Radboud University Nijmegen
Faculty of Arts
Bachelor Thesis
LET-TWB300A

Radboud University

Assessment in speech intelligibility

What does the ideal intelligibility experiment look like?

Student: M.C.A. van Vorselen
Student number: S1021443
Date: Spring 2019
Course: Pre-master General Linguistics
Supervisor: W.A.J. Strik

Preface

In order to obtain my pre-master's certificate, I conducted research into assessment in speech intelligibility and collected the findings into a thesis. The writing of this thesis took place from February 2019 to July 2019 and is linked to the research of the Centre for Language Studies connected to the Radboud University Nijmegen.

The following literature study focusses on the conduction of speech intelligibility research. This thesis will look at the factors involved in an intelligibility experiment and will describe the ideal experiment based on examples in the literature.

I would like to start this thesis with a word of gratitude towards my supervisor, Mr. H. Strik, and the supervising PhD student, Ms. W. Xue, for their guidance, helpful contributions and insightful suggestions. I would also like to extend my gratitude to my family, friends and acquaintances, who have supported and helped me during the process of this thesis. A special thank you goes out to my peers, Hanna-Lina, Sarah, Tessa and Tim, who provided me with feedback, insights and opinions during the writing of this thesis.

Mignon C.A. van Vorselen
Nijmegen, July 2019

Table of contents

Abstract

Intelligibility is an important part of communication and therefore an interesting topic to study. Dozens of studies on intelligibility have been done in recent years. All these studies had different features, components and backgrounds, and each in their own have their advantages, disadvantages and factors of influence. This text discusses, weighs and observes multiple studies done with pathological speech and language learner speech in order to create a general approach. Additionally, the text uses the general approach to create the ideal intelligibility experiment.

The text focuses on the procedure of research and all its features, which will take a big part when one wants to study intelligibility. The thesis seeks to answer the question: 'What does the ideal intelligibility experiment look like?'. The purpose of this study is to present a review of the available literature concerning doing research in the field of intelligibility. It is hoped that this literature study will inform researchers and professionals about the procedure and current state of intelligibility research by patients and language learners.

1. Introduction

‘‘Intelligibility’ can be defined as a speaker’s ability to convey a message to a listener by means of acoustic signal’ (Lagerberg et al., 2015) Intelligibility is an important part of communication and therefore an interesting topic to study. In the past, many studies have been conducted in the broad field of speech intelligibility. These studies primarily focused on the pathological side of intelligibility. As a result, little is known about the factors of intelligibility among language learners. Pathological research has been focussing mainly on the pathology ‘dysarthria’. Dysarthria is a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to damage of the central or peripheral nervous system. It designates problems in oral communication due to paralysis, weakness, or incoordination of the speech muscles (Darley, Aronson & Brown, 1969).

When one starts to prepare an intelligibility study there are many decisions that need to be made. For example, ‘does one look at intelligibility in dysarthric patients or in healthy participants?’ ’Which method is most common and does this suit the research goal?’ ’What are the important factors to consider and what is their positive or negative impact on the results?’ If one wants to study this field, it very soon becomes clear that the possibilities seem endless. All these studies have different facets, components and backgrounds, and each in their own have their advantages, disadvantages and factors that take a big part. These factors might include for example acoustic elements of the speech sample (Markham & Hazan, 2002), word frequency (Bradlow & Pisono,1999) and rate of speech (Derwing & Munro, 2001), but articulation is shown to be the strongest contributor to intelligibility (De Bodt, 2002).

To bring order out of the chaos of the many possibilities, it seems befitting to discuss, weigh and observe multiple studies. In this thesis, the ins and outs of previous intelligibility studies will be explained and assessed in order to answer the question ‘what makes a good intelligibility experiment?’.

Following this question, a possible code of conduct for studying intelligibility will be formulated. These rules are gathered from a thorough literature study in which multiple articles have been assessed. The main goal of this study is to subtract the main conclusions from these articles, bring them together in the form of a thesis and use this information to create an experiment proposal, the ideal intelligibility experiment.

The general topic of this thesis is a study examining dysarthric speech and language learner speech and will therefore discuss many articles from several researchers. Several studies were consulted during the realization of this thesis (see 5. Bibliography). These studies are arranged in a framework (6. Appendix) and classified according to the characteristics 'subject', 'method', 'purpose', 'important factors', 'advantages' and 'disadvantages'. It has been attempted to create a fair diversity between pathological intelligibility studies and language learner intelligibility studies, but given that language learner studies are not widely conducted a skewness in the source material is noticeable.

This thesis starts with a theoretical framework in which the characteristics of different studies are examined and standardized. After this, a general approach for creating intelligibility

studies is formed. This approach is used to create four 'basic studies' (Appendix 2 Basic studies) that form the basis of the ideal experiment described in the recommendation. This ‘ideal experiment’ focuses on two different target audiences, speakers with dysarthria and speakers learning a second language. The purpose of this thesis is to create a design for the ideal intelligibility experiment.

2. Theoretical framework

When looking at previous listening experiments of speech intelligibility, different methodologies and study designs can be distinguished. Many studies opt for using the orthographic transcription of sample sets to assess speech intelligibility, while others opt for an assessment study in which sample sets are rated with 'Direct Magnitude Estimation', 'the Visual Analogue Scale' or 'the Likert Scale'. Studies also differ in certain characteristics. One of these characteristics is the selection of speakers who produced sample sets for assessing speech intelligibility. Some studies recruit samples of dysarthria patients, other studies recruit samples of language learners.

In order to provide more clarity, an elaboration on the different facets of research are given. Hence, this chapter includes for each research what has been done, what the underlying differences are and how this influences the research. This chapter starts with a literature study in which existing experiment designs, the used materials and the participant groups are examined, weighed and standardized. The chapter ends with a step-by-step guide which highlights the procedure of setting up and carrying out the studies present in the literature. The literature review forms the foundation for the proposed study in chapter 3. The ideal intelligibility experiment.

2.1 Design

The design of a listening experiment is the core of the research. It indicates what the research looks like and leaves its mark on the results. A good design of a listening experiment should be simple, efficient and can be repeated by a third party. When examining the literature, intelligibility studies are categorized based on the types of designs of the listening experiments, namely the orthographic transcription design and the human rating design. These two designs are explained as follows.

Orthographic transcription

A classic example of orthographic transcription is a study published by Breukelman and Yorkston (1979). In this research, Breukelman and Yorkston recorded dysarthric patients reading a passage aloud. All the recordings then were transcribed by listeners and rated on correctness. Therefore, the intelligibility of the speakers was calculated based on the transcriptions where a speaker was most intelligible with the most correct transcription.

This study design was often employed by other researchers, but with different characteristics (Kempler and Van Lancker, 2002; Dinnocenzo, Tjaden and Greenman, 2006; Khustad 2006; Stipancic, Tjaden and Wilding, 2016). Kempler and Van Lancker (2002) chose to include various elicitation tasks (spontaneous speech, repetition, reading, repeated singing and spontaneous singing) in the design, following the approach conducted by Martin (1990) who emphasizes the consistency of deficits across speech task, stating that patients with dysarthria show 'very little difference in articulatory accuracy between automatic-reactive and volitional-purposive speech' (p. 470). The assumption that intelligibility rates are consistent in a patient's speech across different speaking conditions is also contrary to the current treatment protocols for dysarthria (Kalf & van Zundert, 2017).

Dinnocenzo, Tjaden and Greenman (2006) highlight another part of the elicitation process, namely the impact of loudness. Speaking loudly has been shown to increase speech clarity (Duffy, 1995; Rosenbek & LaPointe, 1978) and is the core value of dysarthria therapies, including the Lee Silverman Voice Treatment (LSVT)[1] (Ramig, Countryman, Thompson & Horii, 1995). To test this hypothesis, the researchers took account into loudness and added utterances produced at the patient's maximum level of loudness to the sample of the classic orthographic transcription design.

Hustad (2006), on the other hand, chose to make adjustments to the assessment process of the orthographic transcriptions. Transcriptions were scored using three different paradigms; total word phonemic match, informational word phonemic match and informational word semantic match. In former research, Turner and Tjaden (2000) stated that patients with dysarthria tend to have shorter vowel durations and the first and second formant frequency values are more centralized for informative words as compared to content words. Hustad (2006) used these paradigms to measure the extent to which information was transmitted in addition to overall intelligibility.

Human rating design

Ganzeboom, Bakker, Cucchiarine and Strik (2016) decided to take a different approach. In addition to using a variant of the classical transcription experiment, they conducted a survey. The speech samples were therefore evaluated in two ways, by an orthographic transcription and by a subjective sentence level rating. This section will take a closer look at the use of a rating. A similar approach can be found in multiple studies, e.g. in Tjaden and Wilding (2011) and Weismer and Laures (2002). In those studies, speech samples obtained from patients with dysarthria were rated by listeners on the degree of intelligibility. Higher rating scores referred to more intelligible speech.

Scales

The majority of previous research used a scale for their rating design. A scale is a benchmark for ratios, in this case, intelligibility. There are multiple scales that could be used to rate the degree of intelligibility. Three frequently used methods are the Direct Magnitude Estimation (DME), the Visual Analogue Scale (VAS) and the Likert scale. The usage of the three methods is explained in the following paragraphs.

Direct Magnitude Estimation (DME)

DME has been used frequently as a perceptual scaling technique in studies of the speech intelligibility of people with speech disorders (Weismer & Laures, 2002). The method allows participants to assign a score to stimuli, so the ratios of the numerical assignments reflect ratios of sensory perceptions (Moskowitz, 1977). Listeners are instructed to scale samples of sensory stimuli that are twice as severe at a value of '200' and samples being half as natural have to be scaled a value of '50' (Eadie & Doyle, 2002). Magnitude estimation is a standardly used technique in psychological research to measure the judgements of sensory

[1] LSVT is a treatment program in which a patient with dysarthria is taught to increase his intelligibility by singing and speaking loudly (De Bodt, 2015, P. 273).

stimuli (Stevens, 1975). Schiavetti, Metz and Sitler (1981) suggested that DME is an appropriate scaling measure of speech intelligibility because intelligibility varies along a prothetic continuum and hence it is a fluid given and cannot be framed in numbers.

