ratings was calculated.

Responses were tabulated and subjected to statistical analysis via repeated measures analysis of variance.

## Results

Signal to noise ratios required to achieve a 50% correct performance level were compared between conditions in an effort to determine the relative difficulty of the various test conditions (Figure C-1).

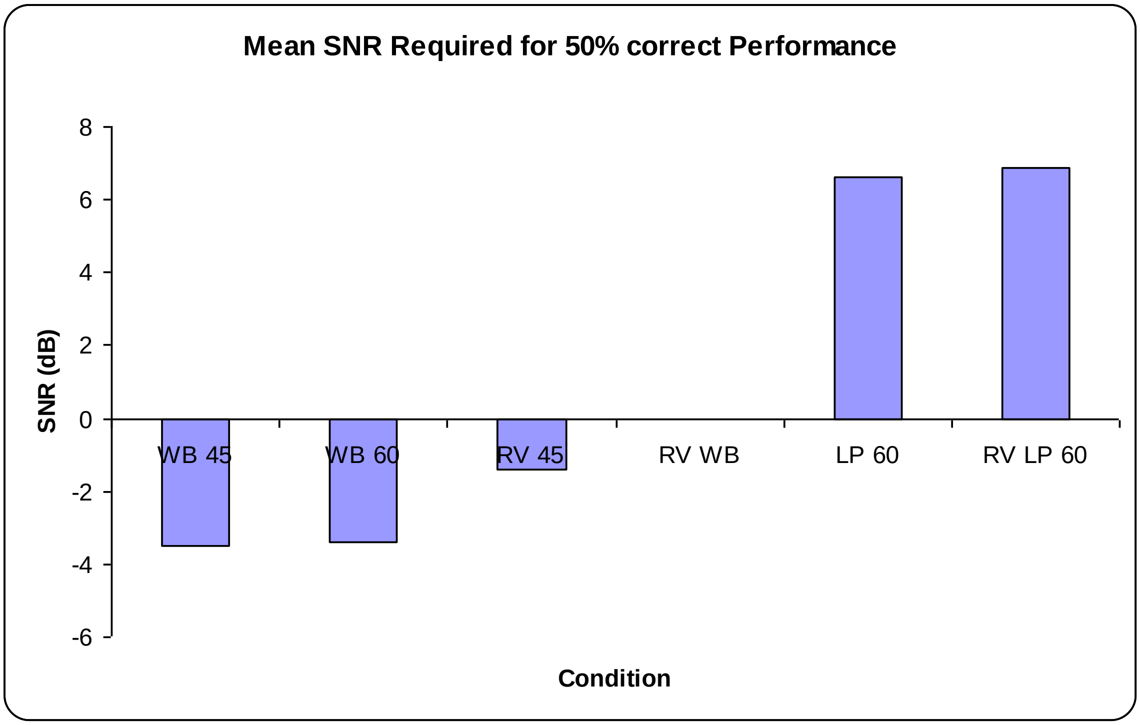

Figure C-1. Mean SNRs required to achieve 50% correct performance under six listening conditions as described in text.

A repeated measures analysis of variance suggested a significant main effect for test condition (*$F_{(5,9)}$=22.387, p<.001*). Post hoc Bonferroni comparisons suggested significant differences between SNRs required to achieve 50% correct performance between wideband and lowpass filtered stimuli (*T=7.162, p<.001*), reverberant and non-reverberant conditions, (*T=2.862, p<.05*), and between 60 and 45 dBA presentation levels (*T=8.891, p<.001*).

Next, individual confidence ratings were examined. While mean response confidence was observed to be highly variable between subjects, repeated measures analysis of variance revealed a significant main effect for confidence rating (*$F_{(1,9)}$=2.42, p<.05*). Mean confidence ratings for each test condition are listed in Table C-1.

