J Am Acad Audiol 29:206–222 (2018)

An Argument for Self-Report as a Reference Standard in Audiology

DOI: 10.3766/jaaa.16128

Andrew J. Vermiglio*
Sigfrid D. Soli†
Xiangming Fang‡

Abstract

Background: The primary components of a diagnostic accuracy study are an index test, the target condition (or disorder), and a reference standard. According to the Standards for Reporting Diagnostic Accuracy statement, the reference standard should be the best method available to independently determine if the results of an index test are correct. Pure-tone thresholds have been used as the ''gold standard'' for the validation of some tests used in audiology. Many studies, however, have shown a lack of agreement between the audiogram and the patient's perception of hearing ability. For example, patients with normal audiograms may report difficulty understanding speech in the presence of background noise.

Purpose: The primary purpose of this article is to present an argument for the use of self-report as a reference standard for diagnostic studies in the field of audiology. This will be in the form of a literature review on pure-tone threshold measures and self-report as reference standards. The secondary purpose is to determine the diagnostic accuracy of pure-tone threshold and Hearing-in-Noise Test (HINT) measures for the detection of a speech-recognition-in-noise disorder.

Research Design: Two groups of participants with normal pure-tone thresholds were evaluated. The King–Kopetzky syndrome (KKS) group was made up of participants with the self-report of speech-recognition-in-noise difficulties. The control group was made up of participants with no reports of speech-recognition-in-noise problems. The reference standard was self-report. Diagnostic accuracy of HINT and pure-tone threshold measures was determined by measuring group differences, sensitivity and specificity, and the area under the curve (AUC) for receiver-operating characteristic (ROC) curves.

Study Sample: Forty-seven participants were tested. All participants were native speakers of American English. Twenty-two participants were in the control group and 25 in the KKS group. The groups were matched for age.

Data Collection and Analysis: Pure-tone threshold data were collected using the Hughson–Westlake procedure. Speech-recognition-in-noise data was collected using a software system and the standard HINT protocol. Statistical analyses were conducted using descriptive, correlational, two-sample *t* tests, and logistic regression.

Results: The literature review revealed that self-report has been used as a reference standard in investigations of patients with normal audiograms and the perception of difficulty understanding speech in the presence of background noise. Self-report may be a better indicator of hearing ability than pure-tone thresholds in some situations. The diagnostic accuracy investigation revealed statistically significant differences between control and KKS groups for HINT performance ($p < 0.01$), but not for pure-tone threshold measures. Better sensitivity was found for the HINT Composite score (88%) than pure-tone average (PTA; 28%). The specificities for the HINT Composite score and PTA were 77% and 95%, respectively. ROC curves revealed a greater AUC for the HINT Composite score (AUC = 0.87) than for PTA (AUC = 0.51).

Conclusion: Self-report is a reasonable reference standard for studies on the diagnostic accuracy of speech-recognition-in-noise tests. For individuals with normal pure-tone thresholds, the HINT demonstrated

*Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC; †House Clinic, Los Angeles, CA; ‡Department of Biostatistics, East Carolina University, Greenville, NC

Corresponding author: Andrew J. Vermiglio, Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC 27834; E-mail: vermiglioa@ecu.edu or vermiglio.av@gmail.com

a higher degree of diagnostic accuracy than pure-tone thresholds for the detection of speech-recognition-in-noise disorder.

Key Words: audiogram, auditory, central auditory processing disorder, clinical entity, diagnostic accuracy, gold standard, pure-tone thresholds reference standard, self-report, sensitivity, specificity, speech recognition in noise, STARD statement, Sydenham–Guttentag criteria, target condition, target disorder

Abbreviations: AMA = American Medical Association; APD = auditory processing disorder; AUC = area under the curve; BKB-SIN = Bamford–Kowal–Bench Speech-in-Noise Test; cABR = complex auditory brainstem response; CPHI = Communication Profile for the Hearing Impaired; HHIE-S = Hearing Handicap Inventory for the Elderly; HINT = Hearing-in-Noise Test; KKS = King–Kopetzky Syndrome; MRI = magnetic resonance imaging; PTA = pure-tone average; PTT = pure-tone threshold; QuickSIN = Quick Speech-in-Noise Test; ROC = receiver-operating characteristic; SNR = signal-to-noise ratio; WIN = Words in Noise test

INTRODUCTION

Pure-tone threshold (PTT) testing has been the primary method for the measurement of hearing since the development of the first audiometer in the 1920s (Jerger, 2009). It has been used to detect the presence of a hearing loss, to determine the need for hearing aids, and for the detection of the harmful effects of noise exposure, or ototoxic medication. Results from PTT testing are related to the ability to recognize speech in a quiet environment. In general, however, these results are poorly related to the presence of a speech-recognition-in-noise disorder. This is true for both normal and elevated PTTs. The ability to recognize speech in a noisy environment must be measured directly and not inferred from the audiogram (Vermiglio et al, 2012).

A speech-recognition-in-noise disorder may be found in the presence of a normal audiogram (King, 1954b; Middelweerd et al, 1990). This condition has been called an auditory processing disorder (APD) by Pryce et al (2010). Vermiglio (2014), using the Sydenham–Guttentag criteria for legitimate disorders, has argued that a speech-recognition-in-noise disorder in the presence of a normal audiogram should not be equated with an APD. The former is a clearly defined disorder, whereas the latter is too vague a concept to provide guidance for clinicians. A speech-recognition-in-noise disorder is measurable. The data from speech-recognition-in-noise testing may be used to determine the need for and the benefit of various forms of intervention including auditory training (Sweetow and Sabes, 2006), a hearing aid with a directional microphone (Kuk et al, 2008; Johnston et al, 2009), or a frequency modulation system (Johnston et al, 2009). While a number of speech-in-noise test protocols are commercially available, the validity or diagnostic accuracy of these tests has not been clearly presented.

The term diagnostic accuracy refers to the ability of an index test to detect the presence or absence of a target condition or disorder (Bossuyt et al, 2003). A reference standard is a test used to determine if the results of the index test are correct. Results from the reference standard are used to determine the group assignment for research participants. The disordered group includes participants with the target condition. Participants without the target condition are assigned to the control group. The reference standard should be the best available method for establishing the presence or absence of the target condition. Diagnostic accuracy of an index test may be established in at least three ways: (a) by determining if the index test results in healthy participants vary from results in patients with the target condition (Bossuyt et al, 2003), (b) by determining the sensitivity and specificity of the index test for the detection of the target condition (Berkson, 1947), and (c) by determining receiver-operating characteristic (ROC) curves for the index test (Peterson et al, 1954). For the present literature review, the main components of a diagnostic accuracy study (the index test, target condition, and reference standard) will be identified even though these terms may not have been used by the authors. For a review of the basics of diagnostic accuracy studies, see the papers by Swets (1988), Swets et al (2000), Bossuyt et al (2003), and Vermiglio (2016).

The Problem

Wilson et al (2007) used PTTs as a reference standard to determine the diagnostic accuracy of four different speech-recognition-in-noise index tests: the Bamford–Kowal–Bench Speech-in-Noise Test (BKB-SIN; Bench et al, 1979; Niquette et al, 2003; Etymotic, 2005), the Quick Speech-in-Noise Test (QuickSIN; Killion et al, 2004), the Words-in-Noise test (WIN; Wilson, 2003; Wilson and Burks, 2005), and the Hearing-in-Noise Test (HINT; Nilsson et al, 1994; Vermiglio, 2008). The target condition was a hearing loss defined as a pure-tone average (PTA) for 500, 1000, and 2000 Hz between 20 and 60 dB HL. The components of the diagnostic accuracy study by Wilson et al (2007) are presented in Table 1. Participants with a hearing loss (PTA >20 and <60 dB HL) were assigned

Table 1. Diagnostic Accuracy Results for Four Speech Recognition in Noise Tests

Index Test	Target Condition	Reference Standard	True Positives (Sensitivity, %)	False Negatives (%)
WIN Test	PTA >20 and <60 dB HL	PTTs	99.0	1.0
QuickSIN	PTA >20 and <60 dB HL	PTTs	90.0	10.0
HINT	PTA >20 and <60 dB HL	PTTs	72.0	28.0
BKB-SIN	PTA >20 and <60 dB HL	PTTs	78.0	22.0

Notes: The target condition was a PTA >20 and <60 dB HL. PTTs were used as the reference standard. Data from Wilson et al, 2007.

to the disordered group. Sensitivity represents the percentage of true-positive results. In the Wilson et al study, it represents the percentage of participants in the disordered group who performed below normal limits on a speech-recognition-in-noise test. The authors did not report the specificity (percentage of true negatives) or the percentage of false positives for their study. The authors found that of the four speech-recognition-in-noise tests, the WIN and QuickSIN protocols showed the highest sensitivity. The WIN test had a sensitivity of 99% and the QuickSIN test had a sensitivity of 90%. The BKB-SIN test and the HINT had sensitivities of 78% and 72%, respectively.

The authors concluded that "The QuickSIN and WIN materials are more sensitive measures of speech recognition performance in background noise than the BKB-SIN and HINT materials." However, this statement is misleading. The authors did not determine the sensitivity of the index tests for the detection of a speech-recognition-in-noise disorder. Instead, they determined the sensitivity of the index tests for a different target condition, a PTA between 20 and 60 dB HL. This is an example of target displacement and it occurs when the diagnostic accuracy of an index test for one target condition is attributed to a different target condition (Vermiglio, 2016). It is more correct to say that the QuickSIN and WIN protocols were more sensitive to a PTA between 20 and 60 dB HL than the BKB-SIN and HINT protocols.