Visual Analogue Scale (VAS)

VAS is often used to evaluate the anaesthetic properties of various treatments and accomplishes this evaluation by measuring either pain relief or pain severity (Langley & Sheppeard, 1984). As seen in figure 1, the VAS is a continuous horizontal or vertical scale for subjective magnitude estimation and consists of a straight line, which carries the limits of a verbal description of each extreme of the statement to be evaluated. The subject is asked to indicate his opinion towards each statement on this line (shown in figure 1 as a red cross).

Question:	VAS-rating		
I have a sore throat after prolonged speaking	Strong rejection		Strong approval
I feel exhausted after prolonged speaking	Strong rejection		Strong approval

Figure 1 Example of a VAS-scale

Likert scale

In the book 'A technique for the measurement of attitudes' (Likert, 1932) Likert introduced a different method of measuring attitudes. Instead of a continuous scale, Likert used a scale that resembles a questionnaire (usually from 0 to 10) to retrieve concrete answers. The Likert-type scale consists of a series of statements, as seen in figure 2. A subject is asked to indicate whether he or she agrees or disagrees which each statement (shown in figure 2 as a red cross). Commonly, five degrees of attachment to the given statement are provided; 'strongly agree', 'agree', 'undecided', 'disagree' and 'strongly disagree' (Arnold, McCroskey & Prichard, 1967).

Question:	Likert-rating				
I have a sore throat after prolonged speaking			X		
	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
I feel exhausted after prolonged speaking				X	
	Strongly disagree	Disagree	Undecided	Agree	Strongly agree

Figure 2 Example of a Likert-scale

2.2 Material

Naturally, material is crucial in order to conduct any research. Due to the nature of the design, a listening experiment for measuring speech intelligibility, the material in the reviewed studies was a sample set with speech utterances and a questionnaire corresponding to the research design.

Samples

Samples of spoken utterances are needed in order to conduct an intelligibility study, whether the experiment uses an orthographic transcription design or a human rating design.
In the earlier days, this was a time- and material consuming job. Breukelman and Yorkston (1979) chose to use a series of listening tapes, while Khaghaninejad (2018) simply made use of a recorder and a laptop. In approximately 40 years, technology has taken a giant leap and it has become much easier to do research with recordings.

The practical components related to recording the samples are interesting, but to compile the ideal experiment one has to look at what these samples consist of. First, a decision has to be made between using spontaneous speech or using bounded speech (fixed utterances). Transcribing spontaneous speech rather than lists of words or sentences is generally agreed to be a highly reliable method that gives a more ecologically valid picture of the speaker's communicative ability (Kwiatkowski and Shriberg 1992; Whitehill 2002). But looking at the literature, it soon became clear that most studies opt for a form of bounded speech. Frearson (1985) showed that dysarthric speakers were more intelligible when reading aloud than speaking spontaneously. In line with this study, Darley, Aronson and Brown (1969) elicited speech using a paragraph called “the Grandfather Passage”(Appendix 3 Grandfather passage), a text that contains nearly all of the phonemes of American English (Fairbanks, 1960).

Although many studies were conducted by using a paragraph as elicitation material, this does not seem to be the norm. Several elicitation methods can be distinguished within the literature namely spontaneous speech, reading aloud and singing. The distinction is further explained in Table 1.

An interesting point of view was represented by Khaghaninejad (2018). In this study, the intelligibility of Iranian second language learners of English was studied. In this study, sentences with minimal pairs were added to the eliciting material, in order to create a measurable parameter for the listeners. The use of minimal pairs for intelligibility research is not a new concept and could be seen in many studies (e.g. in Chin & Finnegan. 1998; Hayes-Harb, Smith, Bent & Bradlow, 2008 and Hodge & Gotzke, 2011).

One bias to be aware of during creating the samples is the context bias. There is a possibility that the listeners are able to understand the speaker based on the context of the speech sample. To eliminate this bias, several studies (Stark & McClelland, 2000; Bowers, 1994; Dorfman, 1994; Hamann & squire, 1997) were conducted using nonsense words. With this addition, the listeners solely make a notion of the sound itself instead of any confounding variables. This could result in a more pure intelligibility score, solely based on pronunciation.

Table 1 Arguments for distinguishing between elicitation methods

Elicitation method	Reason
Spontaneous speech	The results of spontaneous language analysis, in combination with the results of other language tests, can be used to draw up an appropriate therapy plan. Spontaneous language analyses are also used to determine the degree of progress after therapy. (Grande, Hussmann, Bay, Christoph, Piefke, Willmes, et al., 2008)
Reading Aloud	Dysarthric speakers are more intelligible when reading aloud than when speaking spontaneously. This is probably due to the fact that patients do not have to rely heavily on affected brain functions.
Singing	Clinical observations suggested that dysarthria patients are more intelligible when singing than speaking (Waters, 1994).

Questionnaires

The nature of the research, with a very heavy focus on the output of listeners, requires the researchers to focus on the material for the listeners. In both designs, the listeners are asked to answer some sort of questionnaire, such as a transcription-based questionnaire in the first design and a rating scale based questionnaire in the second design.

In both designs, the experiment starts with a biographical questionnaire to determine which listeners can participate in the experiment (see '2.3 Participants' for the ideal participants). The biographical questionnaire was not specifically mentioned in the majority of studies, but this does not detract from the fact that it is an important part of the study since the research cannot take place without correct participants. A good biographical questionnaire predicts the hypothetical behaviour during research based on past behavioural patterns and filters participants on inclusion criteria (Hustad, 2006).

After creating the samples and selecting the correct participants most experiments started with an estimation of intelligibility, which can be done in various ways. Yorkston and Beukelman (1978; 1980) created an interesting perspective in questionnaire designs by using different transcription methods. They chose a transcription design in which the listeners had to transcribe single words, target words and complete sentences with missing words and answer a multiple-choice quiz about the heard samples.

2.3 Participants

In each reviewed article, the authors chose to use two different participant groups; one group of participants were asked to record samples (hereinafter referred to as speakers) and the other group of participants was asked to assess the intelligibility of the samples (hereinafter referred to as listeners). In this section, a discussion on both groups of participants is presented.

Speakers

As mentioned earlier, samples of spoken utterances are needed to do an intelligibility study. To create those samples, speakers are needed. This literature study focuses on speakers with dysarthric speech and speakers who are learning a foreign language. When looking at the literature, it became clear that the most emphasis in the past used to be placed on investigating dysarthric speech. Investigating language learner speech is a relatively new phenomenon. The basic characteristics of the speaker groups in the literature are listed in Table 2.

Table 2 characteristics of speakers

Research	Amount	Age	Remarks
Beukelman and Yorkston (1979)	9 speakers	20 to 71 years (mean 38 years)	Mild to severe ataxic, spastic or hypokinetic dysarthria
Kempler and Lancker (2002)	1 speaker	74 years old	Parkinson's Disease
Ganzeboom, Bakker, Cucchiarini and Strik (2016)	8 speakers	Unknown	Stroke, Parkinson's Disease or traumatic brain injury
Dinnocenzo, Tjaden and Greenman (2006)	1 speaker	29 years old	Closed head injury
Hustad (2006)	12 speakers	18 to 60 years old	Dysarthria secondary to cerebral palsy
Tjaden and Wilding (2011)	12 speakers	42 to 74 years old (mean 63 years)	Parkinson's Disease (hypokinetic dysarthria)
Weismer and Laures (2002)	4 speakers	Unknown	Parkinson's Disease or traumatic brain injury
Khaghaninejad (2018)	5 speakers	23 to 28 years old	Iranian second language learners (of English)

As 12 is the largest number of speakers, it seems that previous studies did not use large groups of speakers to create samples. Although it is preferred to have more speakers, having more speakers means more data, which also means more labour intensive work in the revising of the data. In addition, it can also be difficult to find suitable speakers due to various reasons. A choice for a small speaker group, therefore, seems the more obvious choice.

Another remarkable fact is that the literature examines only three forms of dysarthria (ataxic, spastic and hypokinetic), while medical science distinguishes seven types of dysarthria, namely, bulbar, myogenic, spastic, ataxic, hypokinetic (Parkinsonism), hyperkinetic and mixed dysarthria (Dharmaperwira-Prins, 2005). Various dysarthrias have different characteristics (Table 3) and are clinically distinguishable, and therefore need to be equally represented in the examination (Darley, Aronson & Brown, 1969).

Table 3 types of dysarthria and their characteristics (Based on Dharmaperwira-Prins, 2005).

