Martin Alexander Gulbrandsen

The Development and Pilot of a Norwegian Nonword Repetition Test Following the Framework of Language Impairment Testing in a Multilingual Setting

Master thesis

LOGO345, Master's Programme in Health Sciences - Logopedics

at

UNIVERSITY OF BERGEN

DEPARTMENT OF BIOLOGICAL AND MEDICAL PSYCHOLOGY

FACULTY OF PSYCHOLOGY

SPRING 2021

Preface

This thesis marks the completion of my master's programme in Logopedics at the University of Bergen. When deciding on what to work on as my thesis, I realized that I wanted to specifically create something. Jan de Jong presented me with the opportunity to work with Nonword repetition, focusing on developmental language disorder among both monolingual and bilingual children. To me, this was the perfect thesis. It allowed me to utilize my linguistic background, and to contribute to future work within a clinical field. It has provided me with greater insight into how DLD works, and the problems bilingual children face when acquiring a new language, and how this can further problematize being provided help when needed. To my knowledge, a project such as this, following the same principles, has not been undertaken in a Norwegian setting, which has made it a complicated and challenging process. Nevertheless, it has been extremely rewarding. It is filled with pride and care, with the hope that it can sincerely contribute to future work within this field.

My most sincere thanks to my supervisor, Jan de Jong. He has provided me with support and feedback at even the most difficult of times and has always been more than happy to discuss and reflect. I also would like to thank Aleksander Morland, who has been a great help on most things' phonology. Thank you to Kamila Polišenská for taking the time to discuss the creation of this test with me.

Thank you to everyone who took time out of their busy schedule to partake in this study, both to those who helped gauge how "Norwegian" the items were, and those who participated in the NWR test itself.

Bergen, May 2021

Martin Alexander Gulbrandsen

Contents

Sammendrag

Hovedformålet med denne masteroppgaven var å utvikle og pilotere en norsk nonord repetisjonsoppgave (NWR), som baserer sin teoretiske bakgrunn og design på rammeverket som ble laget for Language Impairment Testing in a Multilingual Setting (LITMUS). Et delmål var å inkludere en adaptert versjon av en allerede etablert NWR oppgave, for å sammenligne resultatene. Den teoretiske bakgrunnen presenterer relevant forskning, samt faktorene som er brukt i utviklingen av NWR oppgaven. Dette lager grunnlaget for både begrunnelsen og metoden bak masteroppgaven. Metoden presenterer de to NWR oppgavene, samt en detaljert forklaring av de tidligere nevnte faktorene for å knytte de opp mot utviklingen av NWR oppgavene. Metoden presenterer også utførelsen av de to oppgavene, samt etiske hensyn. Aspekter av reliabilitet og validitet av prosjektet er representert.

Nøkkelord: DLD, nonord repetisjon, LITMUS, flerspråklighet, språktilegnelse

Abstract

The main purpose of this thesis was to develop and pilot a Norwegian nonword repetition (NWR) test, basing the theoretical background and construction design on the framework created for Language Impairment Testing in a Multilingual Setting (LITMUS). A secondary goal was to include an adapted version of an already established NWR test, to compare results. The theoretical background presents relevant research findings, as well as the factors used in the development of the Norwegian NWR test. This creates the basis for both the purpose and method of this thesis. The method presents the two NWR tests, further expounding on the aforementioned factors, connecting them to the development of the tests. The method also presents the conduction of the tests and ethical considerations. Aspects of reliability and validity of the thesis are presented.

