The development of the Situated Phoneme (SiP) test
The development of the Situated Phoneme (SiP) test
A Swedish test of phonemic discrimination in noise for adult people with hearing loss
Abstract


In the current thesis, a Swedish phoneme-level speech-audiometric test in natural background noise was developed. The test is called the Situated Phoneme (SiP) test. In the first study, different types of word metrics thought to influence lexical access were developed and calculated for more than 800 000 phonetically transcribed Swedish words, which were then assembled in a database called the AFC-list. In the second study, groups of monosyllabic AFC-list words with minimal phonemic contrast were selected as linguistic stimuli for the SiP-test, using a method by which the influence of word frequency, neighborhood density, phonotactic probability and orthographic transparency was controlled. All test words were recorded to sound files, of which the accuracy was validated in a listening experiment with 28 normal-hearing adult native speakers of Swedish. In the third study, a method was developed by which realistic masker sounds, spectrally matched to each set of test phonemes in the SiP-test material, were generated for the SiP-test based on a database of urban outdoor sound events. In the fourth study, the validity of six statistical methods for significance testing of observed score differences applicable to the SiP-test were investigated. Analyses were based both on computer simulated test sessions and on SiP-test sessions with human participants. In the latter, the SiP-test speech material was presented against the urban outdoor masker sounds at different difficulty levels to 74 people with normal hearing to severe hearing loss in a listening experiment using a multiple-alternative forced choice paradigm. Based on the results, a computational prediction model for the SiP-test was developed, by which the underlying success probability of specific SiP-test trials could be estimated. In turn, this enabled the use of significance-test methods based on the Poisson’s binomial distribution, resulting in improved significance-test validity. In addition, the human SiP-test results were analyzed in terms of test-retest reliability, learning effects, content-, construct- and criterion validity within the remains of the thesis.

Keywords: hearing impairment, speech audiometry, phonemic discrimination, realistic masking noise, critical differences, Swedish

Erik Witte, Audiological Research Centre, Faculty of Medicine and Health, Örebro University, SE-701 82 Örebro, Sweden, erik.witte@oru.se
Included studies

The present thesis is based on the following four studies, referred to in the text by their roman numerals:


II. Witte, E., Ekeroot, J., & Köbler, S. (2020). The development of linguistic stimuli for the Swedish situated phoneme test. [Revised manuscript submitted for publication]. Audiological Research Center, Faculty of Medicine and Health, Örebro University.


# Table of Contents

INTRODUCTION .................................................................................. 11
Consequences of hearing loss ............................................................... 11
Hearing rehabilitation ............................................................................ 12
Evaluation of hearing rehabilitation ....................................................... 13
Limitations of current speech-audiometry tests ....................................... 14
  Test-score variability ........................................................................... 14
  Floor and ceiling effects on the phonemic level ................................... 17
  Learning effects ................................................................................... 18
  The choice of contextual setting ........................................................ 18
  Influences from lexical properties......................................................... 19

A NEW APPROACH IN SWEDISH SPEECH AUDIOMETRY ............. 21
The Situated Phoneme test ................................................................. 21
  The basic design .................................................................................. 21
Swedish word metrics ........................................................................... 22
Selection and creation of linguistic stimuli ........................................... 23
Homogenization of speech-sound difficulty levels ............................... 23
Significance-testing methods ............................................................... 24
Estimation of success probability of specific SiP-test trials .................. 25
Thesis aims .............................................................................................. 26
  Specific aims ........................................................................................ 27

METHODS ............................................................................................. 28
Study I ..................................................................................................... 29
Study II .................................................................................................... 30
Study III .................................................................................................. 31
Study IV .................................................................................................. 32
Test-retest reliability and learning effects ............................................. 34
Comparisons with established audiological measures .......................... 36
Homogenization of test-phoneme difficulty levels ................................ 37
Ethical considerations ............................................................................. 38

RESULTS ................................................................................................ 39
Study I ..................................................................................................... 39
Study II .................................................................................................... 39
Study III .................................................................................................. 39
Study IV .................................................................................................. 41
Test-retest reliability and learning effects ............................................. 41
Introduction

This is a study within the academic field of disability research as outlined by the Swedish Institute of Disability Research (SIDR) during the past twenty years. The academic tradition within the SIDR has fostered studies related to various kinds of disabilities such as different types of mental impairment, hearing impairment, other voice- and speech functions, as well as impairments of neuromuscular skeletal and movement-related functions (Strandberg et al., 2017).

Traditionally, the difficulties experienced by persons with impairments have been attributed either, from a medical perspective, to the physical or mental impairment of the disabled persons, or from a social perspective, to disabling factors in the environment in which the persons with disability reside (Cf. Priestley, 1998). From the critical realist perspective – a perspective that has become an important ontological and epistemological basis for disability research – there is always a complex interplay of mechanisms located at different hierarchically ordered levels of reality. Even though one may define many such ontological levels, Danermark (2002), for instance, describes five levels, namely the molecular, the biological, the psychological, the social and the cultural levels. In order to correctly describe and explain complex phenomena such as human disability, the interplay of mechanisms located at several such levels of reality may often need to be identified through interdisciplinary integration of theories and research methods developed within a range of scientific traditions (Danermark, 2002).

Along the same line of thought, this thesis integrates concepts and methods developed within academic disciplines such as phonetics and phonology, corpus linguistics, natural language processing, psycholinguistics, computational modelling of the human auditory system, acoustics and digital signal processing, technology and computer programming, statistics, human ecology, auditory physiology as well as audiological assessment methodology in order to get a valid and reliable answer to one single question. Namely, how well does the speech perception of persons with hearing loss function in common everyday situations?

Consequences of hearing loss

For hearing communities all-over the world, the ability to perceive spoken language in the auditory domain is of vital importance for successful interpersonal communication. The estimation of the World Health
Organization is that 466 million people today have a disabling degree of hearing loss (World Health Organization, 2020a). In Sweden, 1.5 million people, 16 years or older, (18.3 %) have a self-reported level of hearing loss that causes difficulties in participating in conversations with several people at the same time (Statistics-Sweden, 2019b).

As a consequence of hearing loss, Statistics-Sweden estimates that about 38 000 people aged 16-64 years in Sweden today have a reduced working ability (Statistics-Sweden, 2019a). In addition to increasing the risk of long-term sickness absence in the working population (Friberg et al., 2012), hearing loss may also increase the risk of cognitive decline in the older population (Lin et al., 2013). Thus, hearing loss may have far-reaching both social and psychological consequences for the people affected.

**Hearing rehabilitation**

On a societal scale, unaddressed hearing loss leads to costs within the educational sector and costs due to workforce productivity loss that by large exceed the health care system costs for hearing-loss rehabilitation (World Health Organization, 2017). Expressed as disability-adjusted life years, the total cost of hearing loss is estimated, in monetary terms, to be more than five times as high as the total costs of hearing loss within the health care system. This relation is similar both in the European context and worldwide (WHO, 2017).

The single most common hearing-rehabilitation intervention in Sweden today is hearing-aid fitting. Of the 1.5 million Swedes with self-reported hearing loss referred to above, approximately 409 000 use hearing aids (Statistics-Sweden, 2019b). Modern hearing aids are head born sound amplification systems with advanced built-in digital signal processing abilities, such as frequency-specific amplification and compression, directional microphones, noise-reduction schemes, frequency lowering, and adaptive sound environment classifiers (Dillon, 2012). Normally, a trained audiologist adjusts the signal-processing schemes of hearing aids specifically for each hearing-rehabilitation patient, starting from either of a range of possible presets called prescription methods. The prescription methods, in turn, are based on the pure-tone audiogram of the patient, and additionally sometimes on factors such as age, gender, hearing-aid experience as well as language of the subject (Keidser et al., 2011).
Evaluation of hearing rehabilitation

According to Swedish legislation, professionals who prescribe medical devices such as hearing aids are responsible to ensure that the medical device meets the needs of the user in an adequate way (Socialstyrelsen, 2008). Evaluation of hearing-loss rehabilitation can be performed in different ways, both subjectively, using validated questionnaires, and more objectively using high-quality speech-audiometry tests (ISO 21388, 2020). In the Swedish language today, a number of such speech-audiometry tests are available such as the Sahlgrenska University Speech test (the SUS-test) (Grunditz & Magnusson, 2013), the Swedish Hearing In Noise test (HINT) (Hällgren et al., 2006), the Hagerman’s sentence test, internationally often referred to as the Matrix test (Hagerman, 1982), the Swedish test of phonemically balanced words in speech-weighted noise (SPBN) (Magnusson, 1995), as well as the phonemically balanced 50-item word lists (PB50) originally developed by Lidén and Fant (1954).

The Swedish PB50 speech material has the same format as the well-known American C.I.D. Auditory Test W-22 (Hirsh et al., 1952) and consists of 12 lists each containing 50 monosyllabic words. The test procedure for the PB50 word lists is rather straightforward, and consists of playing test-word recordings, most often at a fixed speech level, and after each test-word presentation the subject is required to give a verbal response as to which test word was perceived. The SPBN and the SUS-test use a subset of the word lists from the PB50 test, along with the same recordings mixed with background noise. The SUS-test uses two test lists mixed at a fixed signal-to-noise (SNR) ratio of + 6 dB with one of the fluctuating ICRA-noises (№ 7) (Dreschler et al., 2001), and the SPBN-test uses six test lists mixed at a fixed SNR of + 4 dB with a speech-weighted stationary noise. The task is the same in the SPBN and the SUS-test as in the PB50 test. The Swedish HINT test is a Swedish-language implementation of the English HINT test (Nilsson et al., 1994). It also uses a verbal response procedure but presents whole sentences taken from 25 phonemically balanced lists each consisting of 10 natural everyday sentences. The HINT sentences are presented against a speech-weighted stationary noise towards which the SNR is altered adaptively between sentence presentations in order to determine the lowest SNR at which correct sentence repetition occurs (Hällgren et al., 2006). The Hagerman material consists of 10 phonetically balanced sentence lists each consisting of 10 sentences. All 10 lists each contain the same words but in different random combinations. Such word shuffling is possible, since the grammatical structure of every sentence is
identical. Also the Hagerman sentences are presented against a speech-weighted noise, which is slightly amplitude modulated to sound less unnatural. In the Hagerman’s sentence test, the aim is to identify the speech-recognition threshold defined as the SNR where 50 % correct sentence repetition occurs. This is done by running several 10-item lists at different fixed SNRs until a result below and above the 50 % limit has been found, from which the approximate SNR representing 50 % speech recognition in noise can be determined by linear interpolation (Hagerman, 1984).

**Limitations of current speech-audiometry tests**

Hearing-rehabilitation interventions sometimes result in large-scale speech-recognition improvements, such as when people with moderate to severe hearing loss are supplied with hearing aids for the first time.

Often however, the expected difference in benefit between different options or conditions is quite small. Examples of such situations could be comparisons of the benefit of two different hearing-aid models, evaluation of an open hearing-aid fitting on someone with only a mild-to-moderate high-frequency loss, or when evaluating results from other hearing-rehabilitation interventions such as auditory-training regimes (Cf. Henshaw & Ferguson, 2013).

The tests described above may all be used by Swedish audiologists to evaluate the effectiveness of interventions in the rehabilitation of people with hearing loss. There are however a number of limitations regarding test-score variability, the presence of floor and ceiling effects, learning effects, the choice of background noise and control over lexical properties that limit their efficiency for use as hearing-rehabilitation evaluation tools, all which will be described below.

**Test-score variability**

Consider the following hypothetical example. Pose that a given hearing-aid signal-processing algorithm on average provides hearing-aid users with a benefit equal to a six percentage points improvement of the true score on a PB50 test, from 70 % to 76 %. The existence of an effect of that size could easily be proven (or disproven) in a research study including measurements on many people. Even though the increased benefit of six percentage points may seem small, it could be largely important from a hearing-rehabilitation perspective, lending real help in everyday situations. However, when trying to evaluate such an effect using any of the tests described above, the first
type of limitation that arises stems from the presence of variability in the test scores.

When comparing the scores from different speech-recognition tests, it is of utmost importance to know the expected degree of variability present in those scores. If the degree of variability is not known, a difference between two scores may be mistakenly interpreted as proof of improved speech recognition, while in reality it is only a result of the normal variability of the test scores. Such incorrect conclusions are generally referred to as statistical type I errors, namely to incorrectly reject a true null hypothesis (Howell, 1996). In comparing speech-recognition scores, the null hypothesis would always be the claim that there is no true underlying difference in the speech-recognition ability of the subject. In the end, falsely rejecting such a hypothesis would mean that the particular hearing-rehabilitation patient may be left with a malfunctioning hearing aid. Similarly, large amounts of test score variability are bound to overshadow any true but smaller benefit that is actually present.

There are many potential sources of variability in speech-recognition scores. Between consecutive test occasions, factors such as levels of motivation, fatigue, stress, and familiarity to the speech material may differ (Dillon, 1982; Egan, 1948). To a large degree, such factors are attributable to the subject and may be controlled. Inherent to all speech-recognition scores are also a certain degree of statistical fluctuation stemming from the binary construction of speech-recognition tests, by which each test stimuli can only be either correct or incorrect. The amount of statistical fluctuations in a speech-recognition test sets the limits of how precisely the underlying speech-recognition ability can be measured. This in turn determines how large a difference between two subsequent tests needs to be in order to be considered statistically significant. The largest non-significant score difference is often referred to as the critical difference, above which the test results may be thought to represent a true difference in speech-recognition ability.

Historically, the amount of statistical fluctuations present in speech-audiometry test scores has been estimated from various statistical tests based on binomial distributions (Carney & Schlauch, 2007; Oleson, 2010; Thornton & Raffin, 1978). The reason that binomial distributions can be used to draw inferences about the variability of speech-audiometry scores is that underlyingly, all speech-audiometry scores are derived from outcomes of binary trials, which are either correct or incorrect (Hagerman, 1976). Thornton and Raffin (1978), Carney and Schlauch (2007) and Oleson
(2010) have all used this property of speech-audiometry scores in order to develop tabled criteria as to how much scores from two consecutive speech-audiometry tests must differ in order for the difference to be considered statistically significant. The size of the critical-difference limits depends on both the number of trials in each test, as well as upon the proportion of correct trials in each test. For the hypothetical example given above in which the score and number of trials in the first test was 70 % and 50 respectively, the criteria in Thornton and Raffin (1978) require that the score of another 50-item test needs to exceed 86 % (or alternatively fall below 52 %) in order to be considered significantly different, using a confidence level of 95 %. This situation makes it next to impossible to test the significance of minor improvements in speech recognition using traditional speech-audiometry tests.