Dysarthria type	Cause	Speech characteristics
Flaccid dysarthria *Bulbar*	Caused by damage in the nucleus of the motor neuron in the brainstem (virus infection/polio, tumour, cerebral infarction, progressive degeneration, congenital abnormality) or caused by damage of the nerves (impact or injury, pressure, intoxication, neuritis).	Continuous breathiness, diplophonia, audible inhalation, short phrases, nasality.
Flaccid dysarthria *Myogenic*	Caused by *myasthenia gravis* (disturbed transmission over the myoneural connection) or dystrophy, inflammation/polymyositis (damaged muscle).	Continuous breathiness, diplophonia, audible inhalation, short phrases, nasality, rapid deterioration and recovery with rest
Spastic dysarthria	Caused by multiple or bilateral cerebral infarctions, infantile cerebral palsy, trauma, extensive brain tumors, multiple sclerosis, progressive degeneration of the brain, or a combination of all of the above.	Inaccurate consonants, monotony, decreased articulation focus, hoarse voice.
Ataxic dysarthria	Caused by a localized lesion (caused by a tumour, cerebral infarction or trauma) or caused by generalized damage (caused by degeneration, encephalitis, intoxication (Syndrome of Korsakov), lung cancer, demyelinate (multiple sclerosis), cerebral infarction.	Inaccurate consonants, excessive and equal articulation emphasis, alternating deterioration of articulation.
Hypokinetic dysarthria (Parkinsonism)	In 97% of cases, this kind of dysarthria is caused by	Monotony, decreased articulation focus, inaccurate

	Parkinson's disease, other causes are post-encephalitis, congenital abnormalities, trauma (caused by boxing), poison, medication use, multi-infarction.	consonants, short fast speaking parts.
Hyperkinetic dysarthria	Caused by hereditary diseases (Huntington's disease, Wilson's disease), Gilles de la Tourette, Kernicterus in rhesus antagonism, streptococci infections, diplegia spastica infantilis, jaundice, oxygen deficiency at birth.	Inaccurate consonants, extended pauses, varying speed of speech, monotony, distorted vowels, hoarse voice, pressed phonation dysphasia.
Mixed dysarthria	A mix of different dysarthria types. Caused by (among other things) Amyotrophic lateral sclerosis (ALS), Multiple Sclerosis (MS), Wilson's disease.	Seriously disturbed articulation, slow and difficult speech, hypernasality, severe hoarseness, impaired loudness control, decreased emphasis, monotony.

Listeners

If one wants to do intelligibility research, one of the most important elements besides design and material is the listener. Listeners are needed to gather data for the actual research. The listener groups involved in the literature were fairly diverse and to create a general image, the basic characteristics of the listener groups are set out in Table 4.

Looking at the statistics in Table 4, it can be seen that a participant group consisting of seventy listeners is the norm. Factors like age and experience do not seem to be of great importance. This is at odds with earlier studies (Weismer, 2006) which suggested that there is evidence for certain listener characteristics affecting the utility of signal-independent ques (p. 212). Kent, Miolo and Bloedel (1994) even dared to go as far as to say 'the idea that a single intelligibility score can be ascribed to a given individual apart from listener and listening situation is somewhat a fiction' (p. 81). The intelligibility of a speaker could be affected by a number of different external variables related to the listeners' experience and their personalities (Lagerberg, Asberg Johnels, Hartelius & Persson, 2015). Prior exposure to some kind of acoustic speech signal is what science calls 'familiarization'. A listener who has experience with listening to abnormal speech tends to give higher scores on intelligibility and thus creates a bias in research (Spitzer, Liss, Caviness & Adler, 2000). Relatedly, Liss, Spitzer, Caviness and Adler (2002) reported higher intelligibility scores for listeners who were familiar with dysarthric speech, comparing to those unfamiliar with dysarthric speech.

Table 4 characteristics of participant groups

Research	Amount	Age	Occupation
Beukelman and Yorkston (1979)	108 listeners	Unknown	University students, clerical staff and health professionals
Kempler and Lancker (2002)	64 listeners	23 to 78 years (mean 44 years)	Unknown
Ganzeboom, Bakker, Cucchiarini and Strik (2016)	36 listeners	19 to 73 years	Unknown
Dinnocenzo, Tjaden and Greenman (2006)	120 listeners	18 to 60 years (mean 28 years)	Unknown
Hustad (2006)	144 listeners	Mean age of 21.25 years	Unknown
Tjaden and Wilding (2011)	70 listeners	Mean age 32	University students and clerical staff
Weismer and Laures (2002)	10 listeners	Unknown	University students of the speech department
Khaghaninejad (2018)	5 listeners	24 to 31	M.A. Students of linguistics

In addition to having prior knowledge, the number of times a listener heard a sample seems to be another impact factor. Nygaard, Sommers and Pisoni (1994) found that perceptual learning of the speaker's voice can facilitate the listener's perception of words, and Lagerberg et al. (2015) found an effect of the number of presentations on listener transcriptions and explored the reliability in the assessment of speech intelligibility in children. They concluded that the mean intelligibility score from all listeners to a given speech material increased by approximately three percentage points each time the listeners transcribed the same material. This result, therefore, shows a learning effect and this effect must be taken into account when drawing up the procedure.

2.4 Procedure

In this section, a general procedure is drawn. This general procedure is a step by step plan and is based on a combination of research procedures from the literature study. This procedure plan will be used in Chapter 3. *The ideal intelligibility experiment* to create the ideal intelligibility experiment. When one looks at the basic steps of research, six general steps can be distinguished, the recruitment of participants, the collection of material, the evaluation and categorization of samples, the rating of the samples, the revising of responses and the elimination of false outcomes and the establishment of the results (as seen in Figure 3). A further explanation of these steps is made in the following paragraphs.

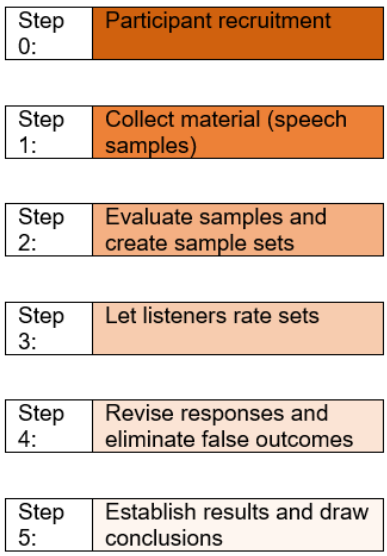

Figure 3 Visualized step-by-step plan for the procedure process

Step 0: Speakers and listeners recruitment

Speakers and listeners are needed to do research into a human function such as speech. The process used to locate and recruit speakers in a qualitative study is important for controlling bias and for efficiently obtaining a representative sample (Arcury & Quandt, 1999).

Looking at the specific speaker target groups, it can be concluded that speakers are often recruited at locations related to their characteristics. For example, the language learners in Khagnaninejad (2018) were recruited as foreign students on an English-language university campus and the dysarthric speakers in Weismer and Laures (2002) were recruited in medical facilities based on their medical file.

Focusing on the other participant group, the listeners, broadly three categories of recruitment locations can be identified: University faculties, professional locations and others. The choice of location has a strong influence on the (intelligibility) results of the research. If one chooses to recruit professionals or paramedical students, there is a good chance that they have experience with abnormal speech creating a familiarization effect (as discussed in Section 2.3 Participants). The intended listeners will have to fill in a biographical questionnaire, which will be used to determine whether they are suitable to participate in the research as a listener.

Step 1: Collect material

When the speakers and listeners are recruited, the material needs to be collected. Various elicitation methods are used in the observed studies (as discussed in Section 2.2 Material). Kempler and Lancker (2002) used different speech production tasks (spontaneous speech, repetition, reading, repeated singing and spontaneous singing) to elicit speech.

In order to elicit these utterances, different methods such as random conversations, reading aloud words and sentences and reading aloud passages like ‘the grandfather passage’ can be used. The characteristics of the samples used in the literature are visualized in Table 5.

Table 5 characteristics of sample sets

Research	Sample sets	Repetition/familiarization	Length
Beukelman and Yorkston (1979)	Two paragraphs and one 50-words list.	No, randomized trails, without familiarization.	
Kempler and Lancker (2002)	30 utterances	Yes, but each utterance in another condition.	3 to 15 words (mean length of 8.2 words)
Ganzeboom, Bakker, Cucchiarini and Strik (2016)	Lists of single words, declarative SUS sentences, interrogative SUS sentences and regular sentences.	Yes, three example speech fragments	
Dinnocenzo, Tjaden and Greenman (2006)	Grandfather Passage, conversational monologue	Yes, the Grandfather Passage and random ordering of words in this Passage	5 to 6 words (mean length of 5,4 words)
Hustad (2006)	Three narrative passages	No, randomized trails, without familiarization.	5 to 8 words (49 different words)
Tjaden and Wilding (2011)	John Passage, monologue	Yes, the utterances were heard twice	
Weismer and Laures (2002)	19 sentences	Yes, a chosen standard used for repetition	8 to 9 syllables
Khaghaninejad (2018)	Conversational monologue, reading aloud sentences	No, randomized trails, without familiarization	

According to literature, a sample set seems to exist of a paragraph and a conversational monologue. It seems that a familiarization set is used to explain the task to the listener and familiarize them with dysarthric speech.

Step 2: Evaluate samples and create sample sets

Whether the samples are single words or full paragraphs, a method of sample grouping should be applied. These sample sets are needed to make the research process as optimal as possible for the researcher and the listener. In most literature, there seems to be a preference for the use of a unit of thirty utterances per set (Bender, Canitto, Murry & Woodson 2004; Kempler & Van Lancker 2002).

To create the sample sets, utterances need to be observed. Often the lengths of the utterances are noted, for example in Weismer and Laures (2002) the number of syllables (8 to 9) played an important role in the usage of the utterance in the sample set. Some studies hold their own characteristics (such as length) as shown in Table 5 in step 1. According to the literature study, the most ideal sample set exists of a passage of some sort, a monologue and reading aloud of words or sentences (as seen in table 5). Kempler and Lancker (2002) indicated that the ideal sample set consisted of only grammatical utterances that were complete phrases and sentences with phonetically identifiable onsets and offsets. The variability (phrase length and topic) in the sample set must be maximized to limit linguistic, grammatical and thematic redundancy, personal details and rare terms have to be eliminated and there should be a minimized repetition of specific words.

Step 3: Let listeners rate sets

After determining the sample sets, the listeners are instructed to listen to the sample sets. All listeners will start the research with a biographical questionnaire, whether they participate in a transcription or an assessment study. The biographical questionnaire determines whether a listener fits the inclusion criteria for participation in the study. A clear example of this is visible in Bender, Cannito and Murry (2004), where a biographical questionnaire checked whether the listeners were inexperienced with dysarthric speech and whether they have any hearing problems.