Keywords: DLD, nonword repetition, LITMUS, bilingualism, language acquisition

Introduction

Developmental language disorder as a clinical challenge

The primary goal of a speech therapist is to help an individual with language problems in remedying or mediating their situation. To begin this process, it is necessary to ascertain whether the individual indeed has difficulties with language, what kind of difficulties, and to what degree. A speech therapist working with children is typically concerned with evaluating the language development of a child, but the tools at their disposal are limited. Some of these tools rely on established literacy skills, target language vocabulary, or cognitive abilities which might not be fully mastered until the child has already enrolled in school. This lack of variation in diagnostic tools for children at a young age is counter-intuitive, as research shows that early intervention is more effective (Schwarz & Nippold, 2002), and the impact of language problems on the early social life of a person is severe (Janus et al., 2017). Not only do the early school years lay the foundation for reading and writing abilities, but they also lay the groundwork for mathematics and logical thinking. These years also provide children with the opportunity of learning aspects of social interaction, and as they grow, they create an image of their identity and self. The longer it takes for a child to receive the help they need, the further behind their peers they end up, which not only affects scholastic development, but also their social skills and sense of belonging. For some individuals, these language problems stem from a known biomedical condition, autism spectrum disorder, or intellectual disability. For others, a deficit in language ability cannot be attributed to any clear cause, such as low non-verbal intelligence, hearing loss, or neurological damage (Leonard, 2014). These latter problems were termed developmental language disorder (DLD) in a consensus study called the CATALISE Consortium (Bishop et al., 2017), and although they have been called many different things in the past, this is the term used in this thesis. The lack of a clear cause makes diagnosing DLD a difficult clinical challenge, further complicated by the heterogeneity and possibility of comorbidity in children with DLD (Bishop, 2004). DLD *can* impact many aspects of language, and which aspects, and the degree to which they are impacted, vary from individual to individual (Van Weerdenburg et al., 2006). Attempts to define subtypes of DLD have not been successful (Conti-Ramsden & Botting, 1999). Currently, DLD refers to difficulties with a wide range of language areas, from phonology, morphology, semantics, and syntax, to pragmatics, discourse, and verbal memory. It is typical in clinical practice to differentiate between children with language production difficulties and children with both language production and comprehension difficulties (Bjerkan, 2000).

DLD and bilingualism

Another issue in screening for DLD is how to aptly screen bilingual children. Research has found that bilingual children with typical development (TD), who have recently started learning their second language, showcase similar linguistic manifestations in their second language as monolingual children with DLD do in their first (Paradis et al., 2003). This means that if a test is normed on monolingual children with TD, the likelihood of bilingual children being scored as having DLD is high if they are tested during the first years of exposure to their second language (Meir et al., 2015). Both bilingual children with TD and children with DLD often show limited activity in language learning situations, perhaps based on a fear of being incorrect, and often have difficulties with expressing themselves or understanding linguistic information (Egeberg, 2016). Many tools rely on a developed vocabulary in the target language, and without this, some areas of language are impossible to assess. Indeed, how can a bilingual child be expected to perform at the same level as a monolingual child on a task such as a vocabulary test, like the British Picture Vocabulary Scale (BPVS-II; a translated and adapted version is often used for Norwegian) if they have yet to have even heard the Norwegian word for an item enough times in their life to store it. It is difficult for a speech therapist to ascertain whether this is caused by a deprivation of stimulus, or if the child struggles with language acquisition. There is also the issue of conveying instructions, as many bilingual children who have recently started learning the target language can have difficulties understanding them, without this being related to DLD. This inherent reliance on target language vocabulary makes assessing the language development of a bilingual child difficult, and it is difficult to determine whether perceived language problems stem from the naturally occurring limited input of target stimulus, or if they are caused by deficient learning mechanisms (Kohnert, 2010; Paradis, 2010). Not only is the child already learning one language at home, but they will also typically have less exposure to the target language than their monolingual peers (Deanda, 2015). The similarity between how bilingual children with TD and children with DLD behave in language settings often leads to misdiagnosis (Paradis et al., 2011). Salameh (2002) found that bilingual children with DLD are often identified later than monolingual children and were more likely to be diagnosed with severe language impairment. There is, however, no reason why bilinguals should be at an increased risk of having DLD, as bilingualism has not been shown to increase the effects of DLD (Paradis et al., 2011). The problem must therefore lie in assessment.

Nonword repetition

A test which can aid in the screening of both monolingual and bilingual children with or without developed literacy skills is nonword repetition (NWR). NWR does not rely on specific target language vocabulary, nor literacy skills, but is rather an elicitation test focused on assessing the areas of phonology and verbal memory. Research has found that NWR deficits correlate with literacy difficulties, despite it not being a literacy task, which may point to the common underlying problem being a limitation in phonological short-term memory (Snowling et al., 2000). It is a process-dependent task which is not specifically reliant on vocabulary and metalinguistic awareness but serves as a simulation of language acquisition. It is seen as an effective predictor for language learning abilities (Gathercole 1995, 2006;), as it closely matches the phonological aspect of word learning, and correlates to the level of the user's phonological working memory (Coady & Evans, 2008). Findings show that NWR had the second highest sensitivity (78%) and a high specificity (87%) when compared to other psycholinguistic markers for identifying DLD in English monolingual children, such as sentence repetition, third singular morpheme, and production of the past tense (Conti-Ramsden et al., 2001). It is therefore possible that poor NWR performance can be a clinical marker of DLD, and only NWR and sentence repetition accurately identified a history of DLD. A history of DLD in this case refers to participants who had mild or resolved language difficulties at the time of testing. The other psycholinguistic markers could only gauge the participants' current language abilities (Conti-Ramsden et al., 2001). This is of importance, as a true marker of DLD should be able to identify a history of language impairments, not only current language abilities (Conti-Ramsden et al., 2001).