As noted already by Hagerman (1976), the basis for using binomial distributions for significance testing of speech-audiometry score differences is not entirely justified. The reason for this is that binomial distributions assume equal success probability of all trials. In other words, the significance testing comes with the assumption that all test words (or sentences) are equally difficult. That this is not the case is obvious. Instead, the difficulty level will differ between different test words, depending on their specific loudness and spectral content, as well as the pure-tone audiogram configuration of each specific listener. Even though the effect of ignoring the assumption of equal difficulty between test trials is conservative in nature (Dillon, 1982) (i.e. decreasing the number of type I errors), the amount to which this affects the validity of the confidence levels used, will differ between different speech-audiometry tests, as well as between different listeners in rather unpredictable ways. Thus, the actual confidence level used (i.e. the observed proportion of statistical type I errors, often referred to as coverage probability) will start to deviate from the nominal confidence level selected (i.e. the expected proportion of statistical type I errors). Depending on the properties of the specific statistical test used, such deviations may be very large for specific combinations of list length and proportion correct trials (Fagerland et al., 2015). Unless the coverage probability is known for all list lengths used and test-score levels attained, the outcome of any statistical significance test used may become next to meaningless. Again, the risk of missing real improvements, or mistaking random fluctuations for real improvements, may have a direct effect upon the outcome of the hearing rehabilitation for the affected individual.
The development of the SiP-test

Floor and ceiling effects on the phonemic level

As noted above, speech-audiometry testing is often performed in the presence of some type of background noise that mask the presented speech material. The degree of masking created by any given background noise, however, varies much between different speech sounds (Meyer et al., 2013; Miller & Nicely, 1955; Phatak & Allen, 2007; Woods et al., 2015). As different speech sounds vary in intensity and frequency content (Ladefoged & Maddieson, 1996), the use of one and the same masker for all speech sounds will necessarily cause louder speech sounds to be clearly audible above the noise, while others become completely inaudible. Thus, at a given SNR, many speech sounds would be far from the auditory perception threshold, either completely inaudible, or fully audible. From a phonemic perspective, inaudible speech sounds thus create floor effects in the speech test, never being actually heard, and on the contrary, fully audible sounds inflict ceiling effects on the test results, never actually missed. The result of such a situation is that only a portion of the speech sounds is effectively tested. Which speech sounds that will belong to this portion will largely depend on the spectrum of the noise in combination with the chosen SNR.

The combined facts that both hearing loss and the acoustic content of human speech sounds tend to be largely frequency specific, means that the degrading effect of hearing loss upon speech-sound audibility will differ between individuals. Thus, for instance, high-frequency hearing loss will mainly decrease the audibility of high-frequency dominated speech sounds, such as, in Swedish, the voiceless fricatives [s], [ʂ] and [ɕ]. For many years, it has been standard procedure to supply hearing-aid gain only for frequency regions affected by the hearing loss. For high-frequency hearing losses, hearing-aid gain is supplied to the corresponding high frequencies. It may seem obvious then that when evaluating such an intervention, what needs to be measured is the perception of the specific speech sounds affected (Cf. Kuk et al., 2010; Woods et al., 2015).

With the phonemically balanced test lists, the possibility to investigate the perception of specific speech sounds is severely limited. That portions of speech materials, such as word lists, are phonemically balanced means that each speech sound occurs at approximately the same rate in each speech-material portion as in the language as a whole. Although it is possible to score the PB50 lists on a phonemic level, their phonemic balance strongly favors a few common speech sounds such as, in Swedish, the vowels [a] and [ɛ], whereas some other sounds, such as [ʂ] and [ʈ] are hardly ever tested (Cf. Häkkinen et al., 2006). Therefore, the most efficient way to deal with
the wide limits of the critical-difference intervals for speech-audiometry test scores described above is to make sure that measurements are only made where improvements can be expected. Thus, investigating the perception of a portion of the speech sounds which happens to be located in the proximity of the speech-in-noise threshold, of which only a few may have actually become more audible by the specific intervention, is inefficient. This was realized already by Egan (1948, p. 960) who referred to such stimuli as “dead weight” to the test. If instead testing could be targeted towards the perception of the specific phonemes affected by the intervention, statistically significant score differences would be considerably easier to attain.

Learning effects
Using fixed lists such as in the PB50-, SUS-, SPBN- and HINT-tests, the number of consecutive administrations to the same subject is clearly limited. The primary reason is that as the same list has been administered once, some learning of which words occur in the list as well as their relative order and position in the list is likely to occur with repeated administrations. Such sequence-learning effects are detrimental to the validity of speech-recognition tests. As subjects eventually learn the test materials by heart, no testing of actual hearing takes place. In effect, the number of times such tests can be administered is limited. Beyond that, hearing-rehabilitation interventions cannot be properly evaluated with fixed-order speech-audiometry lists.

Even though small learning effects have been seen even when the order of presentation is randomized (Foster & Haggard, 1987; Keidser, 1991b, but see also Dubno & Dirks, 1982 who saw no learning effects whatsoever in eight consecutive sessions), the amount of learning effects in such tests seems to level out after a small number of sessions (three or four) (Kollmeier et al., 2015; Walker et al., 1982).

The choice of contextual setting
The typical situation in which speech recognition becomes especially difficult for people with hearing loss is in the presence of background noise (Kochkin, 2010). Therefore, speech audiometry is often performed against a background of disturbing noise. In order for the results from such speech-in-noise tests to be relevant, their background noises need to, at least approximately, sound like sounds that the hearing-rehabilitation patient encounters in his or her everyday life. The question as to what type of background sounds to use in speech audiometry is a matter of which
The development of the SiP-test

The construct validity of a test concerns the extent to which the test can be used to draw conclusions about the construct one aims to measure (Shadish et al., 2002). In the case of speech audiometry, a basic construct of speech recognition in synthetic stationary or fluctuating noise may not be very helpful, as such a construct can hardly be said to reflect the everyday lives of hearing-rehabilitation patients. Instead, the everyday auditory landscape in which people with hearing loss (as well as everyone else in modern society) reside is filled up with all sorts of different sounds, all depending on the local contextual setting. Studies analyzing everyday sound environments of modern western society have shown that common contexts in which conversations often take place are in the presence of background traffic noise, restaurant noise, sounds from public transportations, speech or music and household noise (Smeds et al., 2015; Wagener et al., 2008). Ultimately, therefore, in order to attain high levels of construct validity in speech-audiometry tests, their background noises need to reflect such everyday situations, lest the construct measured risk being totally irrelevant for real life situations encountered outside the testing booth.

Influences from lexical properties

Many lexical factors influence human speech perception. When the perception of single isolated words is of concern, the most important of these factors is the word frequency, indicating how common a word is in the language. The word-frequency effect is facilitative in nature, rendering common words easier to perceive than less common words (Brysbaert et al., 2015). A common metric of word frequency is occurrences per million words. A measure that is better related to the psycholinguistic effect of word frequency is the Zipf-scale, which is a logarithmic scale ranging from approximately one for very uncommon words to about seven for very common words (Brysbaert et al., 2015). Ever since the dawn of speech audiometry, the word-frequency effect has been considered when selecting test words, mainly by excluding infrequent words (Egan, 1948; Lidén & Fant, 1954).

Besides word frequency, there is a range of other lexical metrics that influence word recognition. The phonological and orthographic neighborhood density quantifies the concentration of words that are similar in either spelling or pronunciation. Phonological neighborhood density has been seen to be inhibitory to word perception, creating competition between words, while orthographic neighborhood density seems to be facilitative in
nature (Ziegler et al., 2003). Phonotactic probability quantifies the likelihood by which different phonemes occur in different places of words as well as the likelihood of different phoneme combinations (Storkel, 2004; Vitevitch & Luce, 1998, 1999, 2004) and seems to have a facilitative effect upon word recognition. Orthographic transparency, describing the relation between the phonemes of spoken language and the corresponding graphemes of written language, also seem to be positively correlated to word recognition, even in the auditory domain (Dich, 2014). In addition, there are other factors that influence word recognition such as, contextual diversity/document count (Adelman et al., 2006), number of senses/meanings (Yap et al., 2011), number of letters, phonemes and syllables (Brysbaert et al., 2015), case and status as acronym or abbreviation (Cf. Brysbaert et al., 2009; Martin & Davis, 2019) and even to which word class a word belongs (Coene et al., 2016).

Only the most basic of these other lexical factors, such as word frequency, syllable count, and in Hagerman’s sentences, word-class assignment, have been systematically controlled in the creation of the current Swedish speech-audiometry tests.
A new approach in Swedish speech audiometry

In the light of the discussion above, the efficiency, validity and reliability of speech-audiometry tests could be improved by fulfilling the following criteria. Speech-audiometry tests should:

- Measure speech recognition on a phonemic level, in a way that allows for targeting of specific types of speech sounds
- Ensure that as few speech sounds as possible are subject to floor or ceiling effects by homogenizing the difficulty level between different speech sounds
- Control for relevant lexical factors with psycholinguistic effects upon word recognition
- Allow for extensively repeated use by being constructed in a way that minimizes learning effects
- Use real everyday sounds as background maskers
- Specify methods for significance-testing of test-score differences that have known and stable coverage probabilities

The Situated Phoneme test

The current thesis describes the development of a new speech-audiometry test for the Swedish language that aims to fulfill the criteria laid out above. The new test is henceforth referred to as the Situated Phoneme (SiP) test. The target population for which the SiP-test was developed is adult native speakers of Swedish.

The basic design

In order to test the ability to discriminate between different phonemes, and at the same time enable minimization of learning effects, the SiP-test uses a multiple-alternative forced choice test design. Within the field of speech-perception testing, this type of design was first used by House et al. (1965). In the SiP-test, each test trial consists of an auditory presentation of a real Swedish word, followed by a visual presentation of a number of written response alternatives, which always differ from each other by exactly one speech sound, and of which one is always the correct answer. The task of the participant is to select from the presented response alternatives, which word he or she heard, and guess if uncertain.

When the pronunciation of two or more words with different meanings differ from each other by only a single speech sound, they express minimal
phonemic contrast, and are often referred to as minimal pairs, triplets or quadruples, etc. In the SiP-test, all members of each set of response alternatives are minimal pairs to each other, in the following text, therefore, the groups of response alternatives are sometimes referred to as minimal-variation groups. More often, however, the term test-word group is used instead with the same meaning.

As each test-word group can be regarded as an independent sub-test, SiP-test sessions can be specifically crafted to target only specific phonemic contrasts by including only the test-word groups containing the targeted test phonemes, as well as repeating these any number of times within the same test session. In order to make the SiP-test as flexible and powerful as possible, the test is administered using computer software.

In order for the SiP-test to fulfil the criteria for improved speech-audiometry tests laid out above, some areas required special attention, and are therefore discussed in some detail below.

**Swedish word metrics**

As long as phonemic content is the only property that differs between presented response alternatives in a multiple-alternative forced choice phonemic discrimination test, the only possible way to score better than chance is by auditory phonemic discrimination. Thus, the construct validity in such an experiment is hinged upon the degree as to which phonemic content is the only property that varies between the contrasting response alternatives. As described above, there are many lexical factors other than phonemic content that may differ between words, and which influence the speed and accuracy by which items in the mental lexicon are retrieved, an ability often referred to as lexical access. Even though the influence of such factors may be smaller in closed-set test, studies have indicated that they are not removed (Foster & Haggard, 1987). To illustrate this effect, picture a test situation with a very unfavorable SNR making the test word almost entirely inaudible. Pose then that three response alternatives are present, of which one is very common, and the remaining two are rather infrequently occurring in the language. In this situation, the word-frequency effect will likely create a strong bias to choose the frequent word. Although the psycholinguistic effect of the other word metrics described above may not be as clear, they will likely impose similar biases upon the process of selecting response alternatives. The obvious way of counteracting such biases is to make sure that the most important word metrics are approximately the same between contrasted test words within each test-
word group. The problem in the case of developing the SiP-test was that, with the exception of word frequency data and some orthographic similarity data, there were no resources of this kind available for the Swedish language. Therefore, before selecting test words for the SiP-test, such a word-metric database for the Swedish language would have to be developed.

Selection and creation of linguistic stimuli

Prior to selecting test words for a closed-set forced choice phonemic discrimination test as the SiP-test, two questions naturally present themselves. Firstly, which are the speech sounds for which it would be interesting to measure the ability of discrimination, and therefore should be included as test phonemes? And secondly, to which other speech sounds should each test phoneme be contrasted?

In order for the content validity of the SiP-test to be as high as possible, the response to the first question is that the SiP-test should include as many of the Swedish speech sounds as possible. In response to the second question, one could perhaps argue, in line with the answer to the first question, that as many contrasts as possible should be included. The possible number of phonemic contrasts in the Swedish language (as in most other languages) is however very large. In order to keep the number of response alternatives down and the testing time short, the rationale behind the selection of phoneme contrasts for the SiP-test was to contrast only those Swedish speech sounds that are perceptually most similar. As increasing sound similarity likely also increases the risk of phoneme confusion and subsequent misunderstandings, this strategy focuses testing to the phonemic contrasts that cause most problems in speech understanding.

What is not obvious, however, is the size of the perceptual distance between different Swedish speech sounds. As we knew of no such data for the Swedish language, we devised a measure of phonetic distance based on acoustical analysis of speech-sound recordings, and calculated this measure for all speech sounds occurring in Swedish monosyllabic words.