PTTs as the "Gold Standard" for the Validity of Self-Report Measures

In the field of audiology, PTT testing has been called the "gold standard" for the assessment of the ability to hear (Sindhusake et al, 2001; Shargorodsky et al, 2010; Kiely et al, 2012; Baiduc et al, 2013; Zecker et al, 2013). A gold standard or reference standard should be clearly described to allow for replication. It should appropriately address the research or clinical question, and it should be applied to participants in both the disordered and control groups (Bossuyt et al, 2003).

Diagnostic accuracy studies have been conducted where self-report is used as the index test and PTTs are used as the reference standard (Table 2). In these studies, a true-positive result is one where a study participant reports that they believe they have a hearing loss and the PTA for the better ear or worse ear is >20 or 25 dB HL, depending on the criterion used in the study. A false-positive result occurs when the participant reports hearing difficulty but the PTA is ≤20 or 25 dB HL. For a false-positive result, even though a participant reports hearing difficulty, this claim is considered false because the PTA is within normal limits. This may be classified as a "test-centered" approach to diagnostic accuracy. The assumption, in this example, is that the most accurate method for determining the presence or absence of a hearing disorder is pure-tone audiometry.

Nondahl et al (1998) reported that the single question, "Do you feel you have a hearing loss?" had greater sensitivity to the presence of a hearing loss ($PTA_{0.5,1.0,2.0,4.0\ kHz}$ >20 dB HL) than the Hearing Handicap Inventory for the Elderly (HHIE-S). The single question had a sensitivity of 71% and the HHIE-S had a sensitivity of 34%. The authors reported that for prevalence studies the single question was preferable to the HHIE-S. The same single question was found to be more sensitive to the presence of mild and moderate hearing losses ($PTA_{0.5,1.0,2.0,4.0\ kHz}$ >20 dB HL and <40 dB HL) than the HHIE-S according to Sindhusake et al (2001). For a hearing loss with a PTA >60 dB HL, Sindhusake and colleagues found that both the single question and the HHIE-S had a sensitivity of 100%. Hannula et al (2011) demonstrated that a single question, "Do you have any difficulty with your hearing?" had less sensitivity to a $PTA_{0.5,1.0,2.0,4.0\ kHz}$ than to a $PTA_{0.5,1.0,2.0\ kHz}$.

Classical Threshold of Hearing Studies and Self-Report

Some of the classical studies for the determination of the threshold of hearing have used self-report as the inclusion criterion for the participants. Bunch (1929) included participants with hearing loss if they reported, "While I may not hear quite so well as when I was young, my hearing is as good as that of any one of my age." Steinberg et al (1940) reported on hearing tests conducted at the 1939 World's Fairs in San Francisco and New York. The authors noted that "a person is scarcely aware of a hearing loss of less than 25 db [sic]." It appears that the self-report of the participants determined the 25 dB HL cut point. Steinberg et al (1940) may be the source of the conventional use of 25 dB HL as the cut point for normal PTTs in the current practice of audiology.

Table 2. Diagnostic Accuracy Values for Studies Where PTA Was Used as the Reference Standard for Self-Report of Hearing Difficulties (Index Test)

Study	Index Test(s)	Target Condition	Reference Standard	False Positives (%)	False Negatives (%)	True Positives (Sensitivity, %)	True Negatives (Specificity, %)
Nondahl et al (1998)	Self-report: "Do you feel you have a hearing loss?" A "yes" answer is a positive result	Hearing loss for worse ear (PTA of 500, 1000, 2000, and 4000 Hz >25 dB HL)	PTTs	29.0	29.0	71.0	71.0
Nondahl et al (1998)	Self-report: Hearing Handicap Inventory for the Elderly. A score >8 is a positive result	Hearing loss for worse ear (PTA of 500, 1000, 2000, and 4000 Hz >25 dB HL)	PTTs	5.0	66.0	34.0	95.0
Sindhusake et al (2001)	Self-report: "Do you feel you have a hearing loss?" A "yes" answer is a positive result	Mild hearing loss for better ear (PTA for 0.5, 1.0, 2.0, and 4.0 kHz >25 dB HL)	PTTs	33.0	22.0	78.0	67.0
Sindhusake et al (2001)	Self-report: Hearing Handicap Inventory for the Elderly. A score >8 is a positive result	Mild hearing loss better ear (PTA for 0.5, 1.0, 2.0, and 4.0 kHz >25 dB HL)	PTTs	15.0	42.0	58.0	85.0
Sindhusake et al (2001)	Self-report: "Do you feel you have a hearing loss?" A "yes" answer is a positive result	Moderate hearing loss better ear (PTA for 0.5, 1.0, 2.0 and 4.0 kHz >40 dB HL)	PTTs	44.0	7.0	93.0	56.0
Sindhusake et al (2001)	Self-report: Hearing Handicap Inventory for the Elderly. A score >8 is a positive result	Moderate hearing loss better ear (PTA for 0.5, 1.0, 2.0, and 4.0 kHz >40 dB HL)	PTTs	24.0	20.0	80.0	76.0
Sindhusake et al (2001)	Self-report: "Do you feel you have a hearing loss?" A "yes" answer is a positive result	"Marked" hearing loss better ear (PTA for 0.5, 1.0, 2.0, and 4.0 kHz >60 dB HL)	PTTs	50.0	0.0	100.0	50.0
Sindhusake et al (2001)	Self-report: Hearing Handicap Inventory for the Elderly. A score >8 is a positive result	"Marked" hearing loss better ear (PTA for 0.5, 1.0, 2.0 and 4.0 kHz >60 dB HL)	PTTs	30.0	0.0	100.0	70.0
Hannula et al (2011)	Self-report: "Do you have any difficulty with your hearing?" A "yes" answer is a positive result	Hearing loss for better ear (PTA for 0.5, 1.0, and 2.0 kHz ≥20 dB HL)	PTTs	31.0	23.0	77.0	69.0
Hannula et al (2011)	Self-report: "Do you have any difficulty with your hearing?" A "yes" answer is a positive result	Hearing loss for better ear (PTA for 0.5, 1.0, 2.0, and 4.0 kHz ≥20 dB HL)	PTTs	26.0	31.0	69.0	74.0

A national health survey was conducted by the United States Public Health Service from 1935 to 1936 (NIH, 1938). The survey (also known as the Beasley survey) included a series of questions used to determine the prevalence of hearing loss in the population. Each participant was asked if they had ever experienced hearing difficulty. If the answer was "no," the participant was classified as having "normal hearing." Those who reported no hearing difficulty were in the control group; those who reported various levels of hearing difficulties were assigned to the disordered groups (Beasley, 1940). This is another example where self-report was used as a reference standard for a diagnostic accuracy study of pure-tone audiometry. The results of the Beasley survey were accepted as the American standard for the threshold of hearing by the American Medical Association (AMA) and the American Standards Association in 1951 (Glorig, 1956; Jerger, 2009). In contrast, the self-report of hearing ability has not been used to determine the presence or absence of a hearing disorder in a number of other hearing threshold surveys (Lane, 1922; Zuehl, 1922; Sivian and White, 1933; Dadson and King, 1952; Wheeler and Dickson, 1952; Hinchcliffe, 1961).

Self-Report as the Reference Standard for the Validity of PTT Measures

Merluzzi and Hinchcliffe (1973) used self-report as the reference standard to determine the pure-tone levels for the "threshold of subjective auditory handicap." Four hundred participants were tested. Each participant answered the question "Is your hearing normal, or not as good as it used to be?" Those who answered, "not as good as it used to be" were in the disordered group and those who answered that their hearing was normal were in the control group. According to the authors, for each frequency, the hearing level corresponding to the intersection of the distributions of the thresholds for the two groups was determined to be the threshold of subjective auditory handicap.

Ward (1983) suggested self-report as a possible reference standard for the validation of PTTs when estimating "auditory handicap." Dobie (2011) followed this suggestion and used self-report as the reference standard to determine the validity of the AMA method for the estimation of hearing disability. The AMA method for adults is derived from the $PTA_{0.5,\ 1.0,\ 2.0,\ 3.0\ kHz}$ for each ear (AAO, 1979). The Communication Profile for the Hearing Impaired (CPHI) was used to measure a participant's self-report of communication difficulties. Six of the 18 items referred to conditions with background noise. The other items referred to quiet environments or other settings such as talking to someone on the telephone or someone in another room. Data from 1,001 patients were analyzed. Diagnostic accuracy of PTA was determined using a correlational analysis. There was a modest correlation found between the PTA for the better ear versus scores from the CPHI ($r = -0.385$). This accounts for only about 15% of the variance in the CPHI scores. The p-value for this relationship was not reported.

Self-report has also been used as a reference standard in the literature on pain as a target condition. Stilma et al (2015) used the self-report of pain as the reference standard for the critical-care pain observation tool (index test). Manne et al (1992) wrote that, "Since pain is a subjective perception, self-report should be relied upon as the 'gold standard' for assessing pain." It could also be argued that since hearing is a subjective perception, self-report should be used as a reference standard for behavioral tests in the field of audiology.