The creation of a language-specific NWR test

The purpose of this thesis is to create a language-specific (LS) NWR test, and to test items with monolingual TD children. This NWR test follows the framework made by the COST Action IS0804 (European Cooperation in Science and Technology) for the Language Impairment Testing in a Multilingual Setting (LITMUS) battery. Following this framework allows for parallels to be drawn to other LS NWR tests doing the same, and it provides well researched evidence for the impact different factors may have on an LS NWR. This pilot also employs a crosslinguistic (CL) test created for the COST Action IS0804, which serves as another comparison for the LS NWR test.

Development rationale

There are three main reasons for creating an LS NWR test for Norwegian. Firstly, it becomes a tool specific to the target language, utilizing a range of representative consonants and vowels, and ensuring that they are phonotactically legal. Creating it to be included in LITMUS means that LITMUS will be further bolstered, allowing for testing of bilingual children in Norwegian as well as their first language (if there is an NWR for that language). Secondarily, it is hypothesized that it can be used in conjunction with an NWR which is not LS, such as the CL used here, in aiding in differentiating between bilingual children with DLD, and bilingual children who simply have yet to sufficiently interact with the second language. Thirdly, it allows for the investigation of factors which have proved to have an impact on NWR performance in other languages, expanding our theoretical knowledge on how these different factors impact performance on the target language, in this case Norwegian.

Factors considered in the creation of an NWR test.

The framework highlights the impact the following factors have on the difficulty and language likeness of an NWR. Manipulating these ensures that an NWR test is sufficiently difficult, to avoid floor- and ceiling effects, and it allows control over the influence that a target language has on the NWR test. This section merely aims to explain the different factors the framework deems as the most important when creating an NWR test, whilst the CL and LS sections in the method relates the applicable factors to the two different NRW tests.

Length

The length of a nonword (NW), or more aptly the number of syllables, directly impacts performance. This has been proven in previous studies, as shorter nonwords tax phonological short-term memory to a lesser degree, correlating to a higher accuracy among participants (Stokes et al., 2006, Weismer et al., 2000). This is also found for digit span and immediate serial recall tasks which are both traditionally associated with phonological memory (Archibald & Gathercole, 2006). It is important to note that these other tests utilize the lexicon to a much higher degree, and digit span does not control for number of syllables per se, but rather the number of digits provided at once. It is therefore thought that NWR can provide a purer assessment of phonological storage quality, because lexically based reconstruction processes are unable to compensate for deficits in basic phonological storage when using NWs (Gathercole, 2006). Initially, the similarity between these tests was seen as evidence for NWR being a pure measure of short-term memory capacity. This notion was quickly revised, as it became evident that the prosodic and segmental structure of items had a significant impact on performance, and so nonword repetition became a test of phonological abilities rather than pure memory (Chiat, 2015). As Chiat (2015) also points out, one could argue that the two are inseparable, as memory capacity is reliant on the efficiency and efficacy of processing information stored in the memory. Likewise, the ability to process material stored in memory is reliant on the information staying in memory for long enough to complete the process. The difficulties children with DLD experience in producing NWs is now therefore explained as an impairment to their phonological working memory (Weismer et al., 2000), also referred to as the phonological loop (Gathercole & Baddeley, 1990). Usually, newer NWR tests consist of items with 2-5 or 2-4 syllables. The number of syllables depends on which other factors the NWR test controls for. This is also impacted by how common longer words are in the target language. Norwegian is a good example of this. In Norwegian, most roots are only 2 syllables long (Kristoffersen, 2000). An NW is therefore not particularly "Norwegian" if it is much longer than 2 syllables. It would still be a non-word, but not a particularly "Norwegian" non-word. Monosyllabic items are also generally not included, as they often produce ceiling effects, and the evidence garnered from disyllabic items is strong enough to allow for the omission of monosyllabic items in most NWR tests (Graf Estes et al., 2007).