Homogenization of speech-sound difficulty levels

In order to minimize the phonemic level floor and ceiling effects in the SiP-test, some method that would homogenize the difficulty levels of different phonemes was required. Separately altering the level of different phonemes would not be a way forward, as by itself the relative levels of different speech sounds may function as a perceptual cue for phoneme identification.
The development of the SiP-test (Samuel, 2010). Nor would it be possible to alter the level of individual test words, as then participants could learn to utilize that level deviation, rather than their phonemic discrimination ability, in order to select the correct response (Hagerman, 2009). Instead, the option attempted in order to achieve speech-sound difficulty-level homogenization in the SiP-test is to apply similar amounts of masking to different phonemes by presenting different types of masker sounds to different test-word groups, at a SNR that describes the sound level relation between the contrasting test phonemes and the background sounds. We refer to this method as the phoneme-to-noise ratio (PNR) procedure. The PNR will be described further in the methods section below.

In order for the SiP-test to reflect real-life contextual settings, its test words are presented against a background of naturally occurring everyday sound events. In this way, the SiP-test aims to create a test situation in which the sound environment is similar to what people may encounter in their everyday lives. In the current thesis, the specific context selected is an urban outdoor environment. Thus, the masker sounds consist of passing busses, trams, sounds from construction sites as well as other sounds that may occur in an urban setting.

In order to make the PNR-procedure as effective as possible in homogenizing the difficulty levels between different test phonemes, background sounds were specifically matched in frequency content to the contrasting phonemes in each test-word group. The background sounds were selected from a large database of recordings of urban outdoor sound events. In order for such a masker-sound selection procedure to be practicable, as well as easily applied also to other sound environments in the future, the selection method was implemented in a fully automatized algorithm.

**Significance-testing methods**

As noted above, the degree of statistical fluctuations in speech-recognition measurements change depending on the test length as well as the test score. The fluctuations increase with shorter tests as well as for scores close to 50 percent, and decrease in the opposite directions. The traditional methods of establishing critical intervals mentioned above account for a large range of variations in score but only for a limited set of equal test lengths (10, 25, 50 and 100 test words). As the SiP-test is constructed to allow for a large variation of test-lengths, as well as test-length combinations in which the length of the first test is not necessarily the same as the length of the second
test, the traditional critical-difference tables are not sufficient. Instead, for the SiP-test, significance-test methods are needed that are accurate for almost any test length, as well as test-length combinations.

As also noted above, the statistical methods commonly used for significance testing of speech audiometry score differences are based on binomial distributions and thereby assume equal difficulty levels for all test items presented in a test session. As this assumption cannot be considered fulfilled, the validity of such significance tests could be questioned. By using Poisson’s binomial distributions (Wang, 1993), which account for specific trial-success probabilities, the amount of statistical fluctuation present in speech-recognition scores can possibly be predicted with more accuracy (González et al., 2016). Of course, such a procedure requires that the success probabilities of specific test trials can be estimated.

Thus, fulfilling the criterion regarding the specification of appropriate statistical significance-testing methods for the SiP-test meant both challenges and potential for improvements. The challenges consisted in the identification of methods that could deal with the range of test configurations that occur as a result of the SiP-test flexibility as well as in contesting their validity. The potential for improvement stemmed from the possibility of enhanced coverage probabilities resulting from applying a new type of distributional assumption on speech-audiometry scores.

**Estimation of success probability of specific SiP-test trials**

As noted above, the specification of the Poisson’s binomial distribution requires estimations of the specific success probabilities, alternatively expressed as difficulty levels, of the observed test trials. In a speech-audiometry test, the difficulty levels of specific test trials are bound to depend on factors such as the speech and noise levels used, as well as the hearing-threshold levels of the subject. Such factors can be used to quantify the audibility of the test words, such as in the calculation of the Speech Intelligibility Index (SII) (ANSI-S3.5, 1997). As discussed above, and also seen in other studies (Winkler et al., 2020), also lexical factors of the test words will be important.

The availability of a model capable of predicting SiP-test trial-success probability would have another clinically very important benefit. Namely, that as the success probability of each specific test word, at any combination of speech and noise levels, can be estimated, the total expected score of any SiP-test session, no matter which test words are included, can be predicted simply by summing the predicted success probabilities for the included trials
(Hagerman, 1976). As this can be done prior to actual testing means that it is possible to determine, in advance, which speech and noise levels are appropriate for a given set of test words. Or alternatively, which set of test words that would be appropriate to test for a given combination of subject hearing-threshold levels and speech and noise levels. In addition, as test scores for a baseline test session can be predicted, it would similarly also be possible to approximate the critical-difference intervals towards a second test, whereby as a consequence the number of test trials needed in order to likely detect a significant score difference may be optimized. Thus, creating a reasonably reliable prediction model of SiP-test trial-success probability could be greatly beneficial.

In the next sections, the aims of the thesis, the specific methods used to fulfill those aims, a summary of the results, and finally a brief discussion focusing on the validity and reliability of the SiP-test will be presented. The bulk of the material is based upon studies I-IV. In addition, some further analyses and data not included in any of the four studies will be presented here.

**Thesis aims**

The overall aim of the current thesis was to develop a valid, efficient and well-controlled Swedish-language phoneme-level speech-audiometry test in natural background noise for people with hearing loss. The new test is referred to as the Situated Phoneme (SiP) test. Broadly defined, the aims of the different parts of the thesis were:

A. To create a database containing psycholinguistic word metrics relevant for human word recognition in the Swedish language, which could be used to select appropriate linguistic test stimuli for the SiP-test.

B. To create linguistic test stimuli for the SiP-test.

C. To develop speech-spectrum shaped real-word masker sounds appropriate for the linguistic stimuli in the SiP-test.

D. To investigate the quality of statistical significance-test methods that can be used to perform significance testing of differences between observed SiP-test scores.

E. To develop a prediction model for the Swedish SiP-test.

F. To investigate the validity, reliability and learning effects of the SiP-test as described in the current thesis.
G. To investigate the degree of homogenization of the difficulty levels of different phonemes attained by the SiP-test PNR-procedure as described in the current thesis.

**Specific aims**

The first of these aims (A) was addressed in study I. Specifically, the aims of study I were, a) to develop algorithms that calculate different word metrics relevant for the construction of new Swedish speech-perception tests, and b) to create a database in which those word metrics were calculated for a large set of Swedish words.

Aims B and partly F were addressed in study II. Specifically, the aims of study II were a) to develop a set of minimally contrasting, phonetically similar and psycholinguistically well-controlled test-word groups for the SiP-test, and b) to create sound recordings of the selected SiP-test words and to evaluate their accuracy experimentally with normal-hearing people in a silent background.

Aim C was addressed in study III, which had the specific aim to identify and modestly adapt sound recordings of auditory events that were both appropriate in a common real-world auditory setting and spectrally matched to the test-phoneme contrasts of the SiP-test.

Aims D and E were addressed in study IV. More specifically the aim of study IV was to investigate the quality of different methods that can be used to perform significance testing of subsequent speech-recognition measurements at different score levels as well as test-length combinations. A secondary aim of study IV was to develop a computational model for the prediction of trial-success probabilities in the SiP-test.

Finally, extended analyses of aims F and G were added solely in the current chapter.
Methods

An overview of the relation between the different parts of the current thesis is presented in the form of a flowchart in figure 1.

Figure 1. Flowchart relating the different parts in the development of the SiP-test in the current thesis.
The word-metric database developed in study I was named the AFC-list, as it was developed at the Audiological Research Centre in Örebro, Sweden. The construction of the AFC-list required mainly three types of data, namely, word spellings, phonetic transcriptions, and word frequency data. Additionally, in order for the extracted metrics to be representable of the Swedish language as a whole, the amount of data needed to be rather large.

Fortunately, most of these types of data were available in some form for the Swedish language. Word spellings and phonetic transcriptions were drawn primarily from a freely available Swedish lexical database hosted by the national library of Norway (Nasjonalbiblioteket, 2011) referred to as the NST database. In order to suit the needs of the AFC-list, a substantial revision, primarily of the NST phonetic transcriptions, but also of many incorrect spellings, had to be undertaken. In addition, a slight revision of the original NST phoneme inventory was made, as well as a conversion of its phonetic transcriptions to the IPA standard (International Phonetic Association [IPA], 1999). Finally, also a relatively large set of monosyllabic and disyllabic Swedish words were added to the database. Word frequency data were drawn from a large collection of Swedish internet blogs, available at the University of Gothenburg (Borin et al., 2012).

Based on these data, many types of word metrics were extracted. These were categorized as frequency metrics, word length metrics, semantic metrics, neighborhood metrics, phonotactic metrics, and metrics of orthographic transparency. Several metrics, such as the Zipf-scale word frequency metric (van Heuven et al., 2014), the positional segment probability and the position-specific bi-phone probability metrics of phonotactic probability (Storkel, 2004; Vitevitch & Luce, 2004), and the grapheme-to-phoneme metric of orthographic transparency (Berndt et al., 1987) were direct re- Implementations of metrics originally developed for other languages. Several new metrics were also developed in study I, such as the Zipf-scale-weighted phonetic neighborhood density probability (PNDP), the normalized stress- and syllable-based phonotactic probability (SSPP) and the grapheme-initial letter-to-pronunciation orthographic transparency (GIL2P-OT).

In addition to these rather complex metrics, a range of simpler metrics were included, such as the number of letters, graphemes, phonemes, syllables, homographs, homophones, word-classes and senses, the phonotactic type (e.g. CVC for consonant-vowel-consonant words), the existence of special orthographic characters in the spelling, as well as...
categorizations as abbreviation, acronym, or foreign words. All metrics were calculated using computer algorithms implemented in the programming language Microsoft Visual Basic .NET.

Study II

In order to determine, for each test phoneme in the SiP-test, which were the most similar Swedish speech sounds, and thus to which other phonemes it should be contrasted, we devised a method to compute a metric quantifying the phonetic distance between different speech sounds based on audio recordings of those speech sounds. The computation took as input the phonemically contrasting phones in a large set of audio recordings of Swedish minimal pairs. For each comparison, a set of Bark spectra was calculated for each phoneme. These Bark spectra were then run through a dynamic time-warping algorithm to determine the lowest possible temporo-spectral difference, given an allowed amount of time warping. The attained difference value was finally also time-weighted according to the temporal difference between the compared speech sounds. In order to arrive at a useful scale of phonetic distance between Swedish speech sounds, the resulting phonetic distance data were averaged across all comparisons involving the same phoneme pair. Finally, to create sets of appropriate candidate test phones for the SiP-test, we used the phonetic distance data to form clusters of phonetically similar phones. This clustering resulted in 31 sets of phonemically contrasting candidate test phones.

In order to select appropriate test words that could embody those phoneme contrasts, we searched among the monosyllabic words of the AFC-list to identify groups of words expressing the different phoneme contrasts by means of minimal phonemic variation. Among the many groups identified, a manual selection for inclusion in the SiP-test was made of groups within which the variation in four AFC-list word metrics was as low as reasonably possible. These metrics were the Zipf-scale word frequency value, the PNDP, the word-average SSPP and the word-average GIL2P-OT.

All selected test words were then recorded to digital audio files. For each test word, we made five different recordings by two different speakers, one male and one female. In order to trigger a Lombard effect (Cf. Van Summers et al., 1988) in the recorded speech material, the speakers wore headphones in which a speech-weighted noise was presented during the recording at approximately 65 dB SPL. Prior to the recording of each test word, the speakers were presented with a prototype recording of the same test word.
(without the speech-weighted noise). The speakers were asked to try to pronounce each test word in a similar manner as the prototype recording, the purpose being to ensure approximately equal vocal effort and fundamental frequency in all test-word recordings. In order to ensure that the prototype recordings did not differ much, they were also created in a similar manner, using the recording of one single word as prototype.

In order to validate the perceptibility of the test-word recordings, they were presented in a listening experiment to 28 normal-hearing adult speakers of Swedish. The average presentation level was 62.35 dB SPL. The task of the participants was to listen for a monosyllabic word uttered either by a male or by a female voice, and then to indicate on a touch screen which word he or she had perceived. The response alternatives were presented after the completion of the auditory test-word presentation, and always consisted in the three or four members of the test-word group to which the presented word belonged. Each participant was presented with one recording by each speaker of each member word in all test-word groups. All test words were presented in random order. If, for any test-word recording, an incorrect or missing response occurred, two additional trials presenting that test-word recording were randomly inserted among the remaining test trials.

**Study III**

To achieve the aims of study III, a computer algorithm was developed that was able to parse a large library of sound files consisting of sounds that were considered common in an urban environment (Gloaguen et al., 2017).

For each test-word group and speaker in the SiP-test, the algorithm calculated a Bark-weighted average spectrum of the test-phoneme sections of the SiP-test audio recordings. The algorithm then searched the whole soundscape library for the best matching sound sections, with an appropriate duration and sound-level stability. Having identified the best audio sections for each test-word group and speaker, the algorithm ensured that as rich a variability as possible of different types of masker sounds was distributed between the test-word groups. As a last step, the algorithm utilized custom spectral shaping filters to fine tune the spectral content of the masker sounds, for a better fit against the corresponding contrasting test phones.
Study IV

In study IV, six different significance-test methods were compared. The methods were composed from the combination of three different techniques to calculate confidence intervals, as well as two underlying distributional assumptions. The three techniques were, a) so called, exact calculation, b) uncorrected normal approximation and c) Agresti-Caffo corrected normal approximation (Cf. Agresti & Caffo, 2000; Fagerland et al., 2015). The two different distributional assumptions were the Poisson’s binomial (Wang, 1993) and the binomial distributions.

All methods had been described elsewhere, except the Agresti-Caffo corrected technique based on the Poisson’s binomial distribution, which was formulated in study IV.

The quality of the significance-methods was determined in terms of coverage probability, which describes the actual confidence level attained when using a nominal confidence level of 95 %. The closer the coverage probability was to 0.95 across different test-length combinations and underlying true scores the better the significance-testing method was considered to be.