After determining the listener group using a biographical questionnaire in step 0, the listeners are asked to rate the intelligibility of the speech samples. For practical purposes, the literature indicated that the listeners have to be placed in a quiet environment and has to make either a transcription (the listeners hear a speech sample and will be asked to transcribe what they had heard) or a rating (the listeners hear a speech sample and will be asked to rate the intelligibility on a scale). The results formed in this step will enter the correction process.

Step 4: Revise responses and eliminate false outcomes

After collecting the transcriptions and ratings it is necessary to take a closer look at the outcomes. It is important to revise the responses and eliminate false outcomes. This step is very important, as it makes the establishment of the results purer. A clear way of response management can be found in Hustad (2008): if more than 1 in 10 listeners answered a general question about the samples incorrectly, this question was taken from the pool of questions with the reason that the question would be unanswerable for experienced listeners and therefore only be a disturbance in the overall result.

Bender, Canitto, Murry and Woodson (2004) used another approach. After scoring the results with a computer, the files were hand-edited to check for accuracy. This was done by a human to allow more liberal scoring conventions. The responses that the computer counted as wrong (and were tolerated by the human researcher) were phonemic spelling errors, homonyms, subject/verb contractions, verb tense and pluralization. This method of scoring was used to generate a more realistic rate of intelligibility connected to the basic meaning of the utterance.

Step 5: Establishing results and drawing conclusions

The data was collected and checked before the final step. When conducting a study, there are a number of expectations. These expectations are linked to a hypothesis. In this step, a conclusion supported by the results is drawn. Then the revised data is used to prove the earlier made hypothesis or to invalidate them. The hypothesis, research and data are different for each study, so each study in their own will have a unique conclusion.

3. The ideal intelligibility experiment

In this part of the thesis, the previously acquired knowledge from the literature will be used to create four basic experiments. In this chapter, the experiments will be explained following the step-by-step plan for procedure process created in Section 2.4 Procedure. An overview of the experiments created will be highlighted in the conclusion. These basic experiments are generalizations from the literature and can be adapted or combined to optimize the research as required. These experiments focus on two target speaker groups, dysarthric speakers and second language learners. If one looks more closely at the literature, two different experiments can be distinguished per target group, namely a transcription experiment and a rating experiment. Given that, this thesis focuses on two target groups it means that four different experiments can be formed. In Table 6, these four experiments are indicated in four different colours. Throughout this chapter, the various experiments are indicated by the corresponding colour, respectively 'the yellow experiment', 'the red experiment', 'the green experiment' and 'the blue experiment'.

Table 6 The four 'ideal' experiments

Yellow experiment	Transcription experiment with dysarthric speakers
Red experiment	Rating experiment with dysarthric speakers
Green experiment	Transcription experiment with language learners
Blue experiment	Rating experiment with language learners

In the upcoming paragraph, the four experiments will be set out by the steps of the general procedure (figure 3) created in Section 2.4.

Step 0: Speaker and listener recruitment
Recruitment of speaker groups

As stated earlier, speaker participants are often recruited at locations related to their condition. This means that dysarthric speakers are recruited from specialist medical institutions such as hospitals, rehabilitation centres and first line professionals and second language learners will be recruited from international institutions, language schools and immigration procedures.

Dysarthric speakers

In the perfect situation, all seven dysarthric types are represented in the speaker group. Ideally, all these types are represented in the participant group according to their prevalence. However, at this moment, it is unknown what exactly the prevalence of dysarthria in adults internationally is (Sluijmers et al., 2016) and another approach must be considered. In order to be able to make a clear distinction between dysarthria patients in terms of intelligibility, the approach used Weismer and Laures (2002) is therefore chosen. In this study, the speakers with dysarthria were chosen on the basis of their speech characteristics. If this is translated into the experiments that are created in this thesis, this means that the speakers obtained should be categorized by professionals on a scale from 1 to 100 and have to be grouped,

with 0 - 20 being poorly intelligible, 21 - 40 being moderately intelligible, 41 - 60 being insufficiently intelligible, 61 - 80 being somewhat intelligible and 81 - 100 being intelligible. After this classification process, it can be examined whether a speaker participant group can be selected in which a representative of each intelligibility level is available.

Based on the literature, a group of at least 12 speakers is recommended, in which a diverse number of dysarthrias types and severities is represented. For the aforementioned design, this means that there are ideally more representatives per intelligibility level. The decision to gather multiple participants is based on the study published by Ganzeboom et al. (2016) to avoid speaker familiarity influencing the procedure.

Language learners

Because there has been less research focussing on the intelligibility of language learners in the past, there is less mentioned about participants and participant recruitment in the recruitment. However, just as with dysarthria patients, it is wise to collect enough participants so that the sample sets will become as diverse as possible. It would, therefore, be best to use a group of at least 12 participants. Depending on what is being tested, there is the option of varying with the speakers' characteristics (e.g. native language families). For example, the researcher could choose to allow a variation in language families among the participants or choose to create a homogeneous group with the same native language.

Control/ Comparison group

To give the listener group an idea of a 'normal' voice, it is advisable to use a control group as seen in Weismer and Laures (2002). Samples from this group can be used at the start of the experiment to build up a basis of ‘normal’ speech so the listeners can assess how the 'real' sample differs from the 'control' sample.

Recruitment of listener group

To follow the example set by the literature, a listener group of at least 70 participants should be recruited. It is important that a heterogeneous group is created, with a large variation in age and gender, so external factors do not affect the result. In addition, it is important to ensure that the listener group has as little prior knowledge as possible in order to prevent a publicity effect. To achieve this, inclusion- and exclusion criteria will have to be established. Exclusion criteria that matter in the ideal experiments are:

Participants with severe hearing difficulties.
Participants experienced with dysarthric/ language learner speech (including speech- and language therapists, speech- and language pathologists and language teachers).
Participants that have another mother language than the tested language.

Step 1: Collecting material

Samples are needed for all four experiments. These can be the same for both target groups (the speakers with dysarthria and the language learners). The speech samples will be gathered based on a practical method inspired by Johansson et al. (2014).

This means that recordings will be made in a quiet environment. The mouth-to-microphone distance will be 15 centimetres and the samples will be recorded using an audio recorder of some sort. The participant will be instructed to speak in their 'normal talking voice' at a volume of 40 to 60 dB with an average speaking rate (around 100 to 150 words per minute) to collect samples. Based on the literature, the following components should be included:

A scripted interview made into a monologue *(Semi-spontaneous)*

The scripted interview is set after the example from Kempler and Lancker (2002). The participant will have to respond to many conversational prompts such as 'Tell me about your childhood', 'tell me what your house looks like' and 'tell me about your favourite memory' etc.

A word list of 50 words *(Reading aloud)*

For the list of 50 words, the participant is asked to read aloud 50 words that are not related to each other. It is possible for the researcher to add minimal pairs to this word list, in order to create clarity about the intelligibility of various elements of the speech signal (such as the intelligibility of voiced versus voiceless sounds). In the case of ideal designs, this will be used.

A sentence list of 20 sentences *(Reading aloud)*

In the case of sentences, predictability plays a major role. It may be that listeners can correctly predict a sentence based on the context. To counter this, the sentences will be grammatically correct nonsense sentences.

A paragraph (similar to 'the Grandfather passage') *(Reading aloud)*

In this section, the participants are asked to read a text in which all the sounds of the language studied occur. For English, this is for example 'The grandfather passage' or 'The rainbow passage'. It is important that all speakers read the same text. For studying language learners, it is interesting to look at sounds that do occur in the language studied and not in the native language, while the pronunciation of these sounds can affect intelligibility.

Step 2: Evaluate samples and create sample sets

As stated earlier the samples must be cleaned up. This means that incorrect and incomplete utterances must be removed from the samples. Sample sets must be created after this process. When all the data is taken from the literature and placed next to the components from step 1, it seems that the ideal sample set looks like mentioned in Table 7.

In addition to building an ideal sample set, it is also important that the set is randomized. This means that several sample sets should be created so the researcher can switch between listeners.

Table 7 the ideal sample set

5 words spoken by control participant	
20 words	Spoken by at least 5 speakers with different characteristics
5 sentences spoken by control participant	
15 sentences	Spoken by at least 5 speakers with different characteristics
Half a minute of monologue spoken by control participant	
5 minutes of monologue	Spoken by at least 5 speakers with different characteristics
half a minute of paragraph spoken by control participant	
5 minutes of paragraph	Spoken by at least 5 speakers with different characteristics

Step 3: Let listeners rate sets

When the sample sets are created and the listeners are selected, the main part of the experiment will take place. In this step, the two experiment types will be created, namely the transcription type (which corresponds with the yellow and green experiment) and the rating type (which corresponds with the red and blue experiment). The listeners are placed individually in a quiet placed. Here they receive a brief instruction about the task and they start listening to the practice utterances.

The transcription experiment (yellow and green)

In these two experiments, the listeners will dissect the sample sets and make a word by word transcription. This means that the participants will hear a sample set and note down what they have heard. To prevent the task from becoming a labour-intensive process, it is possible to omit the samples from the control group, since the participants do not have to judge the intelligibility personally. The participants are given the opportunity to pause the sample band, but they are not given the opportunity to listen to the sample again to prevent a familiarization effect from happening. The participants are also not able to change their answers after writing them down, because a correction can lead to a false intelligibility score.

The rating experiment (red and blue)

In these two experiments, the listeners will dissect the sample sets and give the sample a score on the VAS-scale (figure 4). The VAS-scale runs from 1 to 100 and will be grouped as follows, with 0 - 20 being poorly intelligible, 21 - 40 being moderately intelligible,

41 - 60 being insufficiently intelligible, 61 - 80 being somewhat intelligible and 81 - 100 being intelligible. Given that it is a VAS scale, the listeners will get the task of giving a liquid score by making a single clear, unambiguous mark on an undifferentiated line. The scores will be grouped afterwards, whereby the given score will be linked to the aforementioned groups.