Segmental complexity

Segmental complexity refers to the complexity of the representations of a given phonological system and the complexity of the sequencing in which they appear. The complexity of how a phonological system is represented refers specifically to the relationship between orthographic and phonemic realizations and is therefore not relevant in this thesis. An example of this is how *back* would be less complex if it were written as *bak*. The complexity of sequencing, however, is pertinent to NWR. An example of this is how clusters are more complex than non-clusters, particularly if the two combined consonants are articulated at different places and in a different manner. The complexity of the segmental sequencing of NWs in an NWR test has recently emerged as a key factor, where it has been shown that children perform consistently better for items containing only single consonants compared to items containing clusters (Archibald & Gathercole, 2006; Jones et al., 2010). This effect is expected to be dependent on familiarity with complex syllabic structures, such as clusters, but to which extent this familiarity affects results is not known. Depending on what purpose an NWR test is created for, clusters are either included, or omitted. If an NWR test is created to be language-specific, then it usually includes clusters when appropriate for the target language, but if it is created to be more generalised then it might not include clusters. This is also dependent on if the developer deems segmental complexity as crucial for measuring the phonological loop. Note that segmental complexity has only *recently* emerged as a key factor, and so controlling for it is quite a recent development (Chiat, 2015). Some recent NWR studies do not include clusters in their LS tests, even if clusters are typical for the target language. The French LS test created in the study by Thordardottir & Brandeker (2013) does not contain clusters, but this is because the items in that study were not designed to be either wordlike or non-wordlike, and their results showed that simple consonant-vowel (CV) NWs were preferable to more complex NWs when testing bilingual children, because they were less affected by bilingual exposure. The inclusion of clusters in a test which aims to test bilingual children can therefore increase the divide in results between bilinguals and monolinguals. This study employs the CL NWR test and an LS NWR test specifically to test the validity of the LS test-set, but also to investigate this divide in the future, when bilingual participants would be involved. Controlling for segmental complexity in the form of clusters increases the difficulty of the test, and according to the framework, segmental complexity is one of the main deciders on whether the test-set should include NWs up to 5 syllables or not (Chiat, 2015). This factor is controlled separately from the next factor, prosody.

Prosody

Prosody refers to the elements of speech which are properties of syllables or larger units, such as words, and include amongst other features, stress, and intonation (Cutler, 1998). They are usually referred to as suprasegmentals and are only properties of spoken language. Stress plays the part of making a syllable prominent, and is signalled with, among others, increased duration, loudness, and pitch. Stress can be controlled for by producing test items following the typical stress patterns of disyllabic and polysyllabic words in the target language (Chiat, 2015). Prosodic characteristics have shown to influence the repetition of words and NWs (Sahlén et al., 1999; Yuzawa & Saito, 2006). Norwegian is primarily a trochaic language, and words which are not loan words almost exclusively assign primary stress to the left-most foot in a word (Kristoffersen, 2000). As has been observed in other languages with trochaic bias, children tend to omit the weak syllables in an NW if they appear in a pre-stressed position. Using an example from the LS test; Norwegian children are more likely to replicate the whole NW if it is produced with initial stress [‘bʉlpɪ] instead of final stress [bʉl’pɪ] (Carter & Gerken, 2003; Gerken, 1994; Reuterskiöld-Wagner et al., 2005). A Swedish NWR study found that children with DLD were 6 times more likely to omit weak syllables when they were presented in a pre-stressed position (Sahlén et al., 1999). Carter et al. (2002) also found that children had a higher proficiency in repeating NWs including primary stress on the first syllable, and they were more likely to omit subsequent syllables instead of adding more. A 3-syllable NW would be more likely to be produced as a 2-syllable NW, rather than a 4-syllable NW. According to the framework, an LS test which also controls for segmental complexity should control for prosodic structure by ascribing typical primary stress to two of the eight NWs at each length, and atypical primary stress to another two of the eight NWs at each length (Chiat, 2015). The framework posits that prosody should be controlled separately from segmental complexity, as controlling for both would mean the inclusion of an item with atypical stress for each unique cluster position at each length. The total number of items would therefore increase from 24 to 36. NWs which control for segmental complexity should primarily be ascribed primary stress for LS tests based on a trochaic language, as this is the most reminiscent of the target language. Intonation is particularly interesting in Norwegian, where most native speakers belong to tonal accents. Intonation in this case refers to pitch movement, and different Norwegian accents belong to different tone dialects, which differ in what is seen as typical and atypical intonation. This will be brought up again in the section on the LS NWR construction in the method. Intonation is not something specifically pointed out in the COST framework, rather being a part of the *prosody* umbrella.