PTTs versus Self-Report

Martin and Champlin (2000) addressed the issue of a PTA cut point (or "high-level") for the presence of a hearing loss. They wanted to determine if the high level for normal hearing should be a PTA of 25 or 15 dB HL. The authors suggested, "instead of posing a philosophical position about what hearing level constitutes the upper limit of normal hearing sensitivity, it is only patients experiencing hearing difficulty who could advise us." In other words, the self-report from patients may be a better indicator of hearing difficulties than PTA. This is in contrast to the studies presented in Table 2, where PTA was used to determine if self-report of hearing ability was accurate.

Martin and Champlin decided to tap into the self-perception of hearing disability by evaluating patients who had purchased hearing aids. They assumed that a patient would not buy a costly hearing aid unless they knew that they were having hearing difficulties. The authors obtained PTA data from Starkey Laboratories for 556,026 ears of clients who purchased hearing aids. The authors wrote that "The fact that 29,333 (5.3%) of over half a million hearing-aid purchasers whose PTAs were <25 dB HL sought assistance in dealing with their hearing impairments is distinct evidence that many people, who may be told that their hearing is normal based on their PTA, would clearly testify that this is not the case." The results indicate that a PTA <25 dB HL is not a sufficient condition for the self-perception of normal hearing ability. Based on these data, the authors proposed that the upper limit for normal hearing should be 15 dB HL. They also stated that while they do not recommend amplification for every patient with a PTA >15 dB HL, the patient's complaints of hearing difficulty should be "recognized and explored." This is consistent with data from Liberman et al (2016). They demonstrated that individuals with normal PTTs may have neural degeneration in the cochlea that may affect the perception of speech in noisy environments.

In an effort to evaluate readiness for amplification, Palmer et al (2009) asked their patients, "On a scale from 1 to 10, 1 being the worst and 10 being the best, how would you rate your overall hearing ability?" The self-report ratings and PTAs were compared to the pursuance of amplification by the patient. Patients with a $PTA_{1.0,\ 2.0,\ 3.0,\ 4.0\ kHz} < 10$ dB HL rated their overall hearing ability from 4 to 10. None of the patients who rated their hearing ability as a 9 or 10 pursued amplification. One hundred percent of patients who rated their hearing ability as a 1 or 2 pursued amplification. Patients with a rating of 3 to 5 had an ~80% rate of pursuing amplification. The authors concluded that the results support the predictive value of the self-report question for the pursuit of amplification. They also stated that their data demonstrated lack of relationship between PTT data and perception of hearing difficulties.

Is the Self-Report of Speech-Recognition-in-Noise Ability Reliable?

Saunders et al (2004) evaluated "subjective" and "objective" speech-recognition-in-noise abilities. They measured HINT Noise Right and Noise Left thresholds and then averaged them together. The HINT thresholds were measured by using two methods. First, the standard HINT protocol (Nilsson et al, 1994) was used to determine the "performance" or "objective" threshold. In the second method, instead of repeating the target sentences, the participants were instructed to tell the tester if they thought that they could hear all of the words for each sentence presentation. This was used to determine a "perceptual" or "subjective" threshold. One hundred and seven participants were tested. The participants were between 24 and 83 yr of age (mean = 58.9 yr). Test–retest reliability was determined for repeated measures that were two weeks apart. For the test–retest measures, the r values were >0.900 for both the performance and perceptual thresholds. Furthermore, a strong relationship was found between performance and perceptual thresholds ($r = 0.95$, $p < 0.005$). The authors concluded, "individuals are remarkably accurate at estimating their own hearing ability." Self-report, therefore, appears to be a reasonable reference standard for the identification of the presence or absence of a speech-recognition-in-noise disorder.

Though Saunders et al (2004) used the terms "objective" and "subjective" for their speech-recognition-in-noise measures, it should be noted that all behavioral assessments in audiology include both objective and subjective components (Table 3). The objective components are those that may be measured without input from the patient. According to the Merriam-Webster dictionary (2016), the word "subjective" relates "to the way a person experiences things in his or her own mind." Certainly, questionnaires and other self-report measures are examples of subjective assessments. It could be argued, however, that since all behavioral testing methods in audiology rely on some form of self-report, they all possess a subjective component. Noble (1988) stated, "There is no more reason a priori to expect people to fake in response to questions about their hearing than in response to tests using tones."

Self-Report as the Reference Standard for Speech-Recognition-in-Noise Index Tests

A number of investigations have used self-report as a reference standard for the identification of the presence or absence of a speech-recognition-in-noise disorder (Table 4). For the most part, the assignment of research participants to the disordered groups was based on the self-perception of hearing-in-noise difficulties. Middelweerd et al (1990) identified the presence of the disorder by the participants' self-report of "diminished speech intelligibility, especially in background noise." The control group apparently did not report any difficulties with the ability to recognize speech in noise. Saunders and Haggard (1989) coined the term Obscure Auditory Dysfunction and defined it as the report of difficulty understanding speech in background noise in the presence of normal PTTs. Lutman and Saunders (1992) determined the diagnostic accuracy of transient-evoked otoacoustic emissions for the detection

Table 3. Subjective and Objective Components for Pure Tone and Speech-Recognition-in-Noise Tests and Self-Report

Measure	Stimuli	Levels of the Stimuli	Subjective Component	Objective Component
PTTs	Pure tones from 250 to 8000 Hz	Calibrated	Self-report of stimulus audibility	Level and frequency of stimulus at threshold
Speech-recognition-in-noise test (performance thresholds)	Speech and noise	Calibrated	Self-report of speech perception (repetition of target speech)	Threshold SNR
Speech-recognition-in-noise test (perceptual thresholds)	Speech and noise	Calibrated	Self-report of speech perception (yes or no)	Threshold SNR
Self-report of speech-recognition-in-noise ability	Speech and noise	Uncalibrated	Self-report of speech perception	Levels of speech and noise are unknown

Note: Test measures used in Saunders et al (2004).

Table 4. Diagnostic Accuracy Studies Where Self-Report Was Used as the Reference Standard for the Detection of a Speech-Recognition-in-Noise Disorder

Study	Index Test(s)	Target Condition	Reference Standard
Middelweerd et al (1990)	Speech-recognition-in-noise test developed by Plomp and Mimpen (1979)	A speech-recognition-in-noise disorder in the presence of a normal audiogram	Self-report
Saunders and Haggard (1989)	Pseudo-free-field in noise test	Obscure auditory dysfunction (speech-recognition-in-noise disorder in the presence of normal PTTs)	Self-report
Lutman and Saunders (1992)	Transient-evoked otoacoustic emissions	Obscure auditory dysfunction (speech-recognition-in-noise disorder in the presence of normal PTTs)	Self-report
Rappaport et al (1993)	Northwestern University-6 word list in steady-state or chopped noise	Ideopathic discriminatory dysfunction (speech-recognition-in-noise disorder in the presence of normal PTTs)	Self-report
Zhao and Stephens (1999)	Audioscan	KKS (speech-recognition-in-noise disorder in the presence of normal PTTs)	Self-report
Zhao and Stephens (2006)	Otoacoustic emissions	KKS (speech-recognition-in-noise disorder in the presence of normal PTTs)	Self-report
Tremblay et al (2015)	Speech in single talker babble test	Hearing difficulty (includes speech-recognition-in-noise problems) in the presence of normal PTTs	Self-report
Tremblay et al (2015)	Otoacoustic emissions	Hearing difficulty (includes speech-recognition-in-noise problems) in the presence of normal PTTs	Self-report

Note: Group differences were found for each study.

of Obscure Auditory Dysfunction. The disordered group was composed of patients, who sought a referral to a medical specialist for their hearing difficulties. There was no indication in this study, however, that the control group was questioned about their speech-recognition-in-noise difficulties.

Rappaport et al (1993) used the term Idiopathic Discriminatory Dysfunction to describe the self-report of impaired speech intelligibility in noisy environments for participants with normal PTTs. Zhao and Stephens (1999) used self-report as the reference standard to determine the diagnostic accuracy of the Audioscan (Meyer-Bisch, 1996) as an index test for the detection of King–Kopetzky syndrome (KKS) (target condition). The term King–Kopetzky syndrome was coined by Hinchcliffe (1992). It was named in part for Dr. P. F. King who presented case studies of Royal Air Force members who had psychogenic deafness (King, 1954a). According to Zhao and Stephens (1999), King–Kopetzky syndrome is a term used to describe the condition where an "individual complains of difficulties understanding speech in the presence of background noise but has normal hearing thresholds." Tremblay et al (2015) identified the presence of hearing difficulties by participant responses to four questions regarding hearing ability. Three of the four questions concerned environments where each participant listens to speech in the presence of background noise.

For all of the studies presented in Table 4, self-report was used as the reference standard. This may be classified as a "patient-centered" approach to diagnostic accuracy studies. It is consistent with physician Sir William Osler who said, "Listen to your patient, he is telling you the diagnosis." Alvord (1983) wrote, "Frequently, patients are seen at veterans' and military hospitals who are known to have undergone significant noise exposure and yet have normal hearing for pure tones. Such patients have been known to complain of decreased ability to hear speech in noise and seem surprised to learn that their hearing is normal." Even though their PTTs were normal, Alvord took a cue from his patients and investigated the nature of their complaint. Results of his study indicated the presence of cochlear damage for participants with normal audiograms, a history of noise exposure, and complaints of a decreased ability to hear speech in noise. The self-report of Alvord's patients alerted him to the presence of a hearing disorder.