**Table C-1:** Mean Confidence rating comparisons for reverberation condition, bandwidth and level

| | Low Reverb | High Reverb | WB | LP1500 | 60 dB | 45 dB |
|---|---|---|---|---|---|---|
| Mean | 51.25 | 42.36 | 51.02 | 40.39 | 46.20 | 47.79 |
| SD | 7.96 | 12.38 | 8.50 | 13.55 | 9.78 | 13.90 |

Confidence rating data were grouped by test condition and subjected to an additional repeated measures analysis of variance. Significant differences in confidence were observed between reverberation conditions (*T=2.862, p<.05*), and bandwidth (*T=7.162, p<.001*), but not between presentation levels. In the reverberation and bandwidth manipulation conditions, despite identical measured performance, confidence ratings were significantly lower in the more "challenging" (i.e., more reverberant, smaller bandwidth) conditions (Figure C-2).

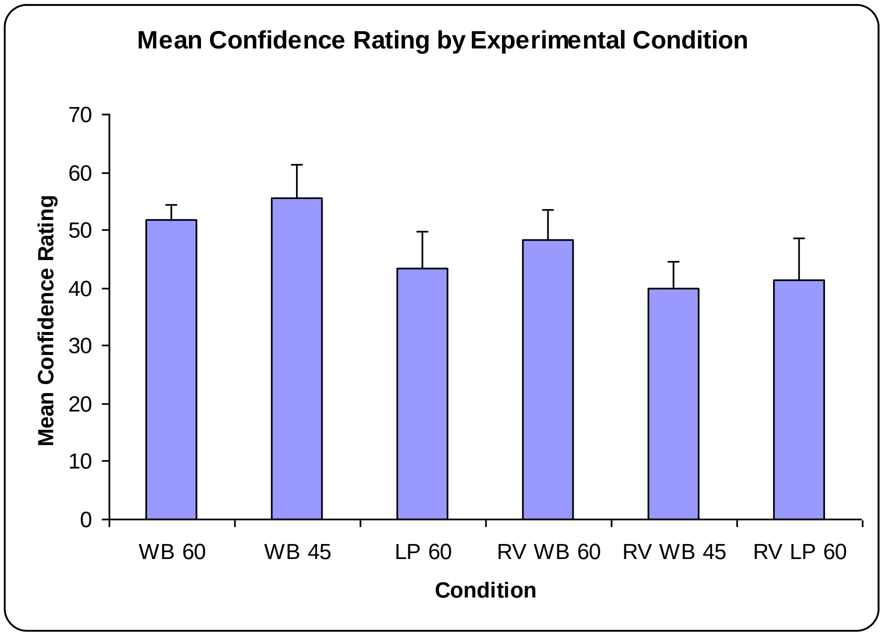

Figure C-2. Group mean confidence ratings obtained in each of six test conditions, 1.) Wideband, Non reverberant, 60dBA (WB 60), 2.) Wideband, non-reverberant, 45 dBA (WB 45), 3.) Low-pass 1500 Hz filtered, non-reverberant, 60 dBA, (LP 60), 4.) Reverberant, wideband, 60 dBA (RV WB 60), 5.) Reverberant, low-pass 1500 Hz filtered, 60 dBA (RV LP 60), and 6.) Wideband, reverberant, 45 dBA (RV 45).

Next, signal to noise ratio was investigated. Repeated measures analysis of variance revealed a significant difference in confidence rating between SNR conditions (*$F_{(2,59)}$=45.54, p<.001*). Post hoc testing suggested that confidence rating increased with SNR change from 50% +3dB noise to 50% correct (*T=6.92, p<.001*) and from 50% to 50%-3 dB noise (*T=3.41, p<.005*) (Figure C-3). These results suggest that the confidence rating measure used in this experiment is indeed sensitive to changes in performance induced through manipulation of SNR.