Phonotactics and phonotactic probability

Phonotactics refers to the possible sequential arrangements phonological units can have in a language and their position (Crystal, 1994). An example which is applicable to both English and Norwegian would be how /spr-/ can occur initially, but /spm-/ would be an "illegal" arrangement and is therefore not allowed in the target language. Traditionally, phonotactics has been regarded in the dichotomous terms of legal and illegal, but the sounds which can be found in the legal category of a language certainly do not occur with the same frequency. The frequency of phonological segments and sequences of those segments occurring in a language is referred to as phonotactic probability (Jusczyk et al., 1994). An NW which has high phonotactic probability consists of segments which are more likely to occur in the target language. An individual's performance on the replication of such a NW will therefore be affected by the user's language experience and knowledge, as they will have interacted with these segments frequently. Phonotactic probability does not take into account the prosodic and syllabic structure of a word and does not differentiate between two phonemes forming an initial cluster of a stressed or unstressed syllable or a sequence of a consonant coda and consonant onset crossing the syllable boundary (Chiat, 2015). An example of this would be how phonotactic probability does not differentiate between how frequently /p/ and /l/ occur together at the onset of a word and how frequently they occur medially between two syllables. [‘pleɪ̯s] *place*, and [‘hɛlp.lɪs] *helpless* both have the same frequency of this sequence, but the first example shows this sequence as an initial cluster, whilst the latter example shows it as occurring between syllables. Performing analysis of phonotactic probability requires access to a corpus which can provide insight into how often these elements occur in sequence. A true analysis of phonotactic probability is impossible without a corpus such as this, but Chiat's framework does recommend an alternative analysis to approximate phonotactic probability, namely wordlikeness.

Wordlikeness

Wordlikeness is the subjective judgement by a native speaker of the extent to which an NW resembles a real word (Chiat, 2015). These ratings are expected to be influenced by the inherent lexical properties of the NWs, such as length, prosody, and frequency of phoneme sequences (Frisch et al., 2000). It is also affected by the inclusion of real target language morphemes and can therefore be manipulated by their inclusion (Casalini et al., 2007). It is therefore an alternative to phonotactic probability, when necessary, but it does not directly gauge objective lexical properties.

Phonemes, phones, and allophones

The aforementioned factors are important in the creation of an NWR test but creating any sort of non-literacy test also includes the selection and controlling of the sounds included in the test. The smallest unit of language which can distinguish meaning is called a phoneme (Abrahamsen & Morland, 2012; Hayes, 2009), and when realised it is called a phone/speech sound (Fintoft, 1982). A phone does not alter the meaning of a morpheme (Hayes, 2009), while a phoneme does, and this distinction is important. Two phones which belong to the same target phoneme are referred to as allophones of the same phoneme. Two speakers which are producing the same word can in many situations produce dissimilar phones to one another, but as long as the phones belong to the same target phonemes, the meaning does not change. A good example in Norwegian is how the /r/ in the word <rot> *root* can be produced with an alveolar tap as [ɾuːt] or with a voiced uvular fricative as [ʁuːt]. The allophone chosen to represent the phoneme can vary based on dialect. If a phone which represents another target phoneme is chosen, however, the meaning of the word changes. For example, if the /r/ is replaced by /m/, the word becomes [muːt] <mot> *courage*. These are called minimal pairs, and although the now two distinct words only have one segmental difference, they do not carry the same meaning (Dodd et al., 2008). This means that a participant's choice of allophonic representation of a phoneme does not impact their results on an NWR test, as it aims to measure their ability to store and replicate the correct target phoneme. To summarize, phonemes are analysed, whilst allophones are disregarded.

The consonants and vowels relevant to Norwegian and the LS NWR test can be found in Appendix D-G (ordering of appendices is based on the article)**,** represented by symbols found in the International Phonetic Alphabet (IPA; International Phonetic Association, 2015) and are based on the overview in Morland (2012), and Kristoffersen (2000).

In this thesis, the orthographic (written) version of a nonword is presented with angle brackets < >, and the phonemic transcription is presented with slashes / /. The recorded NWs are presented through narrow phonetic transcriptions, which provide more phonetical information. Such transcriptions are presented with square brackets [ ]. Should the NWR test created for this pilot be carried on and used in the field, it would be too time-consuming and unnecessary to perform narrow phonetic transcriptions. For example, the difference between /pyltə/ and [pʰyl̪tə] is irrelevant, as reduplication of stimuli on the phonemical level is what NWR gauges. There are some cases when performing qualitative analysis where a narrow phonemic transcription is done to showcase a phonological phenomenon, but for most of the examples in this thesis, a phonetical transcription would have been sufficient.

Goals, research questions, and hypotheses

Aiming to create and pilot a NWR test in a Norwegian environment, this thesis bases itself upon the work done for the COST Action IS0804, and more specifically, utilizing the CL NWR test set in combination with a language-specific set (Chiat & Polišenská, 2016). The main goal of the thesis is to create a language-specific test which takes account of Norwegian phonology, aiming to explore whether the test can be used in Norwegian speech therapy in the future. It should be noted that this thesis aims to be a part of the LITMUS battery, and it follows the COST test's rationale. The creation has been informally aided through e-mail correspondence with COST researchers, which has provided input on how to use the factors in a Norwegian context. Further down the road, this work could be used in conjunction with larger tests for DLD screening in children, particularly bilingual children, and it could also be used to screen for dyslexia in adults, as these groups are particularly prone to experiencing difficulties with NWR (Snowling 1981, Snowling et al. 1986). The hypothesis for this thesis is that the monolingual TD children included in the pilot administration of the test will perform well overall, but that their performance will be impacted by the factors that were considered in the creation of the test, namely length, prosody, segmental complexity, and wordlikeness. This means that it is expected that they will perform better on the CL test (which, for example lacks clusters) as compared to the LS test (which is expected to show the influence of the non-length factors). The LS test is influenced by Norwegian, but it includes and manipulates factors which increase the difficulty of the test, whilst the CL controls only for length, and has a simple phonological inventory. The research questions are as follows:

- To what degree can the internal reliability of the LS NWR test be ascertained at this preliminary stage of LS NWR test development?
- How do the results indicate that the participants perform on the CL test in relation to the LS test?
- Can a correlation between results on the CL test and the LS test be established to support the hypothesis that the LS test can measure language skills?
- Will the different factors utilized to construct the LS NWR test have an impact on performance, and to what degree?
- Can unique error types inform the continued development of the LS NWR test?

Method

The crosslinguistic NWR test

As previously mentioned, this pilot utilizes the CL test, created through the COST Action IS0804 project. This test is designed to test language processing while relying minimally on language-specific knowledge. The main goal of this test is therefore to not put children with limited knowledge of a specific language at a disadvantage, while still testing their language processing skills. This test used to be called a “quasi-universal nonword repetition test” (Chiat, 2015), but the “quasi-universal” term was quickly abandoned in favour of “crosslinguistic”, as it was deemed to be a more appropriate and accessible name. After all, no NWR test can be truly universal, as an NW will always carry language-specific properties that makes it more accessible to speakers of one language than others (Chiat, 2015). Nevertheless, this type of NWR test presents a unique opportunity to compare results on a set of almost identical nonwords across languages, and the results can also be compared with outcomes on an LS NWR test.

Controlled factors

The test controls for length by equally dividing the 16 items into 2-5 syllables. It contains 5-syllables because of the lack of segmental complexity. The phonological inventory is limited, including the consonants /p, b, t, d, k, ɡ, s, z, l, m, n/ and the vowels /a, i, u/. These are presented in simple CVCV structures and is purposely limited to accommodate for word phonology in a large variation of languages. Utilizing simple CVCV structures ensures that the test-set is applicable whether the language allows for more complex segmental contrasts and syllable structures, or if it mostly consists of these simple structures. Simple CVCV structures means that the test does not include clusters, and does therefore not control *for* segmental complexity, rather ensuring that the (low) segmental complexity is the same regardless of the target language. The set includes alternatives for all 16 items to avoid segments containing real words or real word inflections in the particular language the test is used in. This, however, is paradoxical. On one hand the set was made to compare results on almost identical NWs across languages, whilst on the other, it wishes to provide alternatives that can avoid real words in the target language. It is therefore impossible for the set to ensure that it does not contain any real morphemes in any language, as the differences between the alternatives are minimal. Number 9 (see Appendix A) is one of many examples, as the alternatives provide variations on voicing, but each one still includes /li/ which means *hillside* in Norwegian. This was also the case when the set was tested in England, as most of the component syllables (/du/, /mi/, /mu/, /si/, /su/, and /tu/) are real monosyllabic English words (*do, me, moo, see, sue,* and *too*; Chiat & Polišenská, 2016). Prosody is controlled but not controlled for, producing syllables with even length and pitch. The final syllable is lengthened to mark the end of the utterance (Chiat, 2015). This prosody is deemed as the most neutral prosody achievable, although any prosodic patterns, even a neutral one, will be more typical for some languages than others. The CL NWR test does not control for phonotactic probability or wordlikeness, as it is not meant to be influenced by a target language.

Selection of alternatives

The chosen alternatives can be found in Appendix B. When selecting alternatives, the focus was on having a wide phonological representation, and selecting items with the least number of syllables containing real Norwegian words. An alternative containing two or more syllables which constitute a real Norwegian word were excluded entirely. Examples of this are /mʉ:li/ *possible* and /li:ta/ *small*, which can be found in several of the alternatives. Any alternative including the phoneme /z/ was also excluded, as /z/ is not a part of Norwegian phonology. Every consonant presented in the list of alternatives is represented in the selected items, disregarding /z/. This is also true for the vowels, albeit there are only three different vowels in total. The phonemic realization of the phonological inventory consists of the most common Norwegian realisations of the target phonemes [p, b, t, d, k, ɡ, s, l, m, n, ɑ, i, ʉ].