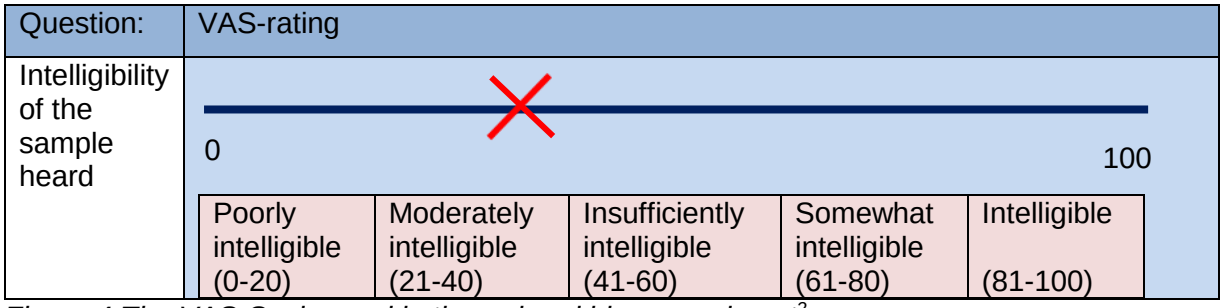

Figure 4 The VAS-Scale used in the red and blue experiment[2]

The participants are not given the opportunity to pause and rewind the sample band. They are also not able to change their answers and to listen to the control group samples at all times, to prevent learning effects from happening (Kreiman, Gerratt, Kempster, Erman & Berke, 1993).

Step 4: Revise responses and eliminate false outcomes

When the listeners have finished transcribing or rating the samples, the responses will be revised. Following the procedure used by Bender et al. (2004), the questionnaires first will be revised with a computer program. This program will be able to make a statistical calculation to state the average intelligibility score per speaker. This program is also able to make a draft rating of the transcription. However, to allow a more liberal scoring convention, the transcriptions will be scored by researchers after the computer rating. This means that phonemic spelling errors, homonym subject/verb contractions, verb tense and pluralization will be allowed in the samples.

The ratings of intelligibility will also be revised with a computer program. This program will be able to match the mark on the VAS scale to the assigned group and then make a statistical calculation to state the average rate of intelligibility per speaker.

Step 5: establishing results and drawing conclusions

When all responses have been checked and cleaned up, results can be determined. This requires less effort for the rating experiment while the standardization process creates a list in which the speakers are ranked on their intelligibility.

Things are different for the transcription study. The transcription study will look at the number of mistakes made in the transcription per speaker. This means that the speaker with the largest number of errors in the transcripts of his speech is the least intelligible speaker. This labelling allows classification to be made in which the most intelligible speaker has the least

[2] The red frame with the scoring groups will not be visible to the listeners, but has been added to provide additional clarity about the layout that the researcher will use for rating the results.

number of transcription errors and the least intelligible speaker has the most transcription errors.

When the results have been established, the conclusions can be drawn. These conclusions depend on the research question, the hypothesis and the target group, so they are different for each study.

4. Conclusion and discussion

This thesis set out to answer the question 'What makes a good intelligibility study?'. In order to assess this question, a literature review was carried out. Several methods were compared and a general course of action has emerged from this. This course consists of a research design, a description of the material and characteristics of the participant groups. Subsequently, the course of action was converted into a step-by-step procedure plan, with which it is possible to see at a glance how an intelligibility study has been constructed.

This step-by-step procedure plan is used in Chapter 3 to create four intelligibility experiments. These experiments seem to steer a middle course between all the different elements mentioned in the literature and together or separately they can provide a clear picture of the ideal intelligibility study. The four experiments involved two designs, a transcription design and a rating design. All experiments have their own special elements and factors.

Table 8 Description of experiments

Yellow experiment	This experiment is a transcription experiment in which the listeners make a transcription of speech samples from a dysarthric speaker. The number of faults made in the transcription is used to calculate the intelligibility of the speakers. Design of the experiment: Transcription experiment
Red experiment	This experiment is a rating experiment in which the listeners make an intelligibility rating on the VAS-scale based on the speech samples from a dysarthric speaker. The average rating per speaker will be used to create a ranking in intelligibility. Design of the experiment: Rating experiment
Green experiment	This experiment is a transcription experiment in which the listeners make a transcription of speech samples from a language learner speaker. The number of faults made in the transcription is used to calculate the intelligibility of the speakers. Design of the experiment: Transcription experiment
Blue experiment	This experiment is a rating experiment in which the listeners make an intelligibility rating on the VAS-scale based on the speech samples from a language learner. The average rating per speaker will be used to create a ranking in intelligibility. Design of the experiment: Rating experiment
Characteristics of the sample sets (used in all of the four experiments)	The sample sets consists of a monologue, a wordlist, a sentence list, a paragraph. These utterances will contain different sentence structures with different, minimal pairs and all the sounds in the target language. To create the opportunity to make a more grounded rating, the sample sets used for the rating experiments (red and blue) will be expanded with speech samples from native speakers to determine a zero value.

To answer the main question ‘what makes a good intelligibility experiment?’ It is important to take a close look at the experiments that have been suggested (Table 6). Four experiments have been created, the yellow experiment, the red experiment, the green experiment and the blue experiment. These experiments can be combined to create a more inclusive study. The experiments are described in Table 8.

A good intelligibility experiment consists of several designs, namely a transcription component and a scale component. In these components, multiple speech samples are used from various target groups, within which a great diversity exists. These samples are listened by a diverse group of listeners with little to no experience with the target group. The listeners then make a transcription or a VAS rating based on the samples.

Discussion

When one looks at the literature list, some questions may arise. The literature used for the theoretical framework is not evenly balanced. The majority of the literature study is based on articles regarding dysarthric speech. One of the main reasons for this uneven framework is the fact that intelligibility research in second language learners is still in their infancy. Little research has been conducted in the field of intelligibility of second language learners. With the most second language learning studies being focused on bilingual children and their development, it is hard to find any research that has been done into the field of intelligibility of an adult language learner (Dörnyei, 2005). The conclusions drawn by the literature are mostly based on dysarthric speakers and are then transferred to second language learners. This makes the experiments created for second language learners (green and blue) less founded, so more research in the field of intelligibility with second language learners is desirable.

Another important factor to keep an eye on is the fact that second language learner studies have the tendency to quickly turn into an assessment of the proficiency to speak a language on a native-like level, instead of an assessment on speech intelligibility and voice quality. To counter this, it is important for the speakers to adhere to fixed utterances and for the researcher to keep a close eye on the proposed purpose of research.

A number of studies have been left out of the literature for practical reasons. This concerns studies that focus on intelligibility research in combination with medication (Johansson, et al. 2014; Sandström, Hägglund, Johansson, Blomstedt & Karlsson, 2015; Bender, Cannito, Murry & Woodson 2004). They provided an interesting perspective on the possible treatment of dysarthria but are not profitable for a standard intelligibility study for several reasons. One of those reasons is that the effect of drugs, such as botox, is often short-lived. In addition, it is not allowed for a standard researcher to conduct studies involving medication without proper training. Furthermore, such research does not pursue the goal that is maintained within this literature study. This does not alter the fact that such studies offer an interesting perspective and can certainly be an addition to intelligibility studies.

In conclusion, in the future, it would, therefore, be interesting to do more research on second language acquisition and to do follow-up research on the influence of medication on intelligibility. Given the purpose of this literature study, the focus is on preparing general research. If a follow-up study is drawn up, it would be interesting to connect the research

more to society. This could, for example, be done by focusing the follow-up research more on creating therapy options, facilitating the rehabilitation process for dysarthria patients or improving the education of second language learners.

5. Bibliography

Arcury, T., Quandt, S. (*1999*) *Participant Recruitment for Qualitative Research: A Site-Based Approach to Community Research in Complex Societies.* Human Organization: Summer 1999, Vol. 58, No. 2, 128-133.

Arnold, W.E., McCroskey, J.C. & Prichard S.V.O. (1967). *The Likert-type scale,* 15:2, 31-33, DOI: 10.1080/01463376709368825

Bender, B. & Cannito, M. (2004). *Speech intelligibility in severe adductor spasmodic dysphonia*. Journal of Speech, language, and hearing research, vol. 47 (February 2004), 21-32.

Beukelman & Yorkston (1978). *A comparison of techniques for measuring intelligibility of dysarthric speech.* Journal of communication disorders 11 (1978), 499-512.

Beukelman & Yorkston (1979). *The relationship between information transfer and speech intelligibility of dysarthric speakers.* Journal of communication disorders 12 (1979), 189-196.

Bowers, J. (1994). *Does implicit memory extend to legal and illegal nonwords?* Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 534-549.

Bradlow, A.R. & Pisoni, D.B. (1999). *Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors*
Journal of the Acoustical Society of America, 106 (1999), 2074-2085

Brauer, T. (2010). *Lee-Silverman-Voice-Treatment bei Morbus Parkinson.* Sprache Stimme Gehör 2010(34), 186

Chin, S.B., Finnegan, K.R. (1998). *Minimal Pairs in the perception and production of speech by Pediatric Cochlear Implant Users.* Indiana University: Research on spoken language processing.

Dagenais, P., Garcia, J., & Watts, C. (1998). *Acceptability and intelligibility of mildly impaired dysarthric speech by different listeners.* In M. Cannito, K. Yorkston, & D. Beukelman (Eds.), Neuromotor speech disorders: Nature, assessment and management, pp. 229–240). Baltimore, MD: Brookes.