Objections to the Use of Self-Report as a Reference Standard

The present literature review has demonstrated a precedent and rationale for the utilization of self-report as a reference standard in the field of audiology. Even so, one may object to this arrangement when considering

anecdotal evidence that a patient's description of their hearing ability may be questionable. This is not an uncommon occurrence, especially when the test results are used as part of a hearing screening for employment or for an evaluation for compensation regarding a job-related hearing disorder. For diagnostic accuracy studies, however, it is assumed that participants in the disordered and control groups are not motivated to exaggerate or understate their hearing ability. It would be inappropriate to use self-report as a reference standard for a diagnostic accuracy study when the participants are prone to bias.

Some may object to the use of self-report as a reference standard because it is a "subjective" as opposed to an "objective" measure of hearing ability. Mendel (2007) identified speech-recognition-in-noise protocols as "objective" measures. However, as shown in Table 3, the major differences between "subjective" and "objective" measures are the calibration of the stimuli and the control of the test environment. All behavioral tests used in audiology rely on the self-report of the participants. There are no "pure objective" behavioral measures in audiology.

Bossuyt et al (2003) wrote that authors of diagnostic accuracy studies should describe the index test and reference standard in sufficient detail to allow readers to assess the potential for bias and to evaluate the generalizability of the results. The reader also needs to know the authors' approach to the study. The studies presented in Table 2 may be considered "test-centered" diagnostic accuracy studies, where a behavioral test (PTTs) was used as the reference standard. The studies presented in Table 4 may be considered "patient-centered" diagnostic accuracy studies, where the self-report of the participant is used as the reference standard. The selection of the reference standard should be based on the target condition of interest. The utility of the results of both types of studies is left to the judgement of the reader.

Summary of the Literature Review

Wilson et al (2007) used PTTs as the reference standard for the validation of index tests used for the assessment of the ability to recognize speech in noise. This research design is questionable since PTTs appear not to be the best available method for establishing the presence or absence of a speech-recognition-in-noise disorder (Fry, 1942; Alvord, 1983; Saunders and Haggard, 1989; Middelweerd et al, 1990; Lutman and Saunders, 1992; Rappaport et al, 1993; Martin and Champlin, 2000; Zhao and Stephens, 2006). The validity of self-report has been determined using PTT measures as the reference standard in a number of "test-centered" investigations (Table 2). In these studies, a self-report of hearing difficulty is considered invalid (or a false-positive result) when made in the presence of a normal PTA. In contrast, Merluzzi and Hinchcliffe (1973) and Dobie (2011) used self-report as a reference standard to determine the validity of PTT measures for the diagnosis of a hearing disability. Martin and Champlin (2000) and Palmer et al (2009) have proposed that self-report may actually be a better indicator of hearing ability than PTTs. Saunders et al (2004) demonstrated that the self-report of speech-recognition-in-noise ability is reliable. A number of studies have been conducted where self-report has been used as a reference standard for the presence or absence of a speech-recognition-in-noise disorder (Table 4). In all of these "patient-centered" studies, participants in the disordered groups reported speech-recognition-in-noise deficits in the presence of normal PTTs.

The Current Study

The purpose of the present investigation was to determine the diagnostic accuracy of the HINT and pure-tone threshold measures as index tests for the detection of a speech-recognition-in-noise disorder (target condition). Self-report was used as the reference standard for this "patient-centered" investigation. Similar to the studies in Table 4, two groups of participants with normal audiograms were tested. Those with the self-report of speech-recognition-in-noise difficulties were assigned to the disordered group, also called the KKS group. The participants without speech-in-noise difficulties were assigned to the control group. Diagnostic accuracy of the index tests was found by determining (a) if the index test results for control participants vary from results for participants in the disordered group, (b) the sensitivity and specificity of the index tests, and (c) ROC curves for the index tests.

The study hypotheses are as follows:

(a) A statistically significant difference will be found between groups for the ability to recognize speech in the presence of background noise using the HINT.
(b) No significant differences will be found between groups for PTT measures.
(c) Better sensitivity and specificity will be found for the HINT than for PTT measures.
(d) ROC curves will reveal a greater area under the curve (AUC) for the HINT than PTT measures.

METHODS

Permission to use the study data was obtained from the internal review board at St. Vincent Medical Center in Los Angeles, California. Audiometric, self-report, and HINT data were collected at the House Ear Institute in Los Angeles. Forty-seven individuals participated in this study. All were native speakers of

American English. Otoscopic visualization revealed clear external ear canals for all participants. PTTs were obtained for 250, 500, 1000, 2000, 3000, 4000, and 6000 Hz. All participants had PTTs ≤25 dB HL (250–6000 Hz). The bilateral PTA was determined for each participant across both ears and for test frequencies 500, 1000, and 2000 Hz. The maximum PTT between ears was also determined. The participants were asked if they had any difficulty understanding speech in a noisy environment, such as a crowded restaurant. Participants who reported some degree of difficulty were assigned to the KKS (disordered) group. Participants with the self-report of no difficulties hearing speech in noise were in the control group. Table 5 shows the similarities between groups for age composition. A two-sample *t* test revealed no significant difference between groups for age ($p = 0.79$).

The American English version of the HINT was used to measure binaural speech perception in steady-state speech-spectrum noise (Nilsson et al, 1994; Vermiglio, 2008). The HINT speech material is an "Americanized" version of the BKB sentences developed by Bench et al (1979) in the United Kingdom. The HINT protocol was modeled in part after the test used by Middelweerd et al (1990) and developed by Plomp and Mimpen (1979). For the HINT administration, speech and noise were presented through headphones using head-related transfer functions from a Knowles Electronics Mannequin for Auditory Research to simulate sound field locations. All virtual sound sources were 1 m from the center of the head in the simulated sound field. Three different noise locations were used: directly in front of the participant ("Noise Front"), 90° to the participant's left ("Noise Left"), and 90° to the participant's right ("Noise Right"). The use of head-related transfer functions allows for the preservation of the head shadow effect when testing the noise side conditions under headphones. The HINT threshold was obtained under each of the three noise conditions to sample a range of binaural directional hearing ability in noise. The standard HINT protocol was used.

The HINT is an adaptive threshold test in which the participant is required to recognize and repeat short English sentences spoken by a male talker. The level of the noise is fixed at 65 dBA. The level of the speech is adaptively varied, depending on the response of the participant. When the participant incorrectly repeats the sentence, the signal-to-noise ratio (SNR) of the next sentence is increased; when the participant correctly repeats the sentence, the SNR for the next sentence is decreased. The SNR is changed in 4-dB step sizes for the first four sentences and in 2-dB step sizes for the remaining sentences. There are 20 sentences in each list. The HINT threshold represents the SNR where the participant correctly recognizes 50% of the sentences.

Table 5. Descriptive Statistics for the Age of the Control and KKS Groups

Group	N	Mean Age (Yr)	Standard Deviation	Minimum (Yr)	Maximum (Yr)
Control	22	36.91	8.28	24	53
KKS	25	36.24	8.65	24	53

The HINT software with a custom digital signal processing sound card was used to present the stimuli in a simulated sound field under TDH-50 headphones (Telephonics Corporation, Huntington, NY). List randomization, stimuli presentation, threshold calculations, polarity matching of the headphones, calibration, and data storage were conducted using the HINT software. The headphone signals were within ±0.5 dB of the desired level. Scoring was based on correct sentence repetition, although substitutions that have minimal effect on meaning were allowed (e.g., verb tense ["is" for "was"] and articles ["a" for "the"]).

From the three noise conditions, a number of derived thresholds or scores were determined. The HINT Composite score is an average of the thresholds for the three noise conditions where the Noise Front threshold is weighted twice using the formula (2 × [Noise Front + Noise Right + Noise Left])/4. This score provides a single index of overall speech recognition in noisy environments. The Average HINT threshold is an average of the thresholds for the Noise Front, Noise Right, and Noise Left conditions. The Directional Advantage refers to the improvement in HINT performance that occurs when the noise is spatially separated from the speech signal. The Directional Advantage (Right) is determined by subtracting the Noise Right threshold from the Noise Front threshold. The Directional Advantage (Left) is determined by subtracting the Noise Left threshold from the Noise Front threshold. The Average Directional Advantage is the average of the right and left Directional Advantages.

All the statistical analyses were performed in JMP Pro 12 (SAS Institute Inc., Cary, NC). The threshold measures were compared between the two groups using two-sample *t* tests. Logistic regression was used to investigate the sensitivity and specificity of each index test. A significance level of 0.05 was adopted for all statistical tests.

RESULTS

The descriptive statistics and the two-sample *t* test (one tailed) results for the HINT and PTT measures are presented in Tables 6 and 7, respectively. For the Noise Front, Noise Right, and Noise Left conditions, the control group had thresholds that were on average 1.14 dB better than the KKS group. A multivariate analysis of variance analysis for these three noise conditions revealed a statistically significant overall group difference ($p < 0.0001$). A significant difference

was found between groups for the HINT Composite score and the Average Noise threshold ($p < 0.0001$). The HINT Composite score for the control group was 1.11 dB better than for the KKS group. This is similar to the results of Middelweerd et al (1990) who found that for the steady-state noise condition, the control group performed 1 dB better than the disordered group. For the HINT, a 1-dB change in threshold corresponds approximately to a 10% change in speech intelligibility in noise (Nilsson et al, 1994; Soli and Wong, 2008). There were no significant differences found between groups for any of the HINT directional advantage conditions. No significant differences were found between groups for the bilateral PTA and maximum PTT measures. The two-sample t test results imply that the HINT noise thresholds are more sensitive to the presence of a speech-recognition-in-noise disorder than the HINT directional advantage measures, and the PTT measures. No significant correlation coefficients were found between the HINT and pure-tone measures. A nonsignificant difference of −0.11 dB was found between the Noise Right and Noise Left thresholds across both groups ($p = 0.5911$). Significant correlations were found between age and bilateral PTA ($r = 0.4518$, $p = 0.0014$) and between age and maximum PTT ($r = 0.5080$, $p = 0.0003$). No significant correlations were found between age and any of the HINT measures.