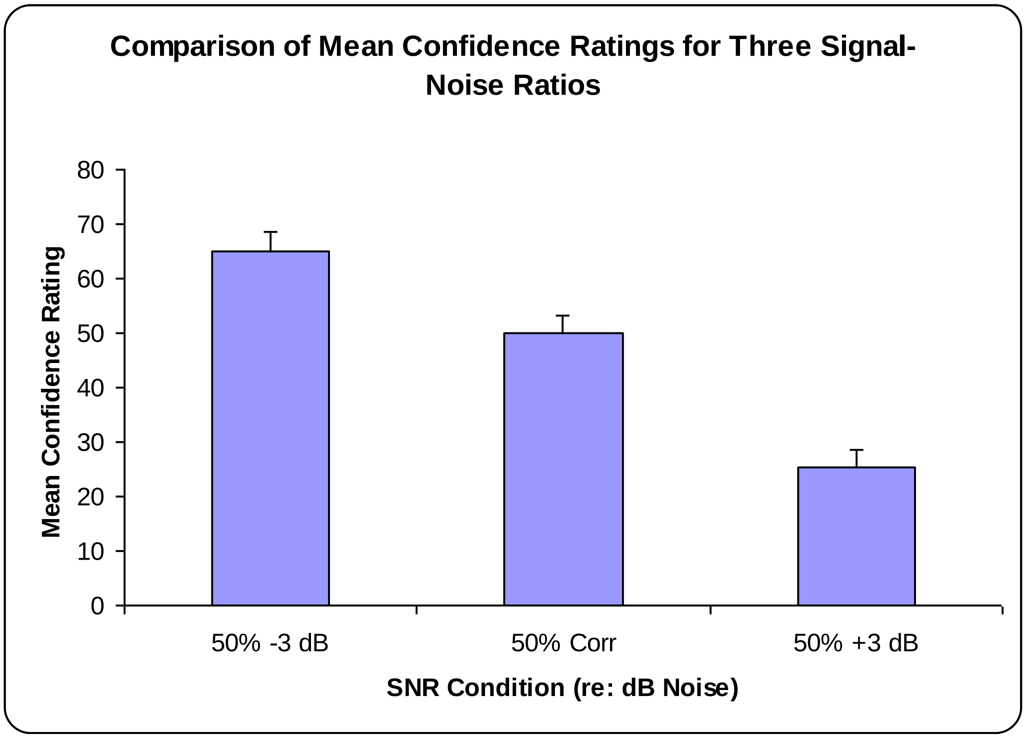

Figure C-3. Mean confidence ratings for *n=10* participants under three listening conditions. Adaptively measured 50% correct for HINT sentences, (50%), 50% correct SNR -3dB (50% +3dB Noise), and 50% correct SNR +3dB (50% -3 dB Noise)

# Discussion

The above results suggest that rated confidence for speech material understanding varied significantly as a function of test condition difficulty for reverberation time and bandwidth, but not for presentation level, despite performance being held at 50% correct. When SNR was worsened, performance was shown to decrease significantly, as was rated confidence. In a similar fashion, improving SNR resulted in higher performance and higher confidence. While both SNR and measured performance were shown to be correlated with confidence ratings, the correlations were relatively weak, likely due to the high variability in response and limited range of performance scores. The range of measured performance scores is an artifact of the SNRs used in the test and the 50% correct criterion. This experimental setup appeared to result in an “all or none” response pattern, wherein variation of the SNR from the 50% correct level resulted in close to 0% performance and low confidence when the noise level was increased or, close to 100% performance and high confidence when the noise level was decreased. The results demonstrating a difference in confidence rating due to reverberation condition and bandwidth suggest that perceived listening situation difficulty, or ease of listening, may influence communication confidence without adversely affecting measured performance. It is not clear that confidence ratings were influenced by audibility in this sample of normal hearing young adults, thus it is difficult to predict from the current study whether similar effects will be observed in a hearing impaired sample when amplification is applied to the speech and noise signals. Nor is it clear that confidence ratings will increase at the same rate between unaided and aided conditions. In an effort to begin to answer these questions, the study detailed in the main document was proposed and conducted.

D: Pilot Experiment test instructions.

# Communication Confidence Worksheet Instructions

## *Background:*

We are interested in determining whether the level of confidence perceived by a listener changes with the difficulty of the listening situation. You will be listening to sentences presented in background noise of varying loudness and attempting to repeat them. After listening to several sentences, you will be asked to rate your confidence in the accuracy of your response.

## *Goals:*

There are two goals to the current experiment.

1. Repeat back as much of each sentence as possible.
2. When directed, indicate your confidence in the accuracy of your response.

## *Directions:*

Listen to each sentence. Repeat back as much as possible. Some conditions may be much easier than others, but it is still important that you repeat everything that you heard. When instructed, indicate the level of confidence you feel in your response by making a mark on the scale below in the appropriate place.