Darley, F.L., Aronson, A.E., Brown, J.R. (1969). *Differential Diagnostic Patterns of Dysarthria.* Journal of Speech and Hearing Research, 12, 246-269.

De Bodt, M. (2015). *Stemstoornissen*. Antwerpen: Garant.

De Bodt, M., Hernandez-Diaz Huici, M.E., Van De Heyning, P.H. (2002*). Intelligibility as a linear combination of dimensions in dysarthric speech.* Journal of communication disorders, volume 35, issue 3, May-June 2002, 283-292.

Derwing, T. & Munro, M.J. (2001). *What speaking rates do non-native listeners prefer?* Applied Linguistics, 22 (2001), 324-337.

Dharmaperwira-Prins, R. (2005). *Dysartrie en verbale apraxie.* Amsterdam: Pearson Benelux B.V.

Dinnocenzo, J., Tjaden, K. & Greenman, G. (2006*). Intelligibility in dysarthria: Effects of listener familiarity and speaking condition.* Clinical Linguistics & Phonetics, 20-9, 659-675.

Dorfman, J. (1994). *Sublexical components in implicit memory for novel words.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 1108-1125.

Dörnyei, Z. (2005). *The psychology of the language learner.* New York: Routledge.

Duffy, J. (1995) *Motor speech disorders: substrates, differential diagnosis and management.* St. Louis: Elsevier.

Eadie, T.L. & Doyle, P.C. (2002). *Direct Magnitude Estimation and Interval Scaling of Naturalness and Severity in Tracheoesophageal (TE) Speakers.* Journal of Speech, Language, and Hearing Research, Volume 45-6, 1088-1096.

Fairbanks, G. (1960). *Voice and articulation drill book*. New York: Harper & Row, 124-139.

Frearson, B. (1985). *A comparison of the Aids Sentence List and Spontaneous Speech intelligibility scores for Dysarthric Speech.* Australian Journal of Human communication disorders. Volume 13, 1985 Issue 1.

Grande, M., Hussmann, K., Bay, E., Christoph, S., Piefke, M., Willmes, K. et al. (2008). *Basic parameters of spontaneous speech as a sensitive method for measuring change during the course of aphasia.* International Journal of Language & Communication Disorders, 43, 408-426.

Hamann, S., & Squire, L. (1997b*). Intact priming for novel perceptual representations in amnesia.* Journal of Cognitive Neuroscience, 9, 699-713.

Hayes-Harb, R., Smith, B.L., Bent, T., Bradlow, A.R. (2008). *The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts.* New York: Elsevier - Journal of Phonetics

Hodge, M.M. & Gotzke, C.L. (2011). *Minimal pair distinctions and intelligibility in preschool children with and without speech sound disorders.* Clinical Linguistics & Phonetics, 25(10), 853-863.

Hustad, K. (2006). *A Closer look at transcription Intelligibility for Speakers with Dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners.* American Journal of Speech-Language Pathology, 15 (August 2006), 268-277.

Hustad, K. (2007). *Effects of Speech Stimuli and Dysarthria Severity on Intelligibility Scores and Listener Confidence Ratings for Speakers with Cerebral Palsy*. Folia Phoniatrica et Logopaedica 2007, 59, 306-317.

Hustad, K. (2008). *The relationship between listener comprehension and intelligibility scores for speakers with dysarthria.* Journal of Speech, language, and hearing, 51(3), 562-573.

Johansson, L., Möller, s., Olofsson, k., Linder, J., Nordh, E., Blomsted, P., Van Doorn, J. & Karlsson, F. (2014). *Word-level intelligibility after caudal zona incerta stimulation for Parkinson's disease.* Acta Neurologica Scandinavica, 130, 27-3.

Kalf H. & S. van Zundert (2017). *Logopedie bij de ziekte van Parkinson.* Woerden: Nederlandse Vereniging voor Logopedie en Foniatrie.

Kempler, D. & Van Lancker, D. (2002). *Effect of Speech Task on Intelligibility in Dysarthria: A Case Study of Parkinson's Disease.* Brain and Language, 80, 448-464.

Kent, R. D., Miolo, G., & Bloedel, S. (1994). *The intelligibility of children's speech: A review of evaluation procedures*. American Journal of Speech-Language Pathology, 3(2), 81-95.

Khaghaninejad, M. (2018). *Intelligibility of Language Learners to Native Speakers: Evidence from Iranian ESL Learners conversing with Canadians.* International Journal of English language & Translation Studies, 6(1), 93-104.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., Berke, G.S. (1993). *Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research.* Journal of Speech and Hearing Research, 36, 21-40.

Kwiatkowski, J., & Shriberg, L. D. (1992). *Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss.* Journal of Speech, Language, and Hearing Research, 35(5), 1095-1104.

Lagerberg, T.B., Asberg Johnels, J., Hartelius, L., Persson, C. (2015) *Effect of the number of presentations on listener transcriptions and reliability in the assessment of speech intelligibility in children.* International Journal of Language and Communication Disorders, 50(4).

Langley and Sheppeard (1984). *The visual analogue scale: Its use in pain measurement.* Heidelberg: Springer-Verlag GmbH

Likert, R. (1932). *A technique for the measurement of attitudes.* New York: Archives of Psychology, 22, 5-55.

Markham, D., & Hazan, V. (2002). *Speaker intelligibility of adults and children.* Proceedings of the international conference for spoken language processing, 16(20), 1685–1688.

Martin, F. N. (1990). *Hearing and hearing disorders.* In G. H. Shames & E. H. Wiig (Eds.), *Human communication and its disorders*. 3, 350-392. Columbus: Merrill.

McHenry, M. (2011). *An exploration of listener variability in intelligibility judgments.* American Journal of Speech-Language Pathology, 20(2), 119-123.

Moskowitz, H.R. (1977). *Magnitude estimation Notes on what, how, when, and why to use it.* Journal of Food Quality, 3(1977), 195-227.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). *Speech perception as a talker-contingent process.* Psychological Science, 5(1), 42-46.

Pennington, L., & Miller, N. (2007). *Influence of listening conditions and listener characteristics on intelligibility of dysarthric speech.* Clinical Linguistics & Phonetics, 21(5), 393-403.

Sandström, L., Hägglund, P., Johansson, L., Blomstedt, P. & Karlsson, F. (2015). *Speech intelligibility in Parkinson's disease patients with zona incerta deep brain stimulation.* Brain and Behavior, doi: 10.1002/brb3.394,1-10.

Schiavetti, N., Metz, D.E. & Sitler, R.W. (1981). *Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: evidence from a study of the hearing impaired.* Journal of speech and hearing research, 24(3), 441-5.

Sluijmers, J. Zoutenbier, I., Versteegde, L., Singer, I., & Gerrits, E. (2016). *Prevalentie en incidentie van dysartrie en spraakapraxie bij volwassenen.* Rapport voor NVLF van Lectoraat Logopedie Hogeschool Utrecht.

Spitzer, S.M., Liss, J.M., Caviness, J.N., & Adler, C. (2000). *An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech.* Journal of Medical Speech-Language Pathology, 8, 285-293.

Stark, C.E.L., McClelland, J.L. (2000). *Repetition Priming of Words, Pseudowords, and Nonwords.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(4), 945-972.

Stevens, S. S. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects.* New York: John Wiley.

Stipancic, K., Tjaden, K. & Wilding, G. (2016). *Comparison of Intelligibility Measures for Adults With Parkinson's Disease, Adults With Multiple Sclerosis, and Healthy Controls*. Journal of Speech, Language, and Hearing Research, 59, 230-238.

Waters, C. (1994). *Patient observations.* University of Southern California School of Medicine: Movement Disorders Center.

Weismer, G. & Laures, J. (2002). *Direct Magnitude Estimates of Speech Intelligibility in Dysarthria: Effects of a Chosen Standard.* Journal of Speech, Language, and Hearing Research, 45, 421-433.

Tjaden, K. & Wilding, G. (2011). *Effects of speaking task on intelligibility in Parkinson's Disease.* 25(2), 155-168.

Ramig, L.O., Countryman, S., Thompson L.L. & Horii Y. (1995*). Comparison of two forms of intensive speech treatment for Parkinson disease.* Journal of speech and hearing research, 38(6), 1232-51.

Rosenbek, J. C., & LaPointe, L. L. (1978). *The dysarthrias: Description, diagnosis, and treatment.* In D. F. Johns (Ed.), *Clinical management of neurogenic communicative disorders*, 251-310. Boston: Little, Brown, and Company.

Weismer, G. (2006). *Motor Speech Disorders*. San Diego: Plural Publishing Inc.

Whitehill, T. L. (2002). *Assessing intelligibility in speakers with cleft palate: a critical review of the literature.* The Cleft Palate-Craniofacial Journal, 39(1), 50-58.

Yorkston, K.M., Beukelman, D.R. (1978) *A comparison of techniques for measuring intelligibility of dysarthric speech.* Journal of communication disorders, 11, 499-512.

6. Appendix

Appendix 1 Literature framework

Article	Subject (What has been done)	Method	Why	Important factors	Advantages	Disadvantages
1. D. Beukelman, K Yorkston 1979	The dysarthric patient reads a passage and the authors (Beukelman & Yorkston) ranked speech intelligibility on a seven-point equal appearing interval scale. Those samples (a passage, 50 words) are transcribed by 108 speakers. Those transcriptions are rated.	Intelligibility scores were derived by computing the percentage of words correctly transcribed by naive speakers. Yorkston and Beukelman (1978) compared intelligibility scores derived from a variety of quantification methods including estimates, scaling procedures and six objective	Determine the relationship between information transfer and speech intelligibility	- Different types of dysarthria(different etiology) - Interaction between dysarthria severity and intelligibility on sentences vs words - Influence of connected speech (predict a word in context more accurately)	- The results of this study reveal an index of successfulness of communication. - Intelligibility scores are a functional index of communicative performance. - Samples of connected speech are highly correlated with measures of functional communication (such as information transfer). - Big participant group (108).	- An overall index of communicative performance cannot be universally applied for all clinical purposes. - 108 transcriptions done on paper (labour-intensive for transcription raters).

		measurement techniques.				