The sensitivity, specificity, and AUC are presented in Table 8. The index tests have been rank ordered according to AUC first, and then by sensitivity. According to this analysis, the HINT Composite score has the highest and the Directional Advantage (Right) has the lowest diagnostic accuracy. All of the HINT conditions (with the exception of the directional measures) were found to be significant predictors of a speech-recognition-in-noise disorder ($p < 0.05$). Neither of the PTT measures were found to be significant predictors of a speech-recognition-in-noise disorder ($p > 0.05$). The sensitivity and specificity values were found for the threshold or score where the greatest difference was found between 1 − specificity and sensitivity. This value represents the cut point between normal and disordered speech-recognition-in-noise ability. The ROC curves for the HINT Composite score and the bilateral PTA are presented in Figure 1.

The quartiles (25th, 50th, and 75th percentiles) for the standard HINT norms for the HINT Composite score are −5.8, −6.4, and −7.0 dB SNR, respectively (Vermiglio, 2008). Figure 2 shows that the HINT Composite scores for 86% of the control participants were above the second quartile (50th percentile). The HINT Composite scores for 92% of the KKS participants were below the third quartile. Recall that the groups were created according to the participants' reported ability to understand speech in noisy environments, such as a crowded restaurant. The area of overlap between the two distributions may be a reflection of the listening skills and experiences of the participants. For example, in daily life, KKS participants with the best HINT Composite scores may not use visual and/or contextual cues as effectively as the control participants with the poorest scores. No HINT Composite scores for the control group were found below the lowest quartile. Additionally, it is assumed that a participant will report difficulties with speech recognition in noise only if this difficulty exists in his or her daily activities. A participant with little or no exposure to conversations in noise may not report a speech-recognition-in-noise deficit even though one may exist.

Table 6. Descriptive Statistics and Group Differences for HINT Thresholds, Composite Score, and Directional Advantage

Variable	Group	n	Mean (dB SNR)	Standard Deviation	Maximum (dB SNR)	Minimum (dB SNR)	Group Difference (dB)	p
Noise Front threshold	Control	22	−3.41	0.81	−1.90	−5.00	1.00	0.0001
	KKS	25	−2.41	0.87	−0.90	−4.20		
Noise Right threshold	Control	22	−10.74	1.20	−8.10	−12.60	1.09	0.0033
	KKS	25	−9.64	1.42	−6.00	−11.30		
Noise Left threshold	Control	22	−10.76	1.07	−9.10	−13.00	1.34	0.0004
	KKS	25	−9.41	1.47	−5.30	−11.90		
HINT Composite score	Control	22	−7.08	0.63	−5.83	−8.25	1.11	<0.0001
	KKS	25	−5.97	0.88	−3.85	−7.40		
Average Noise threshold	Control	22	−8.30	0.66	−6.87	−9.60	1.15	<0.0001
	KKS	25	−7.16	0.98	−4.67	−8.73		
Directional Advantage (Right)	Control	22	7.32	1.32	10.00	5.00	−0.09	0.5900
	KKS	25	7.23	1.39	9.40	4.20		
Directional Advantage (Left)	Control	22	7.34	1.27	10.00	5.20	−0.34	0.7962
	KKS	25	7.00	1.52	9.50	3.40		
Average Directional Advantage	Control	22	7.33	1.03	9.40	5.55	−0.22	0.7363
	KKS	25	7.12	1.29	9.10	4.15		

Table 7. Descriptive Statistics and Group Differences for Bilateral PTA and Maximum PTT

Variable	Group	n	Mean (dB HL)	Standard Deviation	Maximum (dB HL)	Minimum (dB HL)	Group Difference (dB)	*p*
Bilateral $PTA_{0.5,\ 1.0,\ 2.0\ kHz}$	Control	22	4.62	3.82	15	−0.83	−0.09	0.5284
	KKS	25	4.53	4.60	15	−1.67		
Maximum PTT between ears	Control	22	16.36	5.60	25	5	−0.76	0.6803
	KKS	25	15.60	5.46	25	5		

DISCUSSION

The results of the present study reveal that for individuals with normal PTTs, the HINT has greater diagnostic accuracy than PTT measures for the identification of a speech-recognition-in-noise disorder. This has been demonstrated in three ways. First, as shown in Table 6, significant differences ($p < 0.01$) between control and disordered groups were found for HINT results (Noise Front, Noise Right, Noise Left, HINT Composite score, and the Average Noise thresholds). This is consistent with previous studies that have shown significant differences between control and disordered groups for speech-recognition-in-noise ability (Saunders and Haggard, 1989; Middelweerd et al, 1990). No significant differences were found between groups for the PTT measures (Table 7). This is in contrast with Saunders and Haggard (1989) and Zhao and Stephens (2006) who found significantly poorer PTTs for the KKS group when compared to the control group. Second, the HINT conditions with the greatest diagnostic accuracy were the HINT Composite score, Average HINT threshold, and the Noise Front threshold. The sensitivity for these dependent variables ranged from 80% to 88%, and the specificity ranged from 68% to 86% (Table 8). The sensitivity for the bilateral PTA and maximum PTT was 28% and 56%, respectively. The specificity for bilateral PTA and maximum PTT was 95% and 55%, respectively. Third, according to the ROC curve analysis, all of the HINT conditions (with the exception of the directional measures) were found to be significant ($p < 0.05$) predictors of a speech-recognition-in-noise disorder (Table 8). The HINT Composite score had the highest diagnostic accuracy for a speech-recognition-in-noise disorder (AUC = 0.87). Neither of the PTT measures were significant predictors of a speech-recognition-in-noise disorder.

Specifically, the results of this study have shown that for two groups of participants ranging in age from 18 to 54 yr, the standard HINT protocol demonstrated reasonable diagnostic accuracy for a speech-recognition-in-noise disorder when self-report was used as the reference standard. The diagnostic accuracy of the HINT

Table 8. Diagnostic Accuracy of Pure Tone and HINT Measures

Rank Order	Reference Standard	Index Test (Predictor)	Cut Point	Target Condition	Sensitivity (%)	Specificity (%)	*p*	AUC
1	Self-report	HINT Composite score	−6.65 dB SNR	Speech recognition in noise disorder	88	77	0.0012	0.87
2	Self-report	Average HINT threshold	−7.87 dB SNR	Speech recognition in noise disorder	80	86	0.0013	0.86
3	Self-report	HINT Noise Front threshold	−3.00 dB SNR	Speech recognition in noise disorder	80	68	0.0022	0.81
4	Self-report	HINT Noise Left threshold	−9.60 dB SNR	Speech recognition in noise disorder	60	82	0.0043	0.77
5	Self-report	HINT Noise Right threshold	−9.10 dB SNR	Speech recognition in noise disorder	40	96	0.0151	0.66
6	Self-report	Directional Advantage (Left)	7.70 dB	Speech recognition in noise disorder	64	50	0.4052	0.54
7	Self-report	Maximum PTT	15 dB HL	Speech recognition in noise disorder	56	55	0.6305	0.54
8	Self-report	Average Directional Advantage	7.15 dB	Speech recognition in noise disorder	56	59	0.5250	0.53
9	Self-report	Bilateral PTA	0 dB HL	Speech recognition in noise disorder	28	95	0.9423	0.51
10	Self-report	Directional Advantage (Right)	4.90 dB	Speech recognition in noise disorder	12	100	0.8159	0.51

Notes: Index tests are ranked by the AUC first, and then by sensitivity. The cut point represents the threshold or score that separates normal from disordered speech-recognition-in-noise ability.