You will be asked to repeat this process for several sentences at the end of each test condition.

# REFERENCES

Adams JK. (1957). A confidence scale defined in terms of expected percentages. The *American journal of psychology, 70*(3): 432-436.

Adams JK, Adams PA. (1961). Realism of confidence judgements. *Psychological review, 68*(1): 33-45.

ANSI (1997). ANSI S3.5-1997, "American National Standard Methods for Calculation of the Speech Intelligibility Index". American National Standards Institute, New York.

Baranski JV, Petrusic WM. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and psychophysics, 55(4),* 412-428.

Bjorkman M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational behavior and human decision processes, 58*, 386-405.

Carhart R, Tillman TW. (1970). Interaction of competing speech signals with hearing losses. *Arch of otolaryngology, 91,* 274-279.

Cox RM, Alexander GC, Gilmore C. (1987). Development of the connected speech test. (CST). *Ear and Hearing, 8,* 119S-125S.

Cox RM, Alexander GC, Rivera IM. (1991). Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners. *JSHR, 34*, 904-915.

Cox RM, McDaniel DM. (1984). Intelligibility ratings of continuous discourse: Application to hearing aid selection*. JASA, 76*(3), 758-766.

Cox RM, McDaniel DM. (1989). Development of the speech intelligibility rating (SIR) test for hearing aid comparisons. *JSHR, 32*, 347-352.

Dalton DS, Cruickshanks KJ, Klein BE, Klein R, Wiley TL, Nondahl DM. (2003). The impact of hearing loss on quality of life in older adults. *The gerontologist, 43,* 661-668.

Falconer G, Davis H. (1947). The intelligibility of connected discourse as a test for the threshold of speech. *Laryngoscope, 57,* 581-595.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191.

Gatehouse S. (1999). Glasgow hearing aid benefit profile: Derivation and validation of a client-centered outcome measure for hearing aid services. *JAAA, 10,* 80-103.

Gordon-Salant, S., Fitzgibbons, PJ. (1993). Temporal factors and speech recognition performance in young and elderly listeners. *J. Speech and hearing research, 36,* 1276-1285.

Hafter, E., & Schlauch, R. (1992). Cognitive factors and selection of auditory listening bands. In A. L. Dancer, D. Henderson, R. J. Salvi, & R. P. Hamemik (Eds.), *Noise-induced hearing loss.* St Louis, MO: Mosby-Year Book.

Hawkins JE, Stevens SS. (1950). The masking of pure tones and of speech by white noise. *JASA, 22*(3), 6-13.

Hurley RM, Sells JP. (2003). An abbreviated word recognition protocol based on item difficulty*. Ear and hearing, 24*(2):111-118.

Jones DA, Victor CR, Vetter NJ. (1984). Hearing difficulty and its psychological implications for the elderly. *J. Epidemiology and community health, 38,* 75-78.

Jowett S, Ryan T. (1985). Skin disease and handicap: an analysis of the impact of skin conditions*. Social Science in Medicine, 20*(4):425-429.

Holcomb LM, Nerbonne MA, Konkle DF. (2000). The articulation index and hearing handicap. *JAAA, 11,* 224-229.

Kalikow DN, Steven KN, Elliott LL. (1977). Development of a test of speech intelligibility in noise testing sentence materials with controlled word predictability. *J Acoust Soc Am, 61*, 1337–1351.

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ration loss in normal-hearing and hearing-impaired listeners. *The journal of the acoustical society of America, 116,* 2395-2405.

Knutson JF, Lansing CR. (1990). The relationship between communication problems and psychological difficulties in persons with profound acquired hearing loss. J Speech hearing disorders, 55(4):656-64.

Kodman F Jr. (1961). Successful binaural hearing aid users. *Archives of otolaryngology, 74,* 302-304.

Koriat A, Lichtenstein S, Fischhoff B. (1980). Reasons for confidence. Journal of experimental psychology*: Human learning and memory, 6*(2): 107-118.

Kroner S, Biermann A. (2007). The relationship between confidence and self-concept – towards a model of response confidence. *Intelligence, 35*, 580-590.