Coverage-probability values for the different significance-test methods were first calculated for three hypothetical test types that differed in the distribution of success probabilities within each test session. Secondly, the corresponding coverage-probability values were calculated for a set of simulated repeated SiP-test scores for which the underlying trial-success probabilities had been predicted using a multi-level logistic regression model referred to as the SiP-test model. SiP-test scores were estimated for different typical pure-tone audiogram configurations. The SiP-test model was based on an adaptation and extension of the Speech Intelligibility Index (SII) (ANSI-S3.5, 1997) called the Phoneme Discriminability Level (PDL), also developed in study IV, as well as on the duration of the presented test phoneme, test-phoneme type (i.e. vowel or consonant), word frequency, neighborhood density and phonotactic probability of each test words. The word metrics used to represent the last three lexical properties were the Zipf-scale value, the inverse of the PNDP, and the average SSPP taken from the AFC-list. The training data for the SiP-test model were derived from real SiP-test sessions with 74 adult native speakers of Swedish. All participants had normal hearing or symmetric sensorineural hearing loss. Included hearing-threshold levels ranged from normal hearing to severe hearing loss, as defined by the World Health Organization (2020b).
With a few exceptions, each participant took three full-length (84 test words) SiP-test sessions at three different PNRs. The PNRs used differed between participants, and were set so that approximate scores of 60 %, 80 % and 90 % correct would be attained. In a fourth SiP-test session, taken by 72 participants, one of the three initially used PNR values was repeated. The choice of SiP-test voice (male or female), as well as the order of the three different PNR values, was altered between participants.

The sound level ratio referred to as the PNR related the maximum level of the contrasting phonemes (using an integration time of 50 ms) to the sound pressure level of the section of the masker sounds aligned with the test-word presentation.

In each SiP-test session, the test words were presented from a speaker right in front of the listener. At azimuths of -30 and +30 degrees from the front speaker, two different masker sounds from study III were simultaneously presented in separate loudspeakers. Between each test-word
presentation, uncorrelated urban background sounds were presented through the masker speakers at a total level of 60 dB SPL in the listeners’ position, and in the front speaker, a section of the International Speech Test Signal (ISTS) (Holube et al., 2010) appeared at the same average level as the presented test words. Synchronized with the auditory test-word presentation, a yellow circle appeared on a touch screen placed in front of the listener. Five hundred milliseconds after the end of each test-word presentation, three written response alternatives appeared on the touch screen. The task of the listener was to indicate, by tapping on the touch screen, which word he or she had heard, and guess if uncertain. The test setup is illustrated in figure 2.

Finally, coverage-probability values based on the human-participant test-retest session data (i.e. the paired baseline- and repeated test sessions using the same PNR) were compared to the coverage probabilities derived from a large number of consecutive computer simulations of the same human-participant test-retest sessions.

**Test-retest reliability and learning effects**

As noted above, there are many factors besides statistical fluctuations that can potentially influence the stability of speech-recognition tests. In order to investigate the overall test-retest reliability and learning effects of the SiP-test, data from the 72 participants in study IV who took the full (84 trials) SiP-test sessions twice at the same PNR was analyzed. The PNRs used in those test-retest sessions ranged from − 5 dB to + 15 dB and their distribution is presented in figure 3.

In order to determine to what extent the scores from the baseline tests deviated from the scores of the corresponding repeated tests, each score difference was tested using both the Poisson’s binomial Agresti-Caffo corrected method of normal approximation (PBAC) developed in study IV and the binomial Agresti-Caffo corrected method of normal approximation (BAC). In both cases, a confidence level of 95 % was used. In order to determine the amount of test-retest variability present in the test scores that likely stemmed from other sources of variation than random variation, the proportions of test-retest sessions with a score difference that fell outside the boundaries of the estimated critical score differences were calculated. In the case that there should be notable learning effects, effects from fatigue, or from other sources of variation other than random fluctuations between the baseline tests and the retest sessions, more than five percent of the test-
retest score differences should fall outside their corresponding critical intervals.

Even though the presence of learning effects should be apparent from the above analysis, minute learning effects may not be detected due to lack of statistical power. In order to detect smaller learning effects, the grand total test-retest difference was calculated across all participants by comparing the total number of correct trials throughout all 72 baseline tests with the corresponding number of correct trials in all repeated test sessions. In addition, the distribution of test-retest differences was calculated, both as uncorrected percentage scores and as rationalized arcsine units (RAU) (Studebaker, 1985).

![Figure 3. The distribution of PNRs used for the 72 participants described in study IV that took both the baseline and the corresponding repeated test sessions.](image-url)
Comparisons with established audiological measures

In order to compare the outcome of the SiP-test with other outcome measures commonly used in Swedish hearing rehabilitation, such data was collected for the participants of study IV.

Therefore, at the same occasions when the data on the SiP-test was collected in study IV, the SUS-test (Grunditz & Magnusson, 2013) and the first part of the Speech, Spatial and Qualities of Hearing Scale (SSQ) (Gatehouse & Noble, 2004) were administered.

The SUS-test, briefly described in the introduction, was specifically developed for evaluation of hearing-aid benefit in hearing rehabilitation (Grunditz & Magnusson, 2013). In the current study, the SUS-test was presented at 65 dB SPL, and the original SNR of + 6 dB was used. As the SUS-test has been seen to be too difficult to people with severe hearing loss (Grunditz & Magnusson, 2013) two study IV participants with better-ear pure-tone threshold averages for the frequencies 0.5, 1, 2 and 4 kHz (BPTA4) at or above 70 dB HL were excluded from the SUS-test with noise. An additional five participants, all with BPTA4 values at or above 54 dB HL, were also excluded from the SUS-test with noise, as it was too difficult for them to partake in. Finally, one additional subject was excluded due to a high level of familiarity to the particular test lists. Thus, in total, 65 of the 74 participants described in study IV took the SUS-test with background noise.

The SSQ is a questionnaire that measures perceived disability related to speech perception and exists in a Swedish translation by M. Öberg, G. Wänström and S. Arlinger (The English original was developed by Gatehouse & Noble, 2004). In addition to the SiP-test and the SUS-test, we also administered the first part of the SSQ (SSQ1), which focuses on hearing speech in a variety of competing sounds. All 74 participants described in study IV answered the questions in the SSQ1.

As the SiP-test data was not collected at the same difficulty level for all participants in study IV, direct comparisons between the SiP-test scores and those from the SUS-test and the SSQ1 were not possible. Instead, the SiP-test model developed in study IV was used to predict a SiP-test score for each participant at a PNR of zero decibels and a reference level of 68.34 dB SPL (the latter is the same as used with the participants in study IV). Instead of predicting the SiP-test success probabilities on the population level, as can be done using the population-level intercept of the SiP-test model developed in study IV, the specific model intercept for each study-IV participant was used, as this would largely improve the SiP-test score
predictions. Having predicted the success probability of each of the 84 SiP-test trials in one test session per participant in study IV, as expected at a PNR of zero, trial outcomes were then sampled from the Bernoulli distribution using the function `rbern` in the R-package `Rlab` (Boos & Nychka, 2012). Predicted session scores for each participant were then calculated as the number of correct trials. The original choice of SiP-test voice (i.e. male or female), which was balanced between the participants of study IV, was retained in the SiP-score predictions here.

In order to compare the predicted SiP-scores to the outcomes of the SUS-test and the SSQ1, Pearson’s correlation was used. In addition to analyzing the whole SSQ speech-hearing subscale, the SiP-test scores were also correlated to each specific SSQ1 subscale item.

As the quality of any conclusions about the SiP-test criterion validity drawn from such estimated SiP-test scores would be heavily hinged upon the quality of the SiP-test model predictions themselves, the latter needed to be investigated. This was done by comparing the degree of deviation from the average success probabilities estimated for each of the 291 SiP-test sessions in study IV to a) the corresponding observed average session scores as well as b) the corresponding average scores expected when the deviations were solely resulting from statistical fluctuations.

Therefore, the underlying trial-success probability of each single test trial in all 291 test-sessions with real human participants in study IV were first estimated by the SiP-test model from study IV, using participant-specific model intercepts. Then the average trial-success probability was calculated for each of the 291 sessions. Then the differences between the average trial-success probability of each test session and a) the observed proportion of correct trials in each session, and b) simulated session scores consisting of the proportion of correct responses for each test session as sampled from independent Bernoulli distributions based on the estimated trial-success probabilities within each session.

The distribution of differences between the average estimated success probabilities for each of the 291 test sessions and their corresponding average scores were then calculated for both the observed and for the simulated scores separately.

**Homogenization of test-phoneme difficulty levels**

As has already been discussed, the susceptibility to masking differs largely between different speech sounds, which may result in large floor and/or ceiling effects when test stimuli are scored on a phonemic level. By applying
similar amounts of masking to different phoneme classes, the PNR-procedure presented in study IV had the purpose of homogenizing the difficulty levels between different test phonemes, and thereby reducing the presence of such floor and ceiling effects. In order to investigate whether the PNR-procedure had the anticipated effect on the masking susceptibility of different phoneme classes, empirical cumulative score distributions for different test-word groups and test-phoneme classes were analyzed based on the observational SiP-test data from the 71 participants in study IV that took all four SiP-test sessions. Analyses were both based on the data from all 71 participants irrespective of hearing ability, and then separately within groups determined by the degree of hearing loss, as defined according to the World Health Organization (2020b).

**Ethical considerations**

The listening experiments in study II and IV were conducted in accordance with the declaration of Helsinki. In no case were any participants exposed to any harmful sound levels, thus none of the experiments posed any risk for the participants. In both experiments, participation was voluntary, written informed consent was acquired from all participants, and their personal information was treated confidentially.

The reference numbers to the decisions of the regional ethical review board in Uppsala concerning the relevant parts of the current thesis were *Dnr 2015/477* and *Dnr 2018/475*. 


Results

Study I
All results of study I were gathered in a Microsoft Access database file, as well as in a pure UTF-8 formatted text file. The material was named the AFC-list, and contains all available word-metric data extracted for a total of 816,404 phonetically transcribed Swedish words. The AFC-list has been made available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

Study II
In table 1, spellings, phonetic transcription, and contrasting phonemes of all test words within the 28 test-word groups selected for the SIP-test are presented. The test-word groups do not fully embody all 31 sets of candidate test phones, identified for the Swedish language, as it was not possible to find appropriate real-word minimal-variation groups expressing all desired contrasts. Thus, four contrast sets were removed, three sets slightly adjusted. In addition, two sets of high-frequency dominated voiceless fricatives were reduplicated into two test-word groups each.

The variation in the Zipf-scale value, the PNDP, the word-average SSPP and the word-average GIL2P-OT within the selected test-word groups was relatively small. On a relative scale, the remaining variability was largest for the word frequency (Zipf-scale value) and phonological neighborhood density (PNDP).

The sound recordings of the SiP-test words have been made available online at https://osf.io/y4nqb under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license (https://creativecommons.org/licenses/by-nc/4.0/).
Table 1. Spellings, phonetic transcriptions and contrasting phones in the selected minimal-variation groups. The categories ‘Long’ and ‘Short’ correspond to the phonetic length of the contrasting phones.

<table>
<thead>
<tr>
<th>Spelling</th>
<th>Phonetic transcription</th>
<th>Contrasting phones</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vowels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sitt, sytt, sött</td>
<td>[sitː], [svːt], [søːt]</td>
<td>[i], [y], [ø]</td>
</tr>
<tr>
<td>sätt, sitt, sytt</td>
<td>[sɛtː], [sitː], [svːt]</td>
<td>[ɛ], [i], [y]</td>
</tr>
<tr>
<td>satt, sätt, sött</td>
<td>[satː], [sɛtː], [søːt]</td>
<td>[a], [ɛ], [ø]</td>
</tr>
<tr>
<td>mark, märk, mörk</td>
<td>[mark], [mærk], [mœrk]</td>
<td>[a], [æ], [œ]</td>
</tr>
<tr>
<td>bland, blönd, blund</td>
<td>[blænd], [blœnd], [blʊnd]</td>
<td>[a], [o], [œ]</td>
</tr>
<tr>
<td>sarg, sorg, sörj</td>
<td>[sarːɡ], [sɔrːɡ], [sœrːɡ]</td>
<td>[a], [o], [œ]</td>
</tr>
<tr>
<td>rätt, rott, rött</td>
<td>[rɔtː], [rʊtː], [røːtː]</td>
<td>[ɔ], [ʊ], [ø]</td>
</tr>
<tr>
<td><strong>Long</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pir, pur, pyr</td>
<td>[piːr], [pʊr], [pyːr]</td>
<td>[i], [u], [y]</td>
</tr>
<tr>
<td>red, råd, Ryd</td>
<td>[reːd], [rɛːd], [ryːd]</td>
<td>[e], [ɛ], [y]</td>
</tr>
<tr>
<td>Klas, kläs, klös</td>
<td>[klɑːs], [klɛːs], [kløːs]</td>
<td>[o], [ɛ], [œ]</td>
</tr>
<tr>
<td>mas, mäs, mös</td>
<td>[mɑːs], [moːs], [møːs]</td>
<td>[o], [o], [œ]</td>
</tr>
<tr>
<td>mår, mor, mur</td>
<td>[mɔːr], [mʊr], [mœr]</td>
<td>[o], [u], [œ]</td>
</tr>
<tr>
<td><strong>Consonants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hy, hyf, hys, hyrs</td>
<td>[hyː∅], [hyːf], [hyːs], [hys]</td>
<td>[∅], [f], [s], [ʃ]</td>
</tr>
<tr>
<td>arm, armarm, larm</td>
<td>[∅ɑːrm], [fɑːrm], [lɑːrm]</td>
<td>[∅], [f], [l]</td>
</tr>
<tr>
<td>yr, fyrr, skyr, syr</td>
<td>[∅yːr], [fɪːr], [ʃyːr]</td>
<td>[∅], [f], [ʃ], [s]</td>
</tr>
<tr>
<td>å, tå, sjå, så</td>
<td>[∅ɑː∅], [fɔː∅], [ʃɔː∅]</td>
<td>[∅], [f], [ʃ], [s]</td>
</tr>
<tr>
<td>all, hall, pall, tall</td>
<td>[∅ɑlː], [hɑːl], [pɑːl], [tɔːl]</td>
<td>[∅], [h], [p], [t]</td>
</tr>
<tr>
<td>il, kil, fil, sil</td>
<td>[∅iːl], [ɕiːl], [fiːl], [siːl]</td>
<td>[∅], [ɕ], [f], [s]</td>
</tr>
<tr>
<td>ur, bur, dur, mur</td>
<td>[∅ʊrː], [bʊrː], [dʊrː], [mʊrː]</td>
<td>[∅], [b], [d], [m]</td>
</tr>
<tr>
<td>ko, kon, korn</td>
<td>[kuː∅], [kuːn], [kuːɳ]</td>
<td>[∅], [n], [ɳ]</td>
</tr>
<tr>
<td>ed, led, ned</td>
<td>[∅ɛd], [lɛd], [nɛd]</td>
<td>[∅], [l], [n]</td>
</tr>
<tr>
<td>kval, kvarn, kvar</td>
<td>[kvɑːl], [kvɑːn], [kvɑːɾ]</td>
<td>[l], [n], [ɾ]</td>
</tr>
<tr>
<td>kval, kvarn, kvav</td>
<td>[kvɑːl], [kvɑːn], [kvɑːv]</td>
<td>[l], [n], [v]</td>
</tr>
<tr>
<td><strong>Long</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tuff, tuss, tusch</td>
<td>[tufː], [tʊʃː], [tœːʃː]</td>
<td>[f], [s], [ʃ]</td>
</tr>
<tr>
<td>sopp, satt, sort</td>
<td>[sɒpː], [sɔtː], [sɔːt]</td>
<td>[p], [t], [t]</td>
</tr>
<tr>
<td>sock, sätt, sort</td>
<td>[sɔkː], [sɔtː], [sɔtː]</td>
<td>[k], [t], [t]</td>
</tr>
<tr>
<td>tugg, tum, tung</td>
<td>[tʊŋː], [tʊmː], [tœŋː]</td>
<td>[ɡ], [m], [ŋ]</td>
</tr>
<tr>
<td>paj, pall, pang</td>
<td>[paɟː], [pɑlː], [paŋː]</td>
<td>[j], [l], [ŋ]</td>
</tr>
</tbody>
</table>