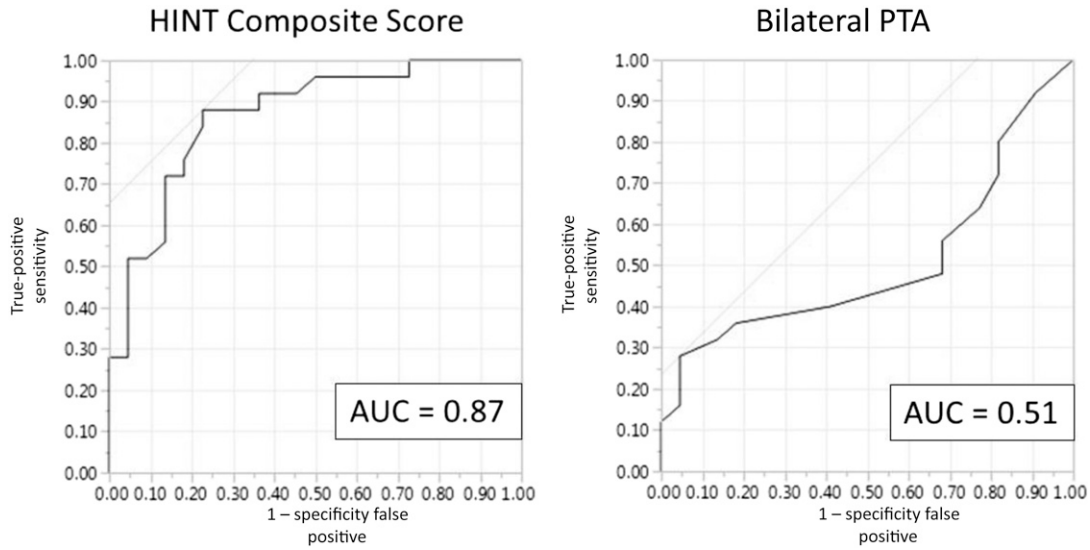

Figure 1. ROC curves for the HINT Composite score and the bilateral PTA.

for a speech-recognition-in-noise disorder cannot be applied to any other target disorders, such as cochlear, VIIIth nerve, auditory brainstem, or auditory cortex sites of lesion. Separate studies would need to be conducted for alternate target conditions. Additionally, appropriate reference standards would need to be determined for each of these studies.

The Question(s) Used for Self-Report

The self-report of hearing ability has been determined through the use of a single question or with questionnaires. Recall that the reference standard should be the best method to determine the presence or absence of a target condition. Therefore, the question(s) should be relevant to the target condition. The question used in the present study is specific to a speech-recognition-in-noise disorder. Questions such as "Do you feel you have a hearing loss?" (Nondahl et al, 1998; Sindhusake et al, 2001) and questionnaires that review multiple aspects of audition may not be entirely relevant to a single target condition. Since it is possible for an individual to have no deficit for the ability to hear in quiet, and yet have difficulty with the ability to hear in noise (Vermiglio et al, 2012), the conflation of these two components of audition in a self-report questionnaire may be misleading.

There are two separate components of the ability to hear that have been addressed in the literature. Carhart (1951) identified two aspects of hearing: acuity and clarity. A loss of acuity refers to a loss of audibility. A deficiency in clarity refers to a loss of intelligibility when the signal is audible. In subsequent years, Plomp (1978; 1986) adopted the terms "audibility" and "distortion." He proposed a model to characterize audibility and distortion losses based on the speech reception threshold. Audibility refers to hearing sensitivity. Distortion refers a loss of clarity, where the signal is audible but unintelligible. This distortion may occur in quiet

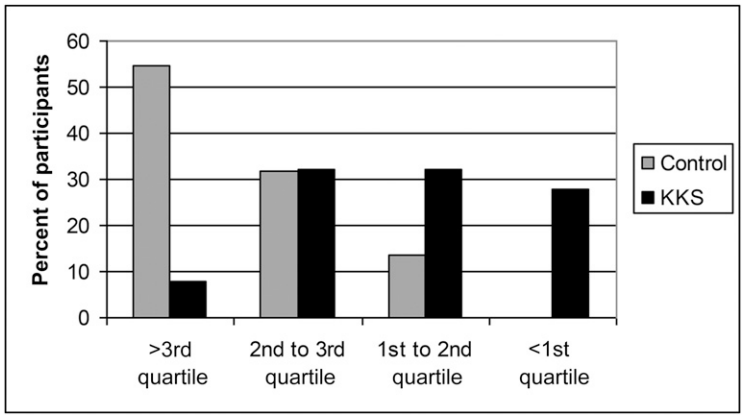

Figure 2. Bar graph of interquartile distributions of HINT Composite scores for control and KKS groups. The HINT Composite score is the average of the thresholds for the three noise conditions where the threshold for Noise Front is weighted twice.

or in noisy environments. Ward (1983) observed two classes of listening errors: (a) failure to hear an acoustic event and (b) the incorrect perception of an audible acoustic event. Vermiglio et al (2012) showed the relationship between PTA (attenuation) and speech-recognition-in-noise ability (distortion). The results revealed that $PTA_{0.5,\ 1.0,\ 2.0\ kHz}$ was not significantly correlated with the ability to recognize speech in steady-state noise. However, $PTA_{0.5,\ 1.0,\ 2.0\ kHz}$ was significantly correlated with the ability to recognize speech in quiet ($r = 0.800$, $p < 0.05$). The authors presented data showing that the presence of normal PTTs ≤15 dB HL (2.0–6.0 kHz) was not a sufficient condition for normal speech-recognition-in-noise performance.

Which Reference Standard Is the "Best?"

For diagnostic accuracy studies, the "best" reference standard is one that most accurately determines the presence or absence of a target condition. The clinical or research question should provide guidance when identifying an appropriate reference standard. If the question is in regard to how well an index test identifies an audibility disorder for tonal stimuli, then PTT measures would be a reasonable reference standard in this "test-centered" approach. If the question is in regard to how well an index test identifies a speech-recognition-in-noise disorder that is perceived by the participant, then self-report would be a reasonable reference standard for this "patient-centered" approach.

Recall that Wilson et al (2007) attributed the diagnostic accuracy of the speech-recognition-in-noise tests used for the identification of one target condition (pure-tone hearing loss) to a different target condition (a speech-recognition-in-noise disorder). The assumption that the diagnostic accuracy of an index test for one target condition is relevant for the diagnostic accuracy of that index test for a different target condition is misleading. For example, Anderson et al (2013) evaluated the efficacy of the complex auditory brainstem response (cABR) versus a speech-recognition-in-noise test for the prediction of a self-reported speech-recognition-in-noise disorder. They argued that an objective measure such as the cABR is needed to predict "real-world" speech-recognition-in-noise performance because the cABR is unaffected by cognitive status, as may be found with behavioral measures. The authors wrote, "We assessed SIN [speech in noise] performance with the QuickSIN because of its widespread clinical use and its superior ability to separate performance between groups of participants with normal hearing and groups of participants with hearing impairment compared with other tests containing sentences, such as the BKB-SIN or HINT (Wilson et al, 2007)." The fallacy of target displacement is found in this sentence. The authors assessed speech-recognition-in-noise performance not with a test known for its diagnostic accuracy for a speech-recognition-in-noise disorder, but for its diagnostic accuracy to identify a disorder of hearing sensitivity. Diagnostic accuracy of an index test for one target condition must be measured directly; it cannot be inferred from a study for a different target condition. Results of the present investigation have demonstrated that for individuals with normal audiograms, PTT measures have a very poor level of diagnostic accuracy for the detection of a speech-recognition-in-noise disorder when compared to the HINT.

It is not possible to select an appropriate reference standard unless the target condition is clearly understood. Vermiglio (2014) has proposed the Sydenham–Guttentag criteria for the identification of legitimate target conditions, also known as clinical entities. According to these criteria, a clinical entity (legitimate diagnostic target, or target condition) is a disorder with an unambiguous definition (Sydenham, 1676 quoted in Meynell, 2006; FDA, 2000), it represents a homogeneous patient group (Sydenham, 1676 quoted in Meynell, 2006; Guttentag, 1949; 1950; FDA, 2000), it represents a perceived limitation for the patient (Guttentag, 1949), and it facilitates diagnosis and intervention (Sydenham, 1676 quoted in Meynell, 2006; Guttentag, 1949; FDA, 2000). Vermiglio has argued that a speech-recognition-in-noise disorder is a clinical entity according to the Sydenham–Guttentag criteria. Moreover, because this target condition is clearly understood, the procurement of a reasonable reference standard is attainable.

The Law of the Instrument and PTTs

According to Wilson and Margolis (2015) "the term 'normal hearing' as used in clinic and research reports, is almost exclusively based on PTTs." They noted, "The practice of defining hearing loss based on hearing sensitivity measures is the expected result of the Law of the Instrument, usually attributed to Abraham Maslow." Maslow (1966) wrote, "I remember seeing an elaborate and complicated automatic washing machine for automobiles that did a beautiful job of washing them. But it could do only that, and everything else that got into its clutches was treated as if it were an automobile to be washed. I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."

Buffett (1984) used this expression when criticizing studies of financial markets that incorporate inappropriate mathematical techniques. He said, "It isn't necessarily because such studies have any utility; it's simply that the data are there and academicians have worked hard to learn the mathematical skills needed to manipulate them. Once these skills are acquired, it seems sinful not to use them, even if the usage has no utility or negative utility. As a friend said, to a man with a hammer, everything looks like a nail." The same could be said of PTTs when they are used as the "gold standard" for the

assessment of the ability to hear, even when their limited utility has been shown through numerous studies (NIH, 1938; Fry, 1942; 1961; Merluzzi and Hinchcliffe, 1973; Saunders and Haggard, 1989; Middelweerd et al, 1990; Lutman and Saunders, 1992; Rappaport et al, 1993; Martin and Champlin, 2000; Dobie, 2011; Tremblay et al, 2015; Zhao and Stephens, 2006).