The Norwegian language-specific NWR

The need for an LS NWR test

The creation of this test follows the framework for creating an LS NWR test (Chiat, 2015), ensuring that it is in tune with recent research, and for it to be a well-founded contribution to the LITMUS test battery. It includes 24 items influenced by Norwegian (See Appendix C). Creating an LS NWR test allows for the manipulation of language-specific properties, allowing for investigation of the impact of different properties on the NWR proficiency of an individual with DLD. When an NWR is influenced by a language the test can contain complex and specific phonology, prosody, and segments. Tests and items containing more complex phonology are more informative about skills and deficits (Chiat, 2015). This also means that, in theory, it should be a more difficult test than the CL test.

Phonological inventory and their representation

The LS NWR test aims to have a varied phonological inventory, whilst minimizing regional advantages or disadvantages. Appendix D shows all the consonants which appear in the Norwegian language, and Appendix E shows which consonants were included in the test-set. Similarly, Appendix F & G show vowels used in Norwegian and which ones were included in the LS NWR test. Following is a detailed breakdown of the inclusions and exclusions, based on the ecological validity for a “universal” Norwegian LS test-set. The main goal in selecting phonemes is to use phonemes which are as “region-free”, as possible. Ensuring that the test is made up of phonemes which any Norwegian speaker recognises makes it so that the test can hopefully be used anywhere on any Norwegian speaker. As can be seen in Appendix D & E, no retroflex consonants were included in the final test-set. This is not only because they only appear in certain dialects, but also because they usually replace a cluster in some dialects. The focus when creating this test set was to include as many different typical phonemes as possible, whilst excluding those which only appear in certain dialects. Inclusion and exclusion, in this sense, means that they are either produced or not produced by the researcher in the test-set, but participants are free to produce an allophone to a target phoneme. For example, the NW <pylte> is pronounced as [‘pʏltə] in the video the participants watched, but they are free to pronounce it as [‘pʏɭʈə]. Creating the test in this way ensures that it is maximally flexible for use in Norway, regardless of dialect, and it can be re-recorded with applicable allophones or tonal accent if necessary. It is important to note that this section compares orthographic realisations to phonemical realisations. This is done to relate different dialects to each other through a shared constant (being one of the two written forms of Norwegian, Bokmål). Utilizing a shared constant to compare different dialects is simply done for clarity. An orthographic cluster is not the same as a spoken cluster. An NWR test does not rely on orthographic constructions, but orthographic clusters which are pronounced as a cluster in some dialects and not others are purposely excluded, to ensure that clusters do not appear if the test is re-recorded with allophonic variations to the target phonemes later. For example, the retroflex plosive [ʈ] often occurs in East-Norwegian dialects when /t/ follows /r/, as in the word <bart> *moustache*, which is then pronounced as [bɑʈ]. In some dialects this word is pronounced with a final cluster (Kristoffersen, 2000). Including an item which includes /rt/ orthographically in the test set could therefore result in a situation where the item includes a cluster in some spoken variations of the test, and the target phoneme(s) would have to be changed to still control for segmental complexity.

/r/ allophones

It is important to note that these overviews (Appendix D-G) show all the realistic phones found in Norwegian. Several of these are, as mentioned in the introduction, allophones. Allophones of the same phoneme are different phonetic realisations which are interchangeable without affecting the meaning of the word. <riste> *shake* in East-Norwegian is usually pronounced as [ɾɪstə], whereas someone from Bergen with a West-Norwegian dialect might pronounce the word as [ʁɪstə]. /r, ɾ, ʀ, χ, ɹ, ɣ, ɻ, ɰ/ all have similar functions in Norwegian, and are allophones of the same phoneme /r/, but their usage depends on dialect. Some of these are called positional variants, which means that they only occur in specific environments. /χ/, for example, is an unvoiced /r/ allophone in some dialects, which is usually used initially, finally, or in front of or after unvoiced phones. [ɾ] is the only included /r/ variation in this LS test, which is either voiced or unvoiced depending on its position and phonetic environment. It is always voiced in the LS NWR test, except from in the NW <bigapeler>, as it occurs in the final position. The test-set is constructed in such a way that it is accommodating for allophonic variations of /r/.

/ʃ/ and /ç/

To differentiate between words starting with or including <skj_> and <kj_>, it has been "traditional" to use either the unvoiced alveolar fricative /ʃ/ or the unvoiced palatal fricative /ç/. The word <skjorte> *shirt* is "supposed" to be pronounced with a /ʃ/, and the word <kjole> *dress* is "supposed" to be pronounced with a /ç/. However, as languages change over time, the /ç/ has become unstable, and is being slowly replaced by /ʃ/ (Papazian, 1994; Van Dommelen, 2003). Particularly younger speakers in larger cities such as Bergen, Stavanger, and Oslo often replace /ç/ with /ʃ/. It is also one of the last phones a child learns, and so it was decided that it would not be represented in the LS test (Papazian, 1994; Preus, 1982; Vanvik, 1979). /ʃ/ on the other hand is represented (although only once). The unvoiced palatal affricative /c͡ç/ is also an allophone of /ç/, which can be found in several west-Norwegian dialects. The restrictiveness of this sound excludes it from the LS test.