Note: The character ∅ (zero-phoneme) corresponds to the absence of a phoneme.

When analyzing the results of the listening experiment evaluating the accuracy of the test-word recordings, we ignored single isolated errors, as such could be the results of random mistakes due to loss of concentration etc. However, repeated errors within the same test session occurred for the word “pyr” (both voices) in the group “pir, pur, pyr”, for the word “sytt”
(for the female voice) in the groups “sätt, sitt, sytt” and “sitt, sytt, sött”, and for the word “sjå” (female voice) in the group “å, få, sjå, så”. In addition, a few repeated errors also occurred for the words “kon” (both voices) and “tall” (the female voice). In general, the female voice stimuli were more often misinterpreted than the male voice stimuli.

**Study III**

In study III, 280 custom masker sounds were created; five masker sounds for each of the 28 different test-word groups and the two different SiP-test voices. All masker sounds (along with slightly modified versions of the test-word recordings from study II) have been made available online at https://osf.io/q4rb3 under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license (https://creativecommons.org/licenses/by-nc/4.0/).

Study III also presented comparisons of the spectra of all maskers to their corresponding test phonemes. The study concluded that the frequency matching was generally good, even though there would be room for improvement in a few test-word groups. Such improvement, however, would likely require that the soundscape library used would be extended with more and different types of sounds.

**Study IV**

The results indicated that by using tests based on the Poisson’s binomial distribution instead of the binomial distribution, the coverage-probability values are considerably more accurate, at least when the inequalities in difficulty levels within test sessions are very large. When the coverage probabilities were calculated from the simulated SiP-test sessions, the corresponding differences between the assumed distributions were much less pronounced, ranging from just above zero to two percentage points for different audiogram configurations, test-length combinations and, difficulty levels.

For the exact and uncorrected approximate methods, the coverage probability tended to be highly unreliable for test-length combinations where at least one test was below approximately 25 trials. The Agresti-Caffo corrected methods, however, stayed relatively accurate for test lengths as short as three trials.

The comparisons between the coverage probabilities calculated from actual and simulated SiP-test sessions indicated relatively good agreement.
Test-retest reliability and learning effects

Of the 72 test-retest difference scores included in the analysis of the test-retest reliability of the full SiP-test, two difference scores (i.e. 2.8 %) were large enough to be considered statistically significant using the PBAC method developed in study IV. Figure 4 presents the scores from the baseline tests plotted against the scores from the repeated test sessions, with the two statistically significant test-retest scores marked with empty circles. Using a confidence level of 95 %, approximately 5 % of the test sessions, i.e. three or four session pairs, would be expected to fall outside the bounds of the critical difference. As this only happened in two cases, the variation between the baseline and repeated SiP-test sessions did not exceed that which could be expected solely from random fluctuations.

Figure 4 also shows the results of the BAC significance-testing method, here presented in the classical form indicating the bounds of the critical-difference interval (Cf. Thornton & Raffin, 1978). The BAC intervals are shown by the dotted lines, and their interpretation is that data points that fall at or between the dotted lines represent score differences that are not statistically significant. As indicated by figure 4, using the BAC method, none of the test-retest score differences turned out to be statistically significant. The results thus indicate that the size of the test-retest score differences do not exceed those expected from purely random fluctuations. This is in full agreement with the results of study IV, where the BAC method was seen to be more conservative as it does not account for the inequalities of test-trial difficulty levels.

The general conclusion, regardless of which significance-test method is chosen, is that the test-retest reliability of the full-length SiP-test should be considered very high.

In total, the baseline and the retest sessions contained 6048 trials each (84 trials × 72 participants). The grand total number of correct trials counted across all baseline and retest sessions were 4623 (76.4 %) and 4659 (77.0 %), respectively. Overall, thus, there was an improvement of 0.6 percentage points between the baseline and the repeated test sessions. The distribution of test-retest differences is presented in figure 5 as uncorrected percentage points and as RAUs. Both the uncorrected and RAU-transformed distributions indicate that the test-retest score differences were approximately normally distributed, with standard deviations of 5.5 percentage points and 6.1 RAUs, respectively. Due to the size of these standard deviations, the minute improvement seen from the baseline sessions to the repeated sessions (0.6 percentage points and 0.5 RAUs
respectively) are clearly not statistically significant. Thus, the amount of learning effects present between consecutive administrations of the SiP-test seems to be insignificantly small, at least when analyzing across all test-word groups.

Figure 4. Scatter plot of baseline, against repeated, SiP-test scores. Open and filled circles indicate significant and non-significant score differences, respectively, as tested by the PBAC method developed in study IV. Dotted lines indicate the critical-difference interval as calculated using the BAC method described in study IV. For both methods, a confidence level of 95 % was used. As more than 95 % of all points show non-significant score differences, the test-retest variability can be completely explained by the expected random fluctuations.
Figure 5. Histograms presenting the test-retest differences for the SiP-test deriving from 72 participants in study IV. Uncorrected percentage points (i.e. the percent correct on the baseline test subtracted from the percent correct on the repeated test) are presented to the left, and the corresponding differences in rationalized arcsine units (RAU) to the right.

Comparisons with established audiological measures
Figure 6 presents frequency polygons indicating the distribution of differences between the average estimated success probabilities for each of the 291 test sessions in study IV and a) their corresponding average observed scores (solid red line) as well as b) corresponding simulated session scores sampled from independent Bernoulli trials (dashed blue line). The frequency polygons in figure 6 both use a bin-width of three percentage points. The vertical solid black line intercepts the abscissa at zero percentage points, equivalent to no difference. As is seen in figure 6, there are no large differences between the distributions of observed and simulated scores. In addition, both distributions are centered at zero percentage points. This indicates that the SiP-test model, to a large extent, succeeded in estimating the underlying success probabilities of the SiP-test trials with the human participants in study IV. In turn, this indicates that the SiP-test model can also be used to make valid inferences about expected SiP-test scores at specific values of PNR, especially in the vicinity of the PNR values sampled (Cf. figure 3). Thus, the following results from the analyses of SiP-test
criterion validity, based on SiP-test scores estimated for a PNR of zero dB should be considered reliable.

Figure 6. Frequency polygons presenting the distribution of differences between average estimated success probabilities for each of the 291 test sessions in study IV and their corresponding observed average scores (solid red line) as well as corresponding simulated session scores sampled from independent Bernoulli trials (dashed blue line).
Figure 7 plots the observed scores for the SUS-test with background noise against the SiP-test scores estimated for a PNR of zero dB. Pearson’s correlation between the SUS-test scores and the SiP-test scores was 0.71 ($r(63) = 0.71$, $p < 0.001$, 95 % CI [0.56, 0.81]). Likewise, figure 8 plots the observed scores for the SSQ1 against the estimated SiP-test scores at a PNR of zero dB. Pearson’s correlation between the SiP-test scores and the scores from SSQ1 was 0.64 ($r(72) = 0.64$, $p < 0.001$, 95 % CI [0.48, 0.75]).

Figure 7. SUS-test scores for 65 participants from study IV plotted against SiP-test scores estimated for a PNR of 0 dB. Colors indicate better-ear pure-tone threshold averages for the frequencies 0.5, 1, 2 and 4 kHz (BPTA4).
Figure 8. SSQ1 scores for all 74 participants of study IV plotted against SiP-test scores estimated for a PNR of 0 dB. Colors indicate better-ear pure-tone threshold averages for the frequencies 0.5, 1, 2 and 4 kHz (BPTA4).

Finally, table 2 presents the correlations between the estimated SiP-test scores for a PNR of 0 dB and each of the specific items in the SSQ speech-hearing subscale. The items in table 2 are presented in decreasing order of correlation with the SiP-test scores. The item-specific correlation coefficients range from $R = 0.66$ (CI = 0.50, 0.77) for the item showing the highest correlation with the SiP-test scores (“Having conversation with five people around a table in quiet with vision”) to $R = 0.32$ (CI = 0.10, 0.51) for the item with the lowest correlation to the SiP-test scores (“Follow one person speaking and telephone at same time”).
Table 2. Correlations between estimated SiP-test scores (for a PNR of 0 dB) and items on the SSQ speech-hearing subscale for the 74 participants in study IV.

<table>
<thead>
<tr>
<th>SSQ item (minimal back-translation)</th>
<th>English item №</th>
<th>Presented order</th>
<th>Pearson’s R (CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Having conversation with five people around a table in quiet with vision</td>
<td>8</td>
<td>3</td>
<td>0.66 (0.5, 0.77)</td>
</tr>
<tr>
<td>Talking with one person in quiet room with floormats</td>
<td>14</td>
<td>2</td>
<td>0.63 (0.47, 0.75)</td>
</tr>
<tr>
<td>Having conversation with five people in noisy restaurant with vision</td>
<td>4</td>
<td>4</td>
<td>0.62 (0.46, 0.74)</td>
</tr>
<tr>
<td>Having conversation in echoic environment</td>
<td>6</td>
<td>7</td>
<td>0.62 (0.45, 0.74)</td>
</tr>
<tr>
<td>Follow conversations without missing start of new talker</td>
<td>2</td>
<td>12</td>
<td>0.59 (0.42, 0.72)</td>
</tr>
<tr>
<td>Having conversation with five people in noisy restaurant—no vision</td>
<td>2</td>
<td>6</td>
<td>0.58 (0.41, 0.71)</td>
</tr>
<tr>
<td>Have conversation on telephone</td>
<td>13</td>
<td>13</td>
<td>0.56 (0.38, 0.7)</td>
</tr>
<tr>
<td>Follow one conversation when many people talking</td>
<td>7</td>
<td>11</td>
<td>0.55 (0.37, 0.69)</td>
</tr>
<tr>
<td>Talking with one person in continuous background noise, such as fan or running water</td>
<td>9</td>
<td>5</td>
<td>0.54 (0.36, 0.69)</td>
</tr>
<tr>
<td>Ignore interfering voice of different pitch</td>
<td>12</td>
<td>9</td>
<td>0.51 (0.32, 0.66)</td>
</tr>
<tr>
<td>Talking with one person with TV on</td>
<td>10</td>
<td>1</td>
<td>0.48 (0.28, 0.64)</td>
</tr>
<tr>
<td>Ignore interfering voice of same pitch</td>
<td>11</td>
<td>8</td>
<td>0.44 (0.23, 0.61)</td>
</tr>
<tr>
<td>Talk with one person and follow TV news</td>
<td>3</td>
<td>10</td>
<td>0.44 (0.24, 0.61)</td>
</tr>
<tr>
<td>Follow one person speaking and telephone at same time</td>
<td>1</td>
<td>14</td>
<td>0.32 (0.1, 0.51)</td>
</tr>
</tbody>
</table>

Note: English item № indicates the original item order in (Gatehouse & Noble, 2004), and Presented order indicates the order of items in the Swedish translation. SSQ items are presented in decreasing order of correlation with the SiP-test scores.
Homogenization of test-phoneme difficulty levels

The cumulative score distributions presented in this section (figures 9-12) were all based on data from 284 SiP-test sessions by the 71 participants in study IV which took all four SiP-test sessions and should all be interpreted similarly. Namely, each point in the figures represents the proportion of participants that had an average score – counted from all four SiP-test sessions – that equaled or succeeded the cumulative score indicated by the abscissa. All points that represent the type of stimuli have been connected by lines to aid the interpretability of the figures. Thus, lines that are located further to the left in the figures represent stimuli that were more difficult than those further to the right.

The grand average score in the 284 SiP-test sessions was 79.2 % for the vowels test phonemes and 75.4 % for the consonant test phonemes. Thus, looking across all participants and SiP-test difficulty levels, the vowel test phonemes were somewhat easier than the consonant test phonemes. The average score difference was 3.8 percentage points. Looking at figure 9, in which cumulative score distributions are presented separately for the vowel and consonant test phonemes, it is clear that this difference was consistent throughout all difficulty levels.

In figure 10, the same type of cumulative score distributions for the SiP-test vowel and consonant test phonemes are presented separately for different groups of hearing-threshold level, as defined in accordance to World Health Organization (2020b). Among the normal-hearing participants, the difficulty levels are completely homogeneous when the average scores for consonants and vowels are compared. However, as the grade of hearing loss increases, so does the difficulty level of the consonant test phonemes as compared to the vowel test phonemes.