Lichtenstein S, Fischhoff D. (1977). Do those who know more also know more about how much they know? *Organizational behavior and human performance, 20*, 159-183.

McCoy SL, Tun PA, Cox LC, Colangelo M, Stewart RA, Wingfield A.(2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The quarterly journal of experimental psychology, 58A*(1);22-33.

Mulrow CD, Aguilar C, Endicott JE, Velez R, Tuley MR, Charlip WS, Hill JA. (1990). Association between hearing impairment and quality of life of elderly individuals. J Am geriatriric society, 38(1):45-50.

Mulrow CD, Aguilar C, Endicott JE, Tuley MR, Velez R, Charlip WS, Rhodes MC, Hill JA, DeNino LA. (1990). Quality-of-life changes and hearing impairment. A randomized trial. Annals of Internal Medicine, 113(3):188-94.

Murphy AH. (1973). A new vector partition of the probability score. *Journal of applied meteorology, 12*, 595-600.

Nilsson M, Soli S, Sullivan J. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The journal of the acoustical society of America, 95,* 1085-1099.

Noble W. (2006). Benefits of fitting one versus two hearing aids. *The ASHA Leader, Mar;11*(4),6-7,32.

Neufeldt, V. Ed. (1988). Webster's New World Dictionary (11th ed.). New York: Simon & Schuster.

Paap KR, Chun E, Vonnahme P. (1999). Discrete threshold versus continuous strenth models of perceptual recognition. *Canadian journal of experimental psychology, 53*(4): 277-293.

Parry W, Steen N, Galloway SR,Kenny RA, Bond J. (2001). Falls and confidence related quality of life outcome measures in an older British cohort. *Postgraduate Medical Journal 77.904* (Feb 2001): 103.

Perfect TJ., Hollins TS. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied cognitive psychology, 10*, 371-382.

Pickett JM, Pollack I. (1963). Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt. *Language and Speech, 6,* 151-164.

Preminger JE, Neuman AC, Bakke MH, Walters D, Levitt H. (2000). An examination of the practicality of the simplex procedure. *Ear and hearing, 21*(3), 177-193.

Preminger JE, Van Tasell DJ. (1995). Quanitifying the relationhip between speech quality and speech intelligibility. *Journal of Speech and Hearing Research, 38,* 714-725.

Rankovic CM, Levy RM. (1997). Estimating articulation scores*. JASA, 102*(6), 3754-3761.

Saunders G, Cienkowski. (2002). A test to measure subjective and objective speech intelligibility*. JAAA, 13,* 38-49.

Saunders, G., & Levitt, H. (1991). An automated true-false response time test for evaluating noise reduction systems (abstract). *ASHA, 33,* 164.

Speaks C, Parker B, Harris C, Kuhl P. (1972). Intelligibility of connected discourse. *Journal of Speech and Hearing Research, 15,* 590-602.

Stankov L, Lee J. (2008). Confidence and cognitive test performance. *Journal of educational psychology, 100*(4): 951-976.

Tambs, K. (2004). Moderate effects of hearing loss on mental health and subjective well-being: results for the Nord-Trondelag hearing loss study. Psychosom Med, 66(5):776-782.

Tenney ER, Spellman BA, MacCoun RJ. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility*. Journal of experimental psychology, 44*, 1368-1375.

Wilson RH, Abrams HB, Pillion AL. (2003). A word-recognition task in multitalker babble using a descending presentation mode from 24-dB S/B to 0-dB S/B. *J Rehabil Res Dev 40*:321–328.

Whitney SL, Hudak MT, Marchetti GF. (1999). The activities specific balance confidence scale and the dizziness handicap inventory. A comparison. *Journal of vestibular research, 9*(4):253-259.

Whitney SL, Wrisley DM, Brown KE, Furman JM. (224). Is perception of handicap related to functional performance in persons with vestibular dysfunction? *Otology & Neurotology, 25,* 139-243.

World Health Organization. (1980). International classification of impairments, disabilities and handicaps (ICIDH). Geneva,WHO.

Yardley L, Smith H. (2002). A prospective study of the relationship between feared consequences of falling and avoidance of activity in community-living older people. *The gerontologist, 42,* 17-23.