To say that an index test has high sensitivity and specificity is meaningless unless the target condition and reference standard are clearly described. In addition, just as several studies have been conducted for the validation of PTTs, multiple studies should be conducted to investigate the diagnostic accuracy of speech-recognition-in-noise test protocols. Index tests with reasonably high levels of diagnostic accuracy may be candidates for use as reference standards in subsequent studies. For example, the magnetic resonance imaging (MRI) scan has been used as an index test for various target disorders. Once the diagnostic accuracy of the MRI was found to be relatively high, the next step was to use it as a reference standard for that target condition. House et al (1986) determined the diagnostic accuracy of the MRI for the detection of acoustic neuromas. Observation during surgery served as the reference standard. According to Chandrasekhar et al (1995), the MRI quickly evolved to become the "gold standard" for the diagnosis of acoustic neuromas following the introduction of gadolinium-DPTA in 1987. The same sequence of events may be realized for speech-recognition-in-noise tests. Speech-in-noise tests that exhibit a high level of diagnostic accuracy may be used as reference standards. Anderson et al (2010) used the HINT as a reference standard to determine the diagnostic accuracy of an auditory brainstem response protocol for the detection of a speech-recognition-in-noise disorder. Berlin (2012) in a discussion of Auditory Neuropathy Spectrum Disorder stated that "audiologists generally focus on audiograms because the audiogram has become the 'gold standard' of hearing ability, but the real 'gold standard' may actually be their speech in noise results."

CONCLUSION AND RECOMMENDATIONS

Self-report has been used as a reference standard for many "patient-centered" diagnostic accuracy studies in the field of audiology (Bunch, 1929; NIH, 1938; Steinberg et al, 1940; Glorig, 1956; Merluzzi and Hinchcliffe, 1973; Saunders and Haggard, 1989; Middelweerd et al, 1990; Lutman and Saunders, 1992; Rappaport et al, 1993; Martin and Champlin, 2000; Zhao and Stephens, 2006; Dobie, 2011; Anderson et al, 2013; Tremblay et al, 2015). Saunders and Forsline (2006) have shown that self-report is reliable. Diagnostic accuracy studies have also used behavioral tests as a reference standard for "test-centered" diagnostic accuracy studies (Nondahl et al, 1998; Sindhusake et al, 2001; Wilson et al, 2007; Anderson et al, 2010; Hannula et al, 2011; Koole et al, 2016). Wilson et al (2007) used PTTs as a reference standard to determine the diagnostic accuracy of speech-recognition-in-noise tests. However, the authors inferred that the diagnostic accuracy values for one target condition (elevated PTA) were applicable to a separate target condition (speech-recognition-in-noise disorder). This is an example of target displacement (Vermiglio, 2016).

The present study has shown that for individuals with normal audiograms, the HINT has greater diagnostic accuracy for the identification of a speech-recognition-in-noise disorder than PTT measures. This has been demonstrated by group differences, sensitivity and specificity values, and the AUC for ROC curves. Diagnostic accuracy values are specific to the study's methodology. The best reference standard is one that independently verifies the presence or absence of a target condition. Only legitimate target conditions (clinical entities) allow for the procurement of reasonable reference standards. PTTs may have become the "gold standard" in audiology due to the Law of the Instrument. This occurs when a tool is used even when it possesses limited utility.

Diagnostic accuracy studies for speech-recognition-in-noise tests will allow clinicians to make informed decisions regarding test selection. The use of self-report as a reference standard in a diagnostic accuracy study enables clinicians to compare the self-report of the individual patient, along with their speech-in-noise test results, to the results of groups of research participants with and without the self-report of speech-recognition-in-noise difficulties. For example, if a patient reports no speech-in-noise difficulties but his speech-in-noise performance is below normal, then the patient could be counseled that even though they reported no difficulty, their test performance was similar to a group of participants who reported difficulties. Conversely, if the patient reports a speech-in-noise difficulty but their speech-in-noise performance is within normal limits, the patient could be counseled that their self-report is inconsistent with individuals who also performed within normal limits. In both scenarios, information from diagnostic accuracy studies would be useful for the development of management plans.

Future studies of diagnostic accuracy should include, at a minimum, an analysis of differences between control and disordered groups, sensitivity and specificity values, and ROC curves. It is imperative that the index test protocol be clearly described to allow for replication (Bossuyt et al, 2003). The target condition should meet the Sydenham–Guttentag criteria for a clinical entity (Vermiglio, 2014). This will allow for the procurement of a reasonable reference standard for the independent verification of the presence or absence of the target condition (Vermiglio, 2016). The rationale for the reference standard should be stated. The reference standard should

be clearly described and it should be applied to all of the participants in the disordered and control groups (Bossuyt et al, 2003). When self-report is the reference standard, the question(s) used should address a homogeneous disorder (e.g., speech recognition in noise, speech recognition in quiet, or a sound localization disorder). The description of the reference standard, index test, or target condition should not be ambiguous. Ambiguity is a hallmark of the APD construct. This ambiguity leads to uncertainty in the test protocols, reference standard, and intervention (Vermiglio, 2014; 2016).

Acknowledgments. Andrew J. Vermiglio thanks Brenda Vermiglio, MA, and Gregg Givens, PhD, for their helpful comments. He also thanks Dan Freed, MS, for his very helpful critiques of the early manuscripts and his development of the software for the speech-recognition-in-noise test used in this study.

REFERENCES

American Academy of Otolaryngology (AAO). (1979) Guide for the evaluation of hearing handicap. *JAMA* 241(19):2055–2059.

Alvord LS. (1983) Cochlear dysfunction in "normal-hearing" patients with history of noise exposure. *Ear Hear* 4(5):247–250.

Anderson S, Parbery-Clark A, White-Schwoch T, Kraus N. (2013) Auditory brainstem response to complex sounds predicts self-reported speech-in-noise performance. *J Speech Lang Hear Res* 56:(1):31–43.

Anderson S, Skoe E, Chandrasekaran B, Kraus N. (2010) Neural timing is linked to speech perception in noise. *J Neurosci* 30(14): 4922–4926.

Baiduc RR, Poling GL, Hong O, Dhar S. (2013) Clinical measures of auditory function: the cochlea and beyond. *Dis Mon* 59(4): 147–156.

Beasley WC. (1940) Characteristics and distribution of impaired hearing in the population of the United States. *J Acoust Soc Am* 12:114.

Bench J, Kowal A, Bamford J. (1979) The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *Br J Audiol* 13(3):108–112.

Berkson J. (1947) Cost-utility as a measure of the efficiency of a test. *J Am Stat Assoc* 42(238):246–255.

Berlin CI. (2012) Auditory neuropathy spectrum disorder, OAEs, ABR, and more. www.audiology.org/news/interview-charles-berlin-phd. Accessed August 25, 2016.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG. (2003) ; Standards for Reporting of Diagnostic Accuracy Group. (2003) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Croat Med J* 44(5):639–650.

Buffett WE. (1984) The Superinvestors of Graham-and-Doddsville. *Hermes: the Columbia Business School Magazine*, 4–15.

Bunch CC. (1929) Age variations in auditory acuity. *Arch Otolaryngol* 9:625–636.

Carhart R. (1951) Basic principles of speech audiometry. *Acta Otolaryngol* 40(1–2):62–71.

Chandrasekhar SS, Brackmann DE, Devgan KK. (1995) Utility of auditory brainstem response audiometry in diagnosis of acoustic neuromas. *Am J Otol* 16(1):63–67.

Dadson RS, King JH. (1952) A determination of the normal threshold of hearing and its relation to the standardization of audiometers. *J Laryngol Otol* 66(8):366–378.

Dobie RA. (2011) The AMA method of estimation of hearing disability: a validation study. *Ear Hear* 32(6):732–740.

Etymotic. (2005) *Bamford-Kowal-Bench Speech-in-Noise Test (Version 1.03) [Audio CD].* Elk Grove Village, IL: Etymotic Research.

Food and Drug Administration, US (FDA). (2000) Division of Neuropharmacological Drug Products (DNDP) Issues Paper for March 9, 2000, Meeting of the Psychopharmacological Drugs Advisory Committee Meeting on the Various Psychiatric and Behavioral Disturbances Associated with Dementia. www.fda.gov/ohrms/dockets/dockets/00n0088/bkg0001.pdf. Accessed September 20, 2013.

Fry DB. (1942) A suggestion for a new method of testing hearing in aviation candidates. *J Laryngol Otol* 57(1):11–13.

Fry DB. (1961) Word and sentence tests for use in speech audiometry. *Lancet* 2(7195):197–199.

Glorig A. (1956) Determination of the normal hearing reference zero. *J Acoust Soc Am* 28(6):1110–1113.

Guttentag OE. (1949) On the clinical entity. *Ann Intern Med* 31(3): 484–496.

Guttentag OE. (1950) Two diagrams on the clinical entity. *J Pediatr* 37(4):530–534.

Hannula S, Bloigu R, Majamaa K, Sorri M, Mäki-Torkko E. (2011) Self-reported hearing problems among older adults: prevalence and comparison to measured hearing impairment. *J Am Acad Audiol* 22(8):550–559.

Hinchcliffe R. (1961) Prevalence of the commoner ear, nose, and throat conditions in the adult rural population of Great Britain. A study by direct examination of two random samples. *Br J Prev Soc Med* 15:128–140.

Hinchcliffe R. (1992) King-Kopetzky syndrome: an auditory stress disorder? *J Audiol Med* 1:89–98.

House JW, Waluch V, Jackler RK. (1986) Magnetic resonance imaging in acoustic neuroma diagnosis. *Ann Otol Rhinol Laryngol* 95(1 Pt 1):16–20.

Jerger J. (2009) *Audiology in the USA*. San Diego, CA: Plural Publishing.

Johnston KN, John AB, Kreisman NV, Hall JW 3rd, Crandell CC. (2009) Multiple benefits of personal FM system use by children with auditory processing disorder (APD). *Int J Audiol* 48(6): 371–383.