In order to study the variation of difficulty levels within the groups of consonant and vowel test phonemes, figure 11 presents the cumulative score distributions separately for the different test-word groups in the SiP-test. Cumulative score distributions for test-word groups with consonant test phonemes are presented in the upper pane and the corresponding data for the vowel test phonemes are presented in the lower pane.

Among the vowels, the spread of difficulty levels is relatively homogeneous. Among the consonants, however, there is a large spread of difficulty levels. Looking across all participants at specific test-word groups in the SiP-test, the results thus indicate that the method of homogenization used in study IV did not manage in creating a complete homogenization of all test phonemes in the SiP-test.
Figure 9. Cumulative frequency polygons showing the cumulative score distributions for vowel and consonant test phonemes in the SiP-test. Each point indicates the proportion of participants (as given by the ordinate) who had the specific proportion of correct responses equal to or less than the cumulative score (as given by the abscissa). Connecting lines are added in order to facilitate interpretation.

In order to investigate whether the pattern seen in figure 11 results from collapsing the analysis across all included levels of hearing loss, figure 12 presents the same type of analysis calculated separately for different grades of hearing loss. Even though the resolution is generally lower in figure 12, as each point reflects the outcome of fewer test trials, it is clear that the same general patterns of larger spread in difficulty among the consonants are
present for all grades of hearing loss. In addition, even the order of relative
difficulty between different test-word groups is very similar between
different grades of hearing loss. Thus, given that PNR is set to an
appropriate value for each subject, the same test-word groups seem to be
easy and difficult, respectively, largely independent of the hearing-threshold
levels of the subjects.

Figure 10. Cumulative frequency polygons showing the cumulative score
distributions for vowel and consonant test phonemes in the SiP-test, separately for
different groups of hearing-threshold level. Each point indicates the proportion of
participants (as given by the ordinate) who has the specific proportion of correct
responses equal to or less than the cumulative score (as given by the abscissa).
Connecting lines are added in order to facilitate interpretation.
Figure 11. Cumulative frequency polygons showing the cumulative scores distributions for the different test-word groups in the SiP-test. Consonant groups are presented in the upper pane and vowel groups in the lower pane. Each point indicates the proportion of participants (as given by the ordinate) who had the specific proportion of correct responses equal to or less than the cumulative score (as given by the abscissa). Connecting lines are added in order to facilitate interpretation. Combinations of line types and shapes (as well as colors) indicate test-word group.
Figure 12. Cumulative frequency polygons showing the cumulative scores distributions for the different test-word groups in the SiP-test, separately for different groups of hearing-threshold level. Consonant groups are presented in the upper set of panes and vowel groups in the lower set. Each point indicates the proportion of participants (as given by the ordinate) who had the specific proportion of correct responses equal to or less than the cumulative score (as given by the abscissa). Connecting lines are added in order to facilitate interpretation. Combinations of line types and shapes (as well as colors) indicate test-word group.
Discussion

Validity

Construct validity
Shadish et al. (2002, p. 38) define construct validity as the “validity of inferences about the higher order constructs that represent sampling particulars”. As a tool for evaluating the benefits that individuals with hearing loss have from various interventions in hearing rehabilitation, the objective of the SiP-test is primarily to quantify the quality of the speech-perception process in various realistic situations. Thus, the higher order construct targeted here concerns the two aspects *speech perception* and *realistic situations*.

Speech perception can be broadly defined as the process of deriving a meaning from an acoustic signal. Over the years, there has been much debate as to precisely how this process unfolds in the human mind and nervous system (For an overview, see Diehl et al., 2004; Miller & Eimas, 1995; Samuel, 2010). Basically, however, the process of speech perception involves the transmission of acoustic energy through the peripheral auditory system, neural encoding and transmission through various nuclei of the central auditory system ending up in cortical activity. Throughout this process, temporal and spectral information is utilized to segment representations of phonemic units that link to words stored in the mental lexicon, which are in turn integrated into the current discourse (Miller & Eimas, 1995). The information flow is most likely both bottom-up, i.e. driven by the sensory stimulus, and top-down, allowing lower-level processing to be influenced by higher-level contextual and lexical processing.

The purpose of the SiP-test is not primarily to evaluate the subject’s ability to use contextual information in order to derive meaningful information, but instead quantify changes in the ability to interpret specific speech sounds, with as little influence from contextual clues as possible. Obviously, this process does not cover all aspects of speech perception. Its strength, however, rests in its ability to selectively target the specific problems in speech perception that a sensory hearing loss cause. Namely, the lack of ability to detect and/or discriminate between the different cues of human speech.
Though there has been much debate as to what exactly constitutes those basic constituents of human speech – whether they are phonological features, spectra, gestures, phonemes, syllables, etc. – the existence of the phoneme as a perceptual unit actively involved in the speech-perception process receives strong support, at least from a linguistic point of view (Kazanina et al., 2018).

As discussed previously, a high level of construct validity of the SiP-test requires that what is tested is the ability to differentiate between different phonemes. Phonemes, on the other hand, typically do not occur in isolation. In spoken verbal communication, phonemes always occur as parts of words. This fact raises two rather important issues. Firstly, if in normal verbal conversation phonemes always occur in words that are stored in the mental lexicon, high construct validity of a speech-perception test should spark this process, lest it is questionable whether the test procedure really involves the construct of speech perception. In comparison to other speech-audiometry tests employing non-words or pseudo-word (for an overview of such tests, see Rødvik et al., 2018), this factor was the single most important reason why we chose to use real Swedish words in the SiP-test.

Secondly, on the other hand, real words are notoriously maladapted for use as test stimuli of speech-perception processes. The reason for this is that words come with meanings, connotations, play different syntactic roles, are hugely dissimilar in frequency of occurrence and levels of competition within the mental lexicon. They also come with spellings, for which the level of transparency differs hugely between different words. In addition, some words are formed using very smooth speech-sound combinations, while others use phonotactic structures that are very atypical for their specific languages (Brysbaert et al., 2015; Coene et al., 2016; Dich, 2014; Storkel, 2004; Vitevitch & Luce, 1998; Yap et al., 2011; Ziegler et al., 2003). Thus, when utilizing real words in speech-perception tests targeting phonemic discrimination, many factors other than the contrasting phonemes bias the speech-perception process, and therefore also the construct validity of the particular test.

As described above, we aimed to reduce the influences of such confounding factors in the SiP-test by minimizing the variation in word frequency, phonological neighborhood density, phonotactic probability and orthographic transparency between the contrasting response alternatives in each test-word group. As the available number of minimal-variation groups in a given language is limited, where Swedish is no exception, it will never be possible to remove such differences completely. In addition to the word-
metric variables held under some control in the current study, there are also other important variables that were allowed to vary freely, such as the word class (See Coene et al., 2016 for a study of word-class influences in Dutch speech audiometry.). Therefore, even though there is perhaps a more rigid control over confounding lexical and sub-lexical influences upon the phonemic discrimination task in the SiP-test, there are still reasons to continually question and re-evaluate its level of construct validity.

As noted above, a non-negotiable prerequisite for high construct validity of the SiP-test is that normal-hearing individuals should be able to discriminate between the various test words in silent background. As, in fact, this proved difficult with some phoneme contrasts in study II, these phoneme contrasts should probably not be included in clinical use of the SiP-test. It is, however, rather intriguing that normal-hearing individuals repeatedly confused the phoneme /y/ for /i/, with both the male and the female voice. Several alternative explanations were discussed in the text of study II, such as the impact of a possible spectral tilt resulting from the Lombard-type recording situation and the influence of phonotactic probability. A further possibility could be that this may be initial indications that the /y/-/i/ phonemic distinction may have begun a process of dissolving in certain context of the Swedish language. If such a process would continue among large groups of prospective hearing-rehabilitation patients, the /y/-/i/ contrasts should likely be removed from the SiP-test material on the same grounds as were other dialectally neutralized contrasts already during the test-word selection process in study II.

When attempting to quantify speech-perception ability, properties of the surrounding auditory environment are no trivial matters. The main reason that these aspects are important is that the types of situations in which people with hearing loss typically experience most speech-perception difficulties are those with background noise (Kochkin, 2010). Traditionally, speech-in-noise audiometric testing in the Swedish language has been performed in stationary, or slightly amplitude modulated, speech-weighted random noise (Hagerman, 1984; Hälgren et al., 2006; Magnusson, 1995). Arguably, performing measurements in speech-weighted noise is often more relevant than running the same tests in a silent background, as the addition of noise also has the effect of reducing ceiling effects. From a test-validity perspective, however, the use of random speech-weighted noises may not be the best choice, as these types of noises are rarely encountered in the everyday situations of people with hearing loss. Instead, the types of situations encountered are typically everyday sound environments, namely
those of offices, construction-sites, kindergartens, households, restaurants, and even natural settings such as beaches or forests. As long as tests of speech perception do not incorporate the soundscapes of such everyday environments, the inferences drawn about the subjects’ speech-perception abilities in everyday situations are bound to be rather fragile. On the other hand, using sound from such everyday situations as speech-audiometry maskers may reduce test reliability as test words may sometimes be ‘glimpsed’ through short gaps in the noise (Cf. Cooke, 2006), and sometimes not. In the SiP-test, this problem was avoided by presenting only relatively stable masker sounds. Doing so, likely increased the test reliability, but had the cost of reducing the construct validity, as obviously also highly fluctuating background noises exist in most everyday situations.

In the version of the SiP-test developed in the current thesis, we aimed to increase the construct validity by defining one specific everyday environment in which the testing is assumed to take place, namely an urban outdoor environment. Consequently, the results of this version of the SiP-test should be at least partly generalizable to such a context. This choice should however not be interpreted as if we deemed the urban outdoor environment as particularly important. On the contrary, this is merely one of the many typical situations that people with hearing loss encounter in their daily lives. In order to improve the construct validity of the SiP-test, more such background environments should ultimately be developed in the future. Optimally, the audiologist should then be able to evaluate the benefit of rehabilitation interventions in the contexts most appropriate for each individual patient.

Even though it may sound daunting to develop such new sound environments for the SiP-test, the task is strongly facilitated by the automatic algorithm developed in study III. In fact, in order to develop a new SiP-test environment, the only thing needed is a database of sound recordings from that particular environment. The algorithm described in study III can then automatically create all sound files needed to run the SiP-test with the new environment. Due to the frequency matching inherent in the algorithm, masker sound with spectra similar to those of the contrasting test phones in each test-word group would be selected also for new environments. In turn, this means that maskers for new environments will be spectrally similar to the masker sounds from the urban outdoor environment developed in the current thesis. In turn, this will mean that the SiP-test prediction model developed in study IV will likely work rather well.
also with masker sounds from new environments. Developing such new SiP-test environments is thus a task that could be pursued in the near future.

**Content validity**
The content validity of a test refers to the extent to which its items adequately cover the construct under investigation (Streiner & Norman, 2008). In study II, we selected speech-sound contrasts to use as test phonemes in the SiP-test. In doing so, we grouped the Swedish speech sounds into clusters of similar speech sounds, as determined by a temporospectrally-based computational model of phonetic distance. For each speech-sound contrast, we created word groups containing the minimal phonemic variation needed in order to test the ability to discriminate between the contrasting test phonemes. As this process made sure that most Swedish phonemes were included in the material, the content validity of the SiP-test, with respect to the Swedish phoneme inventory, should be considered to be relatively good.

However, as the included test phonemes are only contrasted to a limited set of phonetically similar phones, the content validity, in terms of phoneme contrasts is somewhat limited. However, an inclusion of more phoneme contrasts would add unwanted complexity to the SiP-test, as it would require the participant to scan a larger number of response alternatives (Cf. Foster & Haggard, 1987) for each trial. In addition, it would decrease the possibility of finding appropriate sets of real words that do not differ much in the psycholinguistic word metrics used for the assembling of appropriate test-word groups. Thus, developing a phonemic perception test such as the SiP-test involves striking the right balance between factors such as to what extent specific particulars of the measured construct should be represented, what can be considered feasible task demands, as well as what level of control over confounding lexical factors that should be exercised.

**Criterion validity**
When developing a new test, it can be enlightening to compare the results of the new test to those of other existing tests, whereby the criterion validity of the new test can be investigated. High correlations in such comparisons indicate that the tests likely measure the same underlying constructs, often referred to as concurrent validity (Shadish et al., 2002). If such correlations are very high, results from the new test may not add any value, in terms of new information, not already present in the previous test results. In such situations, the new test may be superfluous. On the other hand, if a well-
validated “gold standard” exists, high correlations may indicate that the quality of the new test is very high.

In the current thesis, scores from the SiP-test were compared to scores from two other measures of speech perception used in audiological rehabilitation, the SUS-test and the SSQ speech-hearing subscale. As was presented in the results section above, both of those measures correlated positively with the results of the SiP-test. Studying those correlations, however, it should be remembered that what was compared was only the most general composition of SiP-test trials, including all available test phonemes. This is not the way that the SiP-test is intended to be used clinically. Instead, the clinically most appropriate way to use the SiP-test is to utilize the SiP-test prediction model from study IV, in order to predict exactly which test phones that are likely to be located in the proximity of the speech-discrimination threshold at the specific PNR selected for testing. As the construction of the SiP-test allows for testing of only those specific phonemes, it will likely be superior to both the SUS-test and the SSQ in detecting clinically relevant changes in speech-perception ability.

In contrast to the SiP-test score estimations presented in figure 7, relatively large ceiling effects appear to be present among the corresponding SUS-test scores, as many such observations approach 100 % correct. The main reason for these ceiling effects are probably that the fixed SNR (+6 dB) which was used in the SUS-test sessions was too easy for many participants, while in the SiP-test sessions, ceiling effects could be avoided as the presented PNRs were adjusted separately for each participant.