2. D. Kempler, D. Lancker 2002	This study assessed intelligibility in a dysarthric patient with Parkinson's disease (PD) across five speech production tasks: spontaneous speech, repetition, reading, repeated singing and spontaneous singing.	Recording samples of speech from PD patients in five speech production tasks. Participants had to listen to the production and write down what they heard on the answer sheet provided. In each speech task, correctly transcribed words were counted.	Determine the speech intelligibility of dysarthric patients with Parkinson's diseas e	- Signs of dysarthria vary across speech tasks by Parkinson's disease. - Dysarthria patients show longer vowel durations during reading. - PD patients seem to show a longer voice onset times in reading.- PD patients can improve their speech with techniques which consciously slow them down.	- Speaking loudly, Delayed auditory feedback, slower (paced) speech, reading aloud, singing improve intelligibility. - Repetition, reading, repeated singing and seems significantly more intelligible than spontaneous speech. - 64 listeners	- Study only done with the speech samples of one PD patient. - Some tasks are biased (spontaneous singing f.e. asks for increased respiration, speech volume, articulatory clarity and pitch contour). - Singing and speech may be controlled by different brain mechanisms. - Methods that use repetition and reading to assess intelligibility may overestimate conversational intelligibility in some patients and therefore lead to inadequate understanding of their actual communicative function.

3. G. Weismer, J. Laures 2002	Experiment 1: Determine if speech samples of a person with dysarthria and one with normal neurological history are affected by different standards. Experiment 2: Paired comparison task in which participants chose the most intelligible utterance between utterance pairs.	Using DME as a rating system. Direct magnitude estimation (DME) is a method of perceptual ratio scaling in which an observer makes a numerical estimate of the sensory magnitudes associated with a set of stimuli.	Finding out of the use of a DME rating system is comparable to the use of orthographic transcription rating.	-Each of the multiple variables in the speech signal can have a contribution to speech intelligibility. - The standard created by the control individuals has an influence on the perceptual scaling. - Listeners were students from the speech disorder department.	- Scaling estimates of speech intelligibility are a more complete representation of the deficit than the typical word- or sentence- based percentage estimates. - Due to fixed speech samples, the only variable was the identity of the standard and its possible influence on the perceptual scaling of the fixed set of utterances - Listeners do not have experience with the clinical population. - Both experiments give an identical intelligibility ranking. - A modulus scale is used.	- The number of variables in the speech signal has a high contribution to the intelligibility. - Small participation groups (4 dysarthria patients and 3 control individuals). - Small and in diverse listener groups (3 men, 7 women) (1 man and 9 women). - Listener group has been exposed to dysarthria as a speech disorder. - Listeners were aware of their task. This has a biasing effect on the psychophysical estimates. - Absence of fixed standards is a problem. - There is no guarantee of a match between the position of an utterance on a percentage scale

						and on a psychophysical scale.
4. M. Ganzeboom, M. Bakker, C. Cucchiarini, H. Strik 2016	Investigate intelligibility ratings at three different levels of granularity: utterance, word and subword level.	In a web experiment, 50 speech fragments produced by seven dysarthric speakers were rated by 36 listeners in three ways: a score per utterance on a Visual Analogue and a Likert scale and a score per word and subword level by orthographic transcription.	Exploring the possibility of more detailed evaluations to contribute to the forming of diagnosis and measuring or comparing the outcomes of different types of therapy	- Semantically unpredictable sentences were used.	- 3 x 36 listener ratings. - Three ways of rating: Score per utterance on Visual Analogue Scale, Likert scale and orthographic transcription. - The implemented phoneme scoring method proved feasible, reliable and provided a more sensitive and informative measure of intelligibility. - Use of isolated and pseudowords (effect of context can be minimized) - The distinction between measures of accentedness and intelligibility. - Online experiment with	- Use of isolated and pseudowords gives an unnatural context. - it is not clear if human scale ratings are valid indicators of intelligibility. - Non-diverse patient group (male, hypokinetic dysarthria caused by PD). - Non-diverse listener group (8 male and 28 female) - 5 listeners had experience with dysarthric speech.

					randomized ordered questionnaires. - Focuses on the possible implantation of the results.	
5. J. Dinnocenzo, K. Tjaden, G. Greenman 2006	Examine the effects of familiarization and speaking condition on sentence intelligibility for a speaker with dysarthria secondary to a traumatic brain injury (TBI)	A total of 120 listeners orthographically transcribed sentences produced by the speaker with dysarthria in a habitual, slow, fast, or loud condition after receiving no familiarization, paragraph familiarization, or word list familiarization.	The benefit of familiarization to dysarthria is evidenced by higher intelligibility scores for familiarized listeners as compared to unfamiliarised listeners, with some studies reporting as much as a 15% to 20% improvement in intelligibility for familiarized listeners (Liss et al., 2002).	- Familiarization procedures include: comprised of a random ordering of the words in the paragraph. - A magnitude production paradigm was used to elicit the variation in rate or intensity. - Dysarthria is blinded controlled by professionals during the study.	- Familiarization on intelligibility in dysarthria is well-documented. - The subject was screened on cognition - Listener group of 120 randomly selected individuals.	- Small participant group (3 SLP's) to determine the type of dysarthria. - Rather in diverse listener group (77 females and 43 males). - 120 orthographically transcribed sentences on paper (labour-intensive for transcription raters). - Between-group design does not rule out listener variables as an explanation for improved intelligibility. - No studies have directly compared

						the separate effects of familiarization, rate manipulation and increased loudness on intelligibility in dysarthria.
6. K. Hustad 2006	This study addressed the effects of 3 different paradigms for scoring orthographic transcriptions of dysarthric speech on intelligibility scores. The study also examined whether there were differences in transcription accuracy among words from different linguistic classes.	Speech samples were collected from 12 speakers with dysarthria of varying severity. Twelve different listeners made orthographic transcriptions of each speaker, for a total of 144 listeners. Transcriptions were scored using 3 different paradigms: total word phonemic match, informational word phonemic match and informational word semantic match.	Finding out if used paradigms and linguistic classes have an influence on scoring orthographic transcriptions.	- Randomly selected listeners. - The study was experimental (scripted and not spontaneous)	- Addressing of 3 different paradigms (total word phonemic match, informational word phonemic match and informational word semantic match). - Examination of differences among words from different linguistic classes. - 144 listeners (randomly assigned to each speaker). - No possibility of a learning effect. - Listeners in this study were naïve.	- Non-diverse speaker group (12 adults with dysarthria secondary to cerebral palsy). - Non-diverse listener group (23 men, 121 women). - The findings of the study provide no information regarding speaker production performance. - 144 orthographic transcriptions done on paper (labour-intensive for transcription raters.)

		Transcriptions were also coded into 3 linguistic categories: content words, modifiers and functors. The number of words that each listener transcribed correctly within each category was tallied.				
7. K. Hustad 2007	This study examined differences among transcription intelligibility scores and listener confidence ratings for three different types of speech stimuli – single words, unrelated sentences and	Speakers produced stimulus material which was then played for listeners who made orthographic transcriptions of what they heard and made ratings of their confidence in what they had	Insight if listeners expectations match up with their intelligibility scores. (Can listeners correctly predict if they clearly understand a person with dysarthric speech?)	- The present study uses the data of Hustad 2006.	- 144 listeners. - Use of Likert scale to determine confidence.	- A small group of speakers (Twelve). - The study was experimental (scripted and not spontaneous), generalizing to real-life speaking is limited. - 144 orthographic transcriptions done on paper (labour-intensive for transcription raters.)

	sentences forming a narrative – all produced by speakers with dysarthria.	transcribed for each utterance.				- Non-diverse listener group (23 men, 121 women).
8. K. Hustad 2008	This study examined the relationship between listener comprehension and intelligibility scores for speakers with mild, moderate, severe and profound dysarthria	Speech samples were collected from 12 speakers with dysarthria secondary to cerebral palsy. For each speaker, 12 different listeners completed two tasks (for a total of 144 listeners), one task involved making orthographic transcriptions and one task involved answering comprehension questions. Transcriptions were scored for	Finding out if there is a relationship between listener comprehension and given intelligibility scores.	- The present study uses the data of Hustad 2006.	- Results of the study were consistent with other studies.	- Non-diverse speaker group (12 adults with dysarthria secondary to cerebral palsy). - Non-diverse listener group (23 men, 121 women). - No significant relationship between intelligibility scores and comprehension scores when severity effects were removed. - Listeners' goals influence the two tasks. - Short term memory issues could have played a part in the intelligibility tasks.

		the number of words transcribed correctly; comprehension questions were scored on a 3-point scale according to their accuracy.				- The study was experimental. - There were no language
9. K. Stipancic, K. Tjaden, G. Wilding 2016	This study obtains judgments of sentence intelligibility using orthographic transcription for comparison with previously reported intelligibility judgments obtained using a visual analog scale (VAS).	Same procedures as Tjaden, Sussman and Willing (2014). Speakers read Harvard sentences in habitual clear, loud, and slow conditions, Sentence stimuli were equated for peak intensity and mixed with multi-talker babble. A total of 50 listeners orthographically transcribed sentences.	Compare orthographic transcription with VAS.	- Multitalker bable is added to the sentences.	- Same procedures as Tjaden, Sussman and Willing (2014). - Use of Harvard sentences. - High listeners standards (native speakers, educated, no experience and no history of speech, language or hearing problems). - Use of multi-talker babble for the ecologically valid environment. - 78 speakers.	- 50 listeners in the same age group (18 to 30 years) - Orthographic transcription is time-consuming for listeners and scorers. - Scaling tasks are subjective measures. - Intelligibility of dysarthria in background noise is not highly investigated.