Kiely KM, Gopinath B, Mitchell P, Browning CJ, Anstey KJ. (2012) Evaluating a dichotomized measure of self-reported hearing loss against gold standard audiometry: prevalence estimates and age bias in a pooled national data set. *J Aging Health* 24(3):439–458.

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. (2004) Development of a quick speech-in-noise test for measuring

signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 116(4 Pt 1):2395–2405.

King PF. (1954a) Psychogenic deafness. *J Laryngol Otol* 68(9): 623–635.

King PF. (1954b) Psychogenic deafness [Abstract]. *Proc R Soc Med* 47(11):941–942.

Koole A, Nagtegaal AP, Homans NC, Hofman A, Baatenburg de Jong RJ, Goedegebure A. (2016) Using the Digits-In-Noise Test to estimate age-related hearing loss. *Ear Hear* 37(5):508–513.

Kuk F, Jackson A, Keenan D, Lau CC. (2008) Personal amplification for school-age children with auditory processing disorders. *J Am Acad Audiol* 19(6):465–480.

Lane CE. (1922) Minimum sound energy for audition for tones of high frequency. *Phys Rev* 19:492.

Liberman MC, Epstein MJ, Cleveland SS, Wang H, Maison SF. (2016) Toward a differential diagnosis of hidden hearing loss in humans. *PLoS One* 11(9):e0162726.

Lutman ME, Saunders GH. (1992) Lack of association between otoacoustic emissions and hearing difficulty in subjects with normal hearing thresholds. *J Acoust Soc Am* 92(2 Pt 1):1184–1185.

Manne SL, Jacobsen PB, Redd WH. (1992) Assessment of acute pediatric pain: do child self-report, parent ratings, and nurse ratings measure the same phenomenon? *Pain* 48(1):45–52.

Martin FN, Champlin CA. (2000) Reconsidering the limits of normal hearing. *J Am Acad Audiol* 11(2):64–66.

Maslow AH. (1966) *The Psychology of Science: A Reconnaissance*. Chapel Hill, NC: Maurice Bassett Publishing.

Mendel LL. (2007) Objective and subjective hearing aid assessment outcomes. *Am J Audiol* 16(2):118–129.

Merluzzi F, Hinchcliffe R. (1973) Threshold of subjective auditory handicap. *Audiology* 12(2):65–69.

Merriam-Webster Dictionary. (2016) Definition of subjective in English. www.merriam-webster.com/dictionary/subjective. Accessed January 22, 2017.

Meyer-Bisch C. (1996) Audioscan: a high-definition audiometry technique based on constant-level frequency sweeps–a new method with new hearing indicators. *Audiology* 35(2):63–72.

Meynell GG. (2006) John Locke and the preface to Thomas Sydenham's Observationes medicae. *Med Hist* 50(1):93–110.

Middelweerd MJ, Festen JM, Plomp R. (1990) Difficulties with speech intelligibility in noise in spite of a normal pure-tone audiogram. *Audiology* 29(1):1–7.

National Institutes of Health (NIH). (1938) *Hearing Study Series: Normal Hearing for Speech at Each Decade of Life*. Washington, DC: NIH.

Nilsson M, Soli SD, Sullivan JA. (1994) Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am* 95(2):1085–1099.

Niquette P, Arcaroli J, Revit L, Parkinson A, Staller S, Skinner M, Killion M. (2003) Development of the BKB-SIN Test. Paper presented at the American Auditory Socitey, Scottsdale, AZ.

Noble W. (1988) Evaluation of hearing handicap: a critique of Ward's position. *Audiology* 27(1):53–64.

Nondahl DM, Cruickshanks KJ, Wiley TL, Tweed TS, Klein R, Klein BE. (1998) Accuracy of self-reported hearing loss. *Audiology* 37(5):295–301.

Palmer CV, Solodar HS, Hurley WR, Byrne DC, Williams KO. (2009) Self-perception of hearing ability as a strong predictor of hearing aid purchase. *J Am Acad Audiol* 20(6):341–347.

Peterson WW, Birdsall TG, Fox WC. (1954) The theory of signal detectability. *Trans IRE Profession Group Inform Theory* 4(4): 171–212.

Plomp R. (1978) Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am* 63(2):533–549.

Plomp R. (1986) A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *J Speech Hear Res* 29(2): 146–154.

Plomp R, Mimpen AM. (1979) Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 18(1): 43–52.

Pryce H, Metcalfe C, Hall A, Claire LS. (2010) Illness perceptions and hearing difficulties in King-Kopetzky syndrome: what determines help seeking? *Int J Audiol* 49(7):473–481.

Rappaport JM, Phillips DP, Gulliver JM. (1993) Disturbed speech intelligibility in noise despite a normal audiogram: a defect in temporal resolution? *J Otolaryngol* 22(6):447–453.

Saunders GH, Forsline A. (2006) The Performance-Perceptual Test (PPT) and its relationship to aided reported handicap and hearing aid satisfaction. *Ear Hear* 27(3):229–242.

Saunders GH, Forsline A, Fausti SA. (2004) The performance-perceptual test and its relationship to unaided reported handicap. *Ear Hear* 25(2):117–126.

Saunders GH, Haggard MP. (1989) The clinical assessment of obscure auditory dysfunction–1. Auditory and psychological factors. *Ear Hear* 10(3):200–208.

Shargorodsky J, Curhan SG, Curhan GC, Eavey R. (2010) Change in prevalence of hearing loss in US adolescents. *JAMA* 304(7): 772–778.

Sindhusake D, Mitchell P, Smith W, Golding M, Newall P, Hartley D, Rubin G. (2001) Validation of self-reported hearing loss. The Blue Mountains Hearing Study. *Int J Epidemiol* 30(6):1371–1378.

Sivian LJ, White SD. (1933) On minimum audible sound fields. *J Acoust Soc Am* 4:288–321.

Soli SD, Wong LL. (2008) Assessment of speech intelligibility in noise with the Hearing in Noise Test. *Int J Audiol* 47(6):356–361.

Steinberg JC, Montgomery HC, Gardner MB. (1940) Results of the World's Fair hearing tests. *J Acoust Soc Am* 12:533–562.

Stilma W, Rijkenberg S, Feijen HM, Maaskant JM, Endeman H. (2015) Validation of the Dutch version of the critical-care pain observation tool. *Nurs Crit Care.*

Sweetow RW, Sabes JH. (2006) The need for and development of an adaptive Listening and Communication Enhancement (LACE) Program. *J Am Acad Audiol* 17(8):538–558.

Swets JA. (1988) Measuring the accuracy of diagnostic systems. *Science* 240(4857):1285–1293.

Swets JA, Dawes RM, Monahan J. (2000) Better decisions through science. *Sci Am* 283(4):82–87.

Tremblay KL, Pinto A, Fischer ME, Klein BE, Klein R, Levy S, Tweed TS, Cruickshanks KJ. (2015) Self-reported hearing difficulties among adults with normal audiograms: the beaver dam offspring study. *Ear Hear* 36(6):e290–e299.

Vermiglio AJ. (2008) The American English hearing in noise test. *Int J Audiol* 47(6):386–387.

Vermiglio AJ. (2014) On the clinical entity in audiology: (central) auditory processing and speech recognition in noise disorders. *J Am Acad Audiol* 25(9):904–917.

Vermiglio AJ. (2016) On diagnostic accuracy in audiology: central site of lesion and central auditory processing disorder studies. *J Am Acad Audiol* 27(2):141–156.

Vermiglio AJ, Soli SD, Freed DJ, Fisher LM. (2012) The relationship between high-frequency pure-tone hearing loss, hearing in noise test (HINT) thresholds, and the articulation index. *J Am Acad Audiol* 23(10):779–788.

Ward WD. (1983) The American Medical Association/American Academy of Otolaryngology formula for determination of hearing handicap. *Audiology* 22(4):313–324.

Wheeler LJ, Dickson ED. (1952) The determination of the threshold of hearing. *J Laryngol Otol* 66(8):379–395.

Wilson RH. (2003) Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *J Am Acad Audiol* 14(9):453–470.

Wilson RH, Burks CA. (2005) Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: a clinic protocol. *J Rehabil Res Dev* 42(6):839–852.

Wilson RH, Margolis RH. (2015) Hearing loss terminology should be evidence based: a reply to Clark and Martin (2014). *J Am Acad Audiol* 26(5):524–525.

Wilson RH, McArdle RA, Smith SL. (2007) An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *J Speech Lang Hear Res* 50(4):844–856.

Zecker SG, Hoffman HJ, Frisina R, Dubno JR, Dhar S, Wallhagen M, Kraus N, Griffith JW, Walton JP, Eddins DA, Newman C, Victorson D, Warrier CM, Wilson RH. (2013) Audition assessment using the NIH Toolbox. *Neurology* 80(11, Suppl 3):S45–S48.

Zhao F, Stephens D. (1999) Audioscan testing in patients with King-Kopetzky syndrome. *Acta Otolaryngol* 119(3):306–310.

Zhao F, Stephens D. (2006) Distortion product otoacoustic emissions in patients with King-Kopetzky syndrome. *Int J Audiol* 45(1):34–39.

Zuehl BF. (1922) Measurement of auditory acuity with the Iowa pitch range audiometer. *Psychol Monogr* 31(1):83–97.