In contrast to the SiP-test, which in the current thesis is situated in a specific urban outdoor environment, the SSQ speech-hearing subscale asks about the subjects’ experiences of hearing speech in a variation of other, sometimes rather vaguely defined, contexts (Cf. table 2). As the SSQ spans across different contexts than the SiP-test, the resulting scores are supposedly reflecting a related, but more broadly defined, underlying construct. This generality of the SSQ enables a certain degree of discriminability between the constructs measured by the SiP-test (with the urban outdoor background) and the SSQ. The degree to which two tests measure different constructs is often referred to as discriminant validity. In order to be able to claim that two tests really measure different constructs, it should be possible to show that correlations are not too high, as likely then they are measuring the same construct, whereby their construct validity will be impinged (Shadish et al., 2002). Looking at the item-level correlations presented in table 2, the item with the highest correlation
(R=0.66) to the SiP-test scores is “Having conversation with five people around a table in quiet with vision”. As this correlation only explains about 44% (i.e. R-squared) of the variability in the SiP-score, rather different constructs are likely measured by the two tests.

**Reliability**

As discussed in the introduction, speech-audiometry test scores are most often derived from a collection of discrete binary outcomes, such as correct or incorrect phoneme discrimination. This situation effectively makes the random variability of the test scores completely known, as long as all test items are equally difficult, or the difficulty level of each particular test-item, expressed as the probability of success can be correctly estimated. Threats to the reliability of a speech-audiometry test, however, involve not only the predictable amount of random variability but also other more unpredictable factors such as learning effects, fatigue, lapses of concentration, etc. (Dillon, 1982).

The investigation of test-retest reliability in the current thesis indicated that the amount of random variability expected to be present when comparing the scores from two consecutive SiP-test sessions was very close to the actually observed variability. Thus, as the variability expected solely from random fluctuations explained all observed variability there is no room for any other major sources of variability in the SiP-test data. This is a clear indication that the reliability of the SiP-test is very high, at least when presented in its full-length (i.e. 84 test words) format, within which each SiP-test word is presented only once.

**Learning effects**

The data in the current thesis indicate that there are no observable learning effects between consecutive SiP-test sessions. In comparison to earlier studies in which small, but statistically significant learning effects have been seen when stimuli are presented in random order (Keidser, 1991a, 1991b; Kollmeier et al., 2015), the absence of learning effects in the current study may be related to our use of five different sound recordings for each test word. This may have made it more difficult for the listeners to associate specific traits present in some recordings to specific test words. In addition, we used prototype recordings, which the SiP-test talkers imitated during the recording of the SiP-test words. The intent of these prototype recordings was to minimize prosodic differences between the test-word recordings when uttered by the SiP-test talkers. The absence of such prosodic cues may
have reduced learning effects by further preventing the association of different test words with specific stimuli artefacts.

As noted above, the analysis of learning effects in the SiP-test made in the current thesis was based upon full-length SiP-test sessions, within which each SiP-test word was presented only once. That no learning effects were seen in these data seems promising for the clinical use of the SiP-test as a measure of hearing-rehabilitation benefit. However, as specific test-word groups are selected and repeated many times in clinical use, learning effects may eventually appear. At present, this is still uncertain and will have to be evaluated in future studies with multiple consecutive test administrations.

**Critical differences**

As previously discussed, when evaluating the benefit of various hearing-rehabilitation interventions using speech-audiometry tests, it is of utmost importance to be able to distinguish true score differences from random fluctuations. To pursue this aim in a manner applicable to the SiP-test scores was the main topic of study IV.

The significance-test method that showed the highest quality in study IV across the investigated levels of speech-recognition ability and test-length combinations was the Agresti-Caffo corrected method assuming underlying Poisson’s binomial distributions. The method can be successfully used, along with the SiP-test model, to perform significance testing between consecutive speech-recognition scores from the SiP-test.

As the significance-test methods based on Poisson’s binomial distributions in study IV require a prediction model that can estimate specific success probabilities for each test trial, the methods cannot be used with other speech-recognition tests as long as an equivalent model does not exist. However, although the Agresti-Caffo corrected normal approximation to the binomial is somewhat more conservative, it is almost as accurate as its Poisson’s binomial counterpart and may well be used as a substitute when trial-specific success probabilities are not available.

**Homogenization of test-phoneme difficulty levels**

When comparing the cumulative score distributions that describe the difficulty of the consonant test-word groups with those of the vowel groups in figure 9, it appears that the PNR-procedure has been relatively successful in homogenizing the difficulty levels between consonant and vowel test phonemes.
Considering the cumulative score distributions given in figures 11 and 12, there are clear differences in difficulty levels between the different test-word groups, as indicated by the horizontal spread of the lines. The spread is largest among the different consonant groups, but even the vowel groups show some heterogeneity in difficulty levels. In addition, figure 12 indicates that the amount of spread in difficulty levels between test-word groups in the SiP-test are similar between all different levels of hearing included in this study. The remaining heterogeneity in difficulty levels between the different test-word groups in the SiP-test must therefore be attributed to other factors than hearing-threshold levels. Several such factors exist. Firstly, the PNR-procedure relies upon the appropriateness of the phoneme segmentation performed on the sound recordings of the SiP-test words. As different phoneme types differ largely in regards to their temporal enfolding, segmentation strategies will necessarily also differ between different phoneme types. In addition, the amount to which the identity of the test phonemes will be signaled by adjacent phonemes via co-articulation will also differ between test-word groups. Once the test phonemes within the SiP-test word recordings had been manually segmented in the current thesis, an attempt was made in study IV to reduce the effect of differences in temporal enfolding between the test phonemes by defining the contrasting-phonemes level $L_{CP}$, subsequently used to calculate the PNR. The $L_{CP}$ dealt with temporal fluctuations by using temporal weighting with a predefined temporal integration time of 50 milliseconds. Even though this integration time was psycho-acoustically motivated, the PNR-procedure could well be improved by adjusting this constant. However, in order to improve the homogenization of difficulty levels between different phonemes, other factors than audibility possibly need to be considered.

As described above, an attempt was made to reduce the influence of word frequency, phonological neighborhood density, phonotactic probability and orthographic transparency on the choice of response alternative in each SiP-test trial by minimizing the variability of these factors within each test-word group. However, between the test-word groups these factors were allowed to vary quite freely. Consequently, there was quite a large variation in word frequency, phonological neighborhood density, phonotactic probability and orthographic transparency between the different test-word groups. This likely made some test-word groups, and thereby indirectly also their test phonemes, more difficult than others. Certainly, the homogenization of test-phoneme difficulty levels in the SiP-test would have benefitted from keeping all word metrics as constant as possible also between the test-word groups.
However, even though there were about 9 000 monosyllabic words in the AFC-list from which the test words in the SiP-test were selected, the number of suitable, minimally contrasting, word groups that fulfilled those criteria was in fact rather limited. An alternative approach to deal with this problem could have been to, instead of using monosyllabic words, assemble minimally contrasting test-word groups for the SiP-test from the about 102 000 bisyllabic Swedish words that exist in the AFC-list (Witte & Köbler, 2019). Possibly then, all relevant word metrics could have been held relatively constant even between the test-word groups.

The purpose of the PNR-procedure used in the current thesis was to homogenize the susceptibility to masking between different types of phonemes, so that one and the same test setting could be used for all test phonemes. Such a test would have been elegantly simple to administer and relatively free from both floor and ceiling effects. Even though the objectives with the PNR-procedure were not fully obtained, floor and ceiling effects can also likely be avoided if the PNR value is allowed to be altered between different test-word groups during testing. Which PNR value to use for different test-word groups can be estimated by the SiP-test model developed in study IV. Thus, instead of using a single PNR value, it would be possible to base each test setting on a target score, for instance 60 percent correct phoneme discrimination without hearing aids. Consecutive test sessions with and without hearing aids could then be run, presenting the test phonemes of interest. As the PDL incorporated in the SiP-test model takes hearing-aid amplification into account, be it only monaurally on the best side, it is also possible to predict SiP-test scores with hearing-aid amplification. However, the use of the SiP-test model in this way would require further investigations before being put into clinical practice. Though departing from the idea of using a single PNR for all test-word groups in the SiP-test would make administration of the SiP-test more complex, this is something that could easily be handled directly in the SiP-test software.

**Swedish word metrics**

In study I, a new database, the AFC-list, containing a large set of lexical properties often used in psycholinguistic research was developed for the Swedish language. As the database contains more than 800 000 Swedish words, the resource lays the foundation for extensive research into the study of human word perception in the Swedish language.

In addition, the availability of the AFC-list means that the process of creating new well-controlled Swedish speech-audiometry tests is greatly
simplified, as it offers speech- and hearing researchers simple means of looking up word metrics for a very large collection of Swedish words. In case important words are missing from the AFC-list, the supplementary materials to study I enable the researcher to manually calculate several of the included word metrics. Such manual calculations, however, may be rather tedious. Therefore, all algorithms for the word-metric calculations in study I have been integrated into an internet web site (Witte, 2020), by which the word metrics can be easily calculated for any Swedish real word, pseudo-word or non-word.

**Further development and clinical implementation of the SiP-test**

The current study has developed and validated materials and methodological bases for the Swedish SiP-test. However, in order for the SiP-test to be useful to clinical practitioners and hearing researchers it needs to be packaged into a ready-to-use product along with precise documentation and instructions.

As previously described, the SiP-test has been constructed with the purpose of allowing a large degree of customization towards people with different levels of hearing loss. In fact, SiP-test sessions can be constructed using any selection of test-word groups (with their corresponding test phonemes), any test length, any speech level as well as any noise level. Putting this amount of flexibility directly into the hands of clinical audiologists is likely to result in equally large amounts of confusion. A suggestion concerning clinical implementation of the SiP-test, therefore, could be to define a small amount of, three or four, different standard SiP-test presets that can be used with predefined groups of patients in order to evaluate the efficiency of various hearing-rehabilitation measures taken. For instance, a selection of high-frequency dominated consonant phonemes could be used with patients with sloping audiogram configurations and next-to-normal hearing in the low-to-mid frequencies. With more pronounced mid-frequency losses, also voiced consonants and front vowels could be added to the preset. With flat audiogram configurations, as well as with cochlear implant users, all types of test phonemes could be included. Signal and noise levels could optimally be set for each patient to levels where the SiP-test prediction model estimates scores around 60 percent correct – avoiding both floor and ceiling effects. Finally, the number of test items to present would optimally be determined in a power calculation based on the PBAC method developed in study IV, and guided by the size of the test-score difference one desires to be able to detect between different test
conditions. All this functionality should optimally be implemented directly into the SiP-test software, and be directly available to the clinical audiologist.

Even though the SiP-test methodology in the current thesis has focused on stimulus presentation at fixed PNRs along with an interpretation of the results in terms of observed proportions of correct responses, it is likewise possible to use adaptive test-procedures, rendering results in terms of observed PNR values near a given fixed proportion correct responses (Cf. Levitt, 1971). Using adaptive test-procedures with the SiP-test should however require further validation studies before being put into clinical practice.

Describing a limited set of SiP-test presets as outlined above should by no means implicate that an endpoint for the development of the SiP-test should have been reached. Quite on the contrary, with increased experience, growth of available data and the development of further masking-sound environments for the SiP-test, it should be possible to tailor SiP-test sessions even more to the specific needs of individual hearing-rehabilitation patients. Thus, the development of the SiP-test has only just begun.
Conclusions

In the current thesis, foundations for a new valid and reliable Swedish test of phonemic discrimination for people with hearing loss have been developed. The test is referred to as the Situated Phoneme (SiP) test, as it measures phoneme discrimination situated in the auditory background of a specific everyday sound environment. The SiP-test uses real Swedish monosyllabic words, with controlled variation in the lexical properties word frequency, phonological neighborhood density, phonotactic probability and orthographic transparency, based on metrics developed as a part of the thesis. The everyday sound environment used as maskers in the SiP-test represents an urban outdoor soundscape. Even though this is only one of many sound environments in modern society, efficient methods to develop further sound environments for the SiP-test have been devised as a part of the thesis.

The primary intention of the SiP-test is to be used as a tool for evaluation of benefit in speech perception resulting from various hearing-rehabilitation interventions, such as hearing-aid fitting and hearing-aid adjustment. In terms of reliability, the primary advantage of the SiP-test compared to existing Swedish speech-audiometry tests that may be used for the same purpose, is that the SiP-test can be set to target the perception of the specific speech sounds expected to be affected by particular hearing-rehabilitation interventions on an individual basis. This can be done via the use of a prediction model, the SiP-test model, which accounts for both the audiogram configuration of the subject, the speech- and noise levels presented as well as the influence of specific lexical factors of the test words. In comparison to other commonly used speech-audiometry tests, the SiP-test is thus very flexible, allowing for manipulation of both test length and included materials with very little restrictions. A consequence of this flexibility is that established guidelines concerning the lower limits of score differences between consecutive test sessions needed for statistically significant results are not applicable to the SiP-test. In this thesis, therefore, such methods were developed and evaluated specifically for the SiP-test.

In summary, thus, the current thesis has developed everything needed in order to implement a reliable test for evaluation of speech-perception benefit in common everyday situations, adaptable to the needs of specific hearing-rehabilitation patients, and supplied with validated methods to determine the statistical significance of observed score differences.
Acknowledgements

First of all, I would like to thank my Κοινή Greek teacher, Professor Tommy Wasserman, who was the first to convince me that I would be able to handle doctoral-level studies (even though the subject at that time would have been a rather different one).

I would then like to thank my teachers, friends and colleagues at the Summer Institute of Linguistics, SIL. From you guys, I learned to aim high, work hard and not be content with hasty solutions.

I would also like to thank my former main supervisor Professor Emeritus Claes Möller for putting your trust in my ability to handle this PhD project, Professor Emeritus Arne Leijon, for hitting just the right balance between critical and encouraging feedback (and for finding this piece of work interesting enough to read in your spare time!), Björn Hagerman, Henrik Danielsson, Stig Arlinger and Örjan Dahlström for helping to improve this work at the mid- and full-time seminars, and to Henrik and Örjan for introducing me to the world of R at the SIDR summer school. The latter has really been worth the effort!