		Procedures were identical to those for a VAS reported in Tjaden, Sussman and Wilding (2014).				
10. K. Tjaden, G. Wilding 2011	The purpose of this study was to compare intelligibility estimates obtained for a reading passage and an extemporaneous monologue produced by 12 speakers with Parkinson's disease(PD). These utterances were rated by 70 listeners using orthographic transcriptions and direct magnitude estimation.	Speakers were audio recorded while reading a paragraph and producing a monologue. Speech samples were separated into individual utterances for presentation to 70 listeners who judged intelligibility using orthographic transcription and direct magnitude estimation (DME).	Compare intelligibility estimates for reading passages and extemporaneous monologues.		- 12 speakers (6 men, 6 women). - Two different intelligibility measures were obtained for the reading task. - Familiarity effects were minimised by randomizing.	- All the speakers have the same background (Parkinson's disease) - Non-divers listener group (16 males, 44 females). - Small absolute differences in scaled intelligibility have a large impact on the relative ranking of an individual. - Intelligibility tests may not represent how easily extemporaneous speech is understood. - Orthographic transcription is time-consuming for

						listeners and scorers.
11. B. Bender, M. Cannito, T. Murry, G. Woodson 2004	This study compared speech intelligibility in nondisabled speakers and speakers with adductor spasmodic dysphonia (ADSD) before and after botulinum toxin (Botox) injection.	Speech samples from 10 speakers with severe ADSD prior to and following Botox injection and 10 matched healthy adults. Thirty phrases were extracted from the speech samples and arranged in a counterbalanced listening experiment. Listener response was scored for words correctly identified using a liberal scoring criterion yielding a percentage of	Does botox injections improve intelligibility in ADSD		• 10 speakers (9 women, 1 man). • Speakers were carefully selected by criteria. • Ratings were tested by Wilcox signed ranks test • Speech intelligibility is not equally affected by all speakers with severe ADSD.	- All the speakers have the same background (ADSD) - The study has a counterbalanced experimental design to equally distribute the effects of potential confounds to intelligibility such as phrasal predictiveness and speaker-listener familiarity. - The files were randomized. - The effect of Botox is temporary and thus making the 'after' files less objective.

		words correctly identified for each speaker.				
12. L. Sandström, P. Hägglund, L. Johansson, P. Blomstedt, F. Karlsson, 2015	This study investigates the effects of Levodopa (L-dopa) and deep brain stimulation in caudal zona incerta (cZi-DBS).	utterances were extracted from 11 patients with PD preoperatively (off and on L-dopa medication) and 6 and 12 months post bilateral cZi-DBS operation (off and on stimulation, with simultaneous L-dopa medication). Intelligibility was assessed through a transcription task performed by 41 listeners in a randomized and blinded procedure.	What are the effects of L-dopa and cZi-DBS on intelligibility on spontaneous speech?	Study based on Johansson 2014	• 41 listeners (quite divers) • Background noise was added to the recordings. • Stimuli were randomized. • The use of spontaneous speech samples added a higher degree of confidence in ecological validity.	• Spontaneous speech samples introduce several additional sources of information to the listener. • Presence of individual differences in treatment. • Disease progression has an influence.

13. L. Johansson,S. Möller, K. Olofsson, J. Linder, E. Nordh, P. Blomstedt, J. van Doorn, F. Karlsson, 2014	This study investigates the effect of caudal zona incerta-deep brain stimulation (cZi-DBS) on word-level speech intelligibility in patients with Parkinson's disease, under both an optimal listening condition and a simulated more naturalistic listening condition.	Spoken single words were extracted from reading samples collected from 10 bilaterally implanted patients with PD pre- and post-cZi-DBS. Intelligibility was assessed through a transcription task performed by 32 naive listeners under two listening conditions: (i) with low-amplitude conversational speech added as background and (ii) with no added background noise. The listener's responses were scored in terms of agreement	What are the effects of L-dopa and cZi-DBS on intelligibility on word-level?	• Listeners were primed with the stimulus to be attentive to transcribing the target word.	• The task with and without noise	• Relatively small patient group (8 men, 2 women) • 32 listeners (rather a small group)

		with the intended words.				
14. Mohammad Saber Khaghaninejad, 2018	investigate the extent to which phonological characteristics of Farsi interfere with Iranian ESL learners‟ intelligibility when they interact with Canadian native English speakers.	Iranian ESL learners underwent an interview and were asked to read twenty paired sentences that contained missing vowels and consonants in Farsi and ten sentences including consonant clusters aloud while being tape recorded. Then, Canadian native speakers were invited to listen to the tape and declare their degrees of perception accordingly.	Investigate the pronunciation of second language learners	• Focuses on sounds not present in Farsi and present in Canadian English	• Focus on language learners	• Small group (5 Farsi Iranian, 5 Canadians)

Appendix 2 Basic studies

This appendix shows the creation process of the 'ideal experiments' and can be seen as brainstorming. Elements from this appendix have been used for the shaping of table 6 (3. The ideal intelligibility experiment) and table 8 (4. Conclusion and discussion). This appendix shows a short and first draft of the experiment designs and then will give a draft of the experiment questionnaires.

During the writing process of this thesis, it turned out that a number of components (including the use of professionals) did not fit in well with the subject. These components can still be observed in these first drafts but were eliminated from the thesis during the process. This means that not all elements of this appendix can be seen in the main body of the thesis.

Experiment designs

Patient studies:

Yellow experiment (transcription experiment)

- Design experiment:

In this experiment, listeners have to make a transcription of different sample sets that are offered by headphones. These sample sets consist of different utterances produced by dysarthria patients (each with a different pathology). The transcription is part of a questionnaire.

Used material: Samples of utterances, questionnaires

Red experiment (rating experiment)

- Design experiment:

In this experiment professionals and (non-professional) listeners have to rate different sample sets consisting of different utterances produced by dysarthria patients (each with a different pathology) on intelligibility.

The professionals have to assess the samples and create an ICF profile per patient, following the ICF guidelines. Hopefully, this will create a large, accurate database about the content of the speech of the patients that is trustworthy.

The non-professional listeners have to rate the samples on intelligibility according to the Visual Analogue Scale (VAS). The rating is part of a questionnaire.

Used material: Samples of utterances, ICF models/guide, questionnaires

Language learner studies:

Green experiment (minimal pair experiment)

- Design experiment:

In this experiment, listeners have to make a transcription of different sample sets that are offered by headphones. These sample sets consist of different utterances produced by language learners (from different cultural backgrounds) and consist of sentences with minimal pairs in which different phonological features are used. The transcription is part of a questionnaire.

Used material: Samples of utterances (with minimal pairs in it), questionnaires

Blue experiment (lacking vowels and consonants experiment)

- Design experiment:

In this experiment, listeners have to rate different sample sets consisting of different utterances produced by language learners. The samples are created by the elicited speech method, speakers will be asked to read aloud a set of *English* sentences consisting of sounds that are missing in their mother tongue. This demonstrates the likelihood of pronunciation errors.

The listeners have to rate the samples on intelligibility according to the Visual Analogue Scale (VAS). The rating is part of a questionnaire.

Experiment questionnaires:

Patient studies:

Experiment
The listeners get different samples of utterances from different pathology patients. The samples consist of sets of different sentences. The participants have to access the samples and make a transcription of what they heard.
Questionnaire design
1. What does the sample say? (Transcribe the sample as good as possible.) Sample 1:..... Sample 2:..... etc. 2. Is there a difference in the samples? 3. If yes, what do you think the difference is?
Experiment
The listeners get different samples of utterances from different pathology patients. These samples are assessed by a professional and are marked on content factors following the ICF guidelines. The non-professional participants have to access the samples and rate the samples on intelligibility. This measurement will happen on a Visual Analogue Scale (VAS)

Questionnaire design

Professional:

The professional is asked to create an ICF-profile regarding the samples per pathology patient. This ICF-profile will focus primarily on speech (so will use the codes for body functions (b) and body structures (s).

Non-professionals:

1. How would you rate sample [x] on a scale from 0-10?

   0=poor 10=excellent

1	2	3	4	5	6	7	8	9	10

Repeat this question for multiple sets.

2. Make a top [x] of the samples, with 1 being excellent intelligibility and [x] being not intelligible at all.

Repeat this question for multiple sets.

3. Why did you put sample [x] on [x] 'Not intelligible at al'?

4. What made sample [x] the most intelligible?

Learner studies

Experiment

The listeners get different samples of utterances from different language learners. In these samples minimal pairs with different phonological features are used. The participants have to access the samples and make a transcription of what they heard.

Questionnaire design

2. What does the sample say? (Transcribe the sample as good as possible.)

   Sample 1:.....

Sample 2:.....

etc.

2. Is sample [x] produced by a native speaker?

3. If no, where does the speaker come from?

Experiment

The listeners get different samples of utterances from different language learners. These samples are created by the elicited speech method, speakers were asked to read aloud a set of twenty English sentences to demonstrate the likelihood of pronunciation errors as the language of the speaker is missing consonants and vowels that are existing in English.

The participants have to access the samples and rate the samples on intelligibility. This measurement will happen on a Visual Analogue Scale (VAS)

Questionnaire design

2. How would you rate sample [x] on a scale from 0-10?

   0=poor 10=excellent

1	2	3	4	5	6	7	8	9	10

Repeat this question for multiple sets.

2. Make a top [x] of the samples, with 1 being excellent intelligibility and [x] being not intelligible at all.

Repeat this question for multiple sets.

3. Why did you put sample [x] on [x] ‘Not intelligible at al’?

4. What made sample [x] the most intelligible?

5. What is the native tongue of the speaker from sample [x]?

Appendix 3 Grandfather passage

Grandfather Passage

You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.