Of course, a big thankyou also goes to my excellent supervisors Susanne Köbler, Jonas Ekeroot, Karolina Smeds and Elina Mäki-Torkko. You have always provided valuable knowledge, support and joyful company when I’ve needed it and refrained from interfering when I’ve needed time to work on my own.

Thanks also to all the staff at the audiologist program in Örebro, the research community within SIDR, all people at the Audiological Research Centre in Örebro and at ORU Innovation.

Also, thank you Mom and Dad, Karin and Göran, for all your support!

Lastly, to my dear family, Ingrid, Klara and Selma. Thank you! I’m sorry I had to work just this hard! I hope I will be able to make it up to you during many years to come!

Erik Witte
References


ERIK WITTE  *The development of the SiP-test*  69


Studies from the Swedish Institute for Disability Research

1. Varieties of reading disability
   Stefan Gustafson

2. Cognitive functions in drivers with brain injury – anticipation and adaptation
   Anna Lundqvist

3. Cognitive deafness
   Ulf Andersson

4. Att lära sig leva med förvärvad hörselnedsättning sett ur par-perspektiv
   Carin Fredriksson
   ISBN 91-7373-105-6, 2001

5. Signs, Symptoms, and Disability Related to the Musculo-Skeletal System
   Gunnar Lundberg

6. Participation – Ideology and Everyday Life
   Anette Kjellberg

7. Föräldrar med funktionshinder – om barn, föräldraskap och familjeliv
   Marie Gustavsson Holmström
   ISBN 91-7203-500-5, 2002

8. Active wheelchair use in daily life
   Kersti Samuelsson

9. Två kön eller inget alls. Politiska intentioner och vardagslivets realiteter i den arbetslivsinriktade rehabiliteringen
   Marie Jansson
10. **Audiological and cognitive long-term sequelae from closed head injury**  
    Per-Olof Bergemalm  
    ISBN 91-7668-384-2, 2004

11. **Att vara i särklass – om delaktighet och utanförskap i gymnasiesärskolan**  
    Martin Molin  
    ISBN 91-85295-46-9, 2004

12. **Rättvis idrottsundervisning för elever med rörelsehinder – dilemma kring omfördelning och erkännande**  
    Kajsa Jerlinder  
    Licentiate Degree, 2005

    Per-Inge Carlsson  
    ISBN 91-7668-426-1, 2005

14. **Hearing and cognition in speech comprehension. Methods and applications**  
    Mathias Hägggren  
    ISBN 91-85297-93-3, 2005

15. **Living with deteriorating and hereditary disease: experiences over ten years of persons with muscular dystrophy and their next of kin**  
    Katrin Boström  
    ISBN 91-7668-427-x, 2005

16. **Disease and disability in early rheumatoid arthritis**  
    Ingrid Thyberg  
    ISBN 91-85299-16-2, 2005

17. **"Varför får jag icke följa med dit fram?" Medborgarskapet och den offentliga debatten om dövstumma och blinda 1860-1914**  
    Staffan Bengtsson  
    ISBN 91-85457-06-X, 2005

18. **Modalities of Mind. Modality-specific and nonmodality-specific aspects of working memory for sign and speech**  
    Mary Rudner  
    ISBN 91-85457-10-8, 2005
19. Facing the Illusion Piece by Piece. Face recognition for persons with learning disability
Henrik Danielsson
ISBN 91-85497-09-6, 2006

20. Vuxna med förvärvad traumatisk hjärnskada – omställningsprocesser och konsekvenser i vardagslivet. En studie av femton personers upplevelser och erfarenheter av att leva med förvärvad traumatisk hjärnskada
Thomas Strandberg

21. Nycklar till kommunikation. Kommunikation mellan vuxna personer med grav förvärvad hjärnskada och personernas närstående, anhöriga och personal
Pia Käcker

22. ”Aspergern, det är jag”. En intervjustudie om att leva med Asperger syndrom
Gunvor Larsson Abbad

23. Sounds of silence – Phonological awareness and written language in children with and without speech
Janna Ferreira

24. Postponed Plans: Prospective Memory and Intellectual Disability
Anna Levén

25. Consequences of brain tumours from the perspective of the patients and of their next of kin
Tanja Edvardsson

26. Impact on participation and service for persons with deafblindness
Kerstin Möller

27. Approaches to Audiological Rehabilitation with Hearing Aids: studies on prefitting strategies and assessment of outcomes
Marie Öberg
28. Social Interaction and Participation in Activities of Everyday Life Among Persons with Schizophrenia
   Maria Yilmaz
   Licentiate Degree, 2009

29. Focus on Chronic Disease through Different Lenses of Expertise Towards Implementation of Patient-Focused Decision Support Preventing Disability: The example of Early Rheumatoid Arthritis
   Örjan Dahlström

30. Children with Cochlear Implants: Cognition and Reading Ability
   Malin Wass

31. Restricted participation: Unaccompanied children in interpreter-mediated asylum hearings in Sweden
   Olga Keselman

32. Deaf people and labour market in Sweden. Education – Employment – Economy
   Emelie Rydberg

33. Social rättvisa i inkluderande idrottsundervisning för elever med rörelsehinder – en utopi?
   Kajsa Jerlinder
   ISBN: 978-91-7668-726-0, 2010

34. Erfarenheter av rehabiliteringsprocessen mot ett arbetsliv – brukarens och de professionellas perspektiv
   Helene Hillborg

35. Knowing me, knowing you – Mentalization abilities of children who use augmentative and alternative communication
   Annette Sundqvist
36. **Lärare, socialsekreterare och barn som far illa – om sociala representationer och interprofessionell samverkan**  
   Per Germundsson  

37. **Fats in Mind**  
   Effects of Omega-3 Fatty Acids on Cognition and Behaviour in Childhood  
   Ulrika Birberg Thornberg  

38. **”Jobbet är kommunikation”**  
   Om användning av arbetshjälpmedel för personer med hörselnedsättning  
   Sif Bjarnason  

39. **Applying the ICF-CY to identify everyday life situations of children and youth with disabilities**  
   Margareta Adolfsson  

40. **Tinnitus – an acceptance-based approach**  
   Vendela Zetterqvist  

41. **Applicability of the ICF-CY to describe functioning and environment of children with disabilities**  
   Nina Klang  

42. **Bringing more to participation**  
   Participation in school activities of persons with Disability within the framework of the International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY)  
   Gregor Maxwell  

43. **From Eye to Us.**  
   Prerequisites for and levels of participation in mainstream school of persons with Autism Spectrum Conditions  
   Marita Falkmer  
44. Otosclerosis, clinical long-term perspectives
   Ylva Dahlin-Redfors

45. Tinnitus in Context - A Contemporary Contextual Behavioral Approach
   Hugo Hesser

46. Hearing and middle ear status in children and young adults with cleft palate
   Traci Flynn

47. Utrymme för deltagande, beslutsprocesser i möten mellan patienter med ospecifica ländryggsbesvär och sjukgymnaster i primär vård
   Iréne Josephson

48. ”Man vill ju klara sig själv” Studievardagen för studenter med Asperger syndrom i högre studier
   Ann Simmeborn Fleischer

49. Cognitive erosion and its implications in Alzheimer’s disease
   Selina Mårdh

50. Hörselscreening av en population med utvecklingsstörning
    Utvärdering av psykoakustisk testmetod och av OAE-registrering som komplementär metod
   Eva Andersson

51. Skolformens komplexitet – elevers erfarenheter av skolvardag och tillhörighet i gymnasiesärskolan
   Therése Mineur

52. Evaluating the process of change: Studies on patient journey, hearing disability acceptance and stages-of-change
   Vinaya Kumar Channapatna Manchaiah
53. **Cognition in hearing aid users: Memory for everyday speech**  
Hoi Ning (Elaine) Ng  

54. **Representing sounds and spellings Phonological decline and compensatory working memory in acquired hearing impairment**  
Elisabet Classon  

55. **Assessment of participation in people with a mild intellectual disability**  
Patrik Arvidsson  

56. **Barnperspektiv i barnavårdsutredningar – med barns hälsa och barns upplevelser i fokus**  
Elin Hultman  

57. **Internet Interventions for Hearing Loss**  
Examining rehabilitation Self-report measures and Internet use in hearing-aid users  
Elisabet Sundewall Thorén  

58. **Exploring Cognitive Spare Capacity: Executive Processing of Degraded Speech**  
Sushmit Mishra  

59. **Supported employment i en svensk kontext – förutsättningar när personer med funktionsnedsättning når, får och behåller ett arbete**  
Johanna Gustafsson  

60. **Effects of Specific Cochlear Pathologies on the Auditory Functions: Modelling, Simulations and Clinical Implications**  
Amin Saremi  
61. **Children with profound intellectual and multiple disabilities and their participation in family activities**  
   Anna Karin Axelsson  

62. **Lexical and Semantic Development in Children With Cochlear Implants**  
   Ulrika Löfkvist  

63. **Rethinking sound. Computer-assisted reading intervention with a phonics approach for deaf and hard of hearing children using cochlear implants or hearing aids**  
   Cecilia Nakeva von Mentzer  

64. **Assessing cognitive spare capacity as a measure of listening effort using the Auditory Inference Span Test**  
   Niklas Rönnberg  

65. **Employees with Aided Hearing Impairment: An Interdisciplinary Perspective**  
   Håkan Hua  

66. **Prosthetic and Orthotic Services in Developing Countries**  
   Lina Magnusson  

67. **Dealing with digits - Arithmetic, memory and phonology in deaf signers**  
   Josefine Andin  

68. **Time is of the essence in speech recognition: Get it fast or think about it**  
   Shahram Moradi  

69. **Effects of hearing loss on traffic safety and mobility**  
   Birgitta Thorslund  
70. **Aspekter på lärande vid dövblindhet - möjligheter och begränsningar för personer med Alström syndrom**
   Berit Rönnåsen

71. **Memory and communication in typically developing infants and children with Autism Spectrum Disorder: Behavioral and electrophysiological indices**
   Emelie Nordqvist

72. **Cognitive capacities and composite cognitive skills in individuals with Usher syndrome type 1 and 2**
   Cecilia Henricson

73. **Functioning and Disability in Adults with Hearing Loss Preparatory studies in the ICF Core Sets for Hearing Loss project**
   Sarah Granberg

74. **Speech recognition and memory processes in native and non-native language perception**
   Lisa Kilman

75. **To include or not to include: Teachers’ social representations of inclusion of students with Asperger diagnosis**
   Ann-Charlotte Linton

76. **Health and People with Usher syndrome**
   Moa Wahlqvist

77. **Children with mild intellectual disability and their families – needs for support, service utilisation and experiences of support**
   Lena Olsson
78. **Deafblindness**  
Theory-of-mind, cognitive functioning and social network in Alström syndrome  
Hans-Erik Frölander  

79. **Signs for Developing Reading**  
Sign Language and Reading Development in Deaf and Hard-of-Hearing Children  
Emil Holmer  

80. **Speech masking speech in everyday communication**  
The role of inhibitory control and working memory capacity  
Victoria Stenbäck  

81. **Neural and Cognitive Effects of Hearing Loss on Speech Processing**  
Eline Borch Petersen  

82. **A biopsychosocial approach to functioning, oral health and specialist dental health care in children with disabilities – Swedish and international perspectives**  
Johanna Norderyd  

83. **Samverkansprojekt, och sen då? – en uppföljande studie av samverkansprocessen kring barn och unga som far illa eller riskerar att fara illa**  
Ulrika Englund  

84. **Aural rehabilitation programs for hearing aid users**  
Evaluating and clinically applying educational programs, supported via telephone and/or the internet and professionally guided by an audiologist  
Milijana Malmberg  

85. **Tolkning vid förmedlade samtal via Bildtelefoni.net – interaktion och gemensamt meningskapande**  
Camilla Warnicke  
86. Vardagslivets aktiviteter. Handlingen som terapeutiskt redskap för personer med svårare psykiska funktionshinder – Analys av arbetsterapeuters berättelser med utgångspunkt i G H Meads teori om social handling
Marianne Boström

Antonia Pavli

88. Ej utgiven

89. Perception of disturbing sounds
Investigations of people with hearing loss and normal hearing
Åsa Skagerstrand
ISBN: 978-91-7529-229-8, 2018

90. Peer Interaction in Preschool: Necessary, but not Sufficient
The Influence of Social Interaction on the Link between Behaviour Difficulties and Engagement among Children with and without Need of Special Support
Madeleine Sjöman
ISBN: 978-91-88339-13-3, 2018

91. Prosodic and Phonological Ability in Children with Developmental Language Disorder and Children with Hearing Impairment In the context of Word and Nonword Repetition
Simon Sundström

92. From detection to intervention
Psychological aspects of online hearing rehabilitation
Peter Molander
ISBN: 978-91-7685-226-2, 2018
93. Participation in and outside school
Self-ratings by Swedish adolescents with and without impairments and long-term health conditions
Frida Lygnegård
ISBN: 978-91-85835-92-8, 2018

94. Återgång i arbete efter förvärvad hjärnskada
– livskvalitet, möjligheter och hinder
Marie Matérne
ISBN: 978-91-7529-267-0, 2018

95. Climbing up the hearing rehabilitation ladder
Elisabeth Ingo

96. Situationen på den svenska arbetsmarknaden för personer med funktionsnedsättning
En sekundäranalys om sysselsättning, yrkesnivå, diskriminering, stöd och anpassning för personer med olika typer av funktionsnedsättning
Tomas Boman
ISBN: 978-91-7529-302-8, 2019

97. Different is cool!
Self-efficacy and participation of students with and without disabilities in school-based Physical Education
Karin Bertills

98. Fonologisk utveckling hos barn med otitbenägenhet
En longitudinell studie i åldrarna 3;6 till 5;6 år
Helena Stålnacke

99. Life Strategies, Work and Health in People with Usher Syndrome
Mattias Ehn
ISBN: 978-91-7529-322-6, 2020
100. **Time to Plan:**
How to support everyday planning in adolescents with intellectual disability

Lisa Palmqvist

101. **Reasons for Language**
Language and Analogical Reasoning Ability in Children with Cochlear Implants and Children with Typical Hearing

Michaela Socher
ISBN: 978-91-7929-791-6, 2020