/l/ allophones.

All /l/ allophones, [l, ɽ, ɭ, ʎ, ɭ] are only represented as /l/ in the LST. This is because they, like many other phones, are dependent on dialect and positional variation. /l/ is the most common pronunciation, and it is also the only permissible production of /l/ when it is in the word initial position in a CV_ structure. The phone is represented frequently in the LST and should be easily recognized no matter which dialect a participant is used to. /ɭ/ is a lesser used realization of the orthographic <_rl>, where someone with the appropriate dialect might pronounce <farlig> *dangerous* as [fɑɭi], whereas others retain the orthographic cluster in their phonetical representation. Considering how often <_rl> is replaced by /ɭ/ in the appropriate dialects, there are no situations where this combination occurs in the test set. These orthographic considerations are made so that there are no misunderstandings in the future related to re-recording the test set with focus on a specific dialect. If /ɭ/ was included in this LS test, the orthographic realisation of the item would include <_rl>. This would then be troublesome in dialects where this is realised as a cluster. Excluding this phoneme ensures that the properties of the item does not change if adapted to a specific dialect.

Retroflexes

Most retroflex consonants encountered in Norwegian are used to replace a specific consonant cluster in some dialects, such as the voiced and unvoiced retroflex plosives [ʈ] and [ɖ], which replace the written <_rt> and <_rd>, respectively. They have been omitted as they are not used in most west-Norwegian dialects, and so the rationale for exclusion is the same as the rationale for the exclusion of [ɭ]. [ʈ] can also occur in a <_lt> structure, like for example in <gult> *yellow*, which is pronounced as /gʉːʈ/ in several Norwegian dialects. The NW <pylte> does indeed have an <_lt> structure, but the retroflex [ʈ] is phonotactically illegal in this structure.

Glottal stop

The glottal stop [ʔ] does appear in Norwegian, but it only appears before word-initial vowels to provide emphasis. For particularly emphatic speech, it can be inserted medially in stressed syllables which begin with a vowel (Kristoffersen, 2000). The glottal stop is not included in the LST, as it would enforce the stress-pattern of the NW, which is counteractive to the LS test flexibility in dialectal variation. The syllable structure of the NW items is, barring the inclusion of clusters, CVCV(C), and so there is no position in which this phone can occur.

Palatal phones

The voiced and unvoiced palatal plosives /c/ and /ɟ/ are allophones of /t/ and /d/, respectively, and only occur in mid- and north-Norwegian dialects after short vowels. The restrictiveness of these sounds excludes them from inclusion in the LST.

/ɱ/ is a rare allophone of /m/, which only occurs in front of other labiodental sounds. As seen in Appendix D, the only other labiodental Norwegian phones are /f/ and /ʋ/. Neither /f/ nor /ʋ/ appear after /m/ in the test set either, because of wordlikeness considerations, and so /ɱ/ cannot occur. /ɲ/ is used in the same dialects which use the palatal plosives /c/ and /ɟ/, and it appears as an allophone of /n/. This phone is not included on the same basis that those palatal plosives were excluded.

Vowels

All Norwegian vowels are represented in the test set, barring 4. [a] is not included, in favour of [ɑ], as this is a much more common representation of /a/. [a] occurs in some South Norwegian and South East Norwegian dialects but is not common enough to warrant inclusion. Those who are well versed in Norwegian might question the exclusion of /ø, œ, æ/ and they would be right to do so. These phones are very frequent, and there is no inherent reason for them not being included. This is simply an oversight, although I would posit that this does not have a tangible impact, as the test has a varied phonological inventory without them. It is possible that the inclusion of these phonemes could have made it easier to attain sufficient wordlikeness ratings.

Controlling factors

The LS test controls for length by dividing the 24 items equally into the lengths of 2-4 syllables. Monosyllabic items are not included, as real words in Norwegian are not typically monosyllabic, and they in general rarely add the same value to the test as disyllabic items do (Graf Estes et al., 2007). They are also not included in either the CL test or the English LS test (Chiat, 2015; Chiat, Polišenská & Szewczyk, 2012). 5-syllable items are also not included as they typically create a floor effect for young children (Chiat, 2015), and 5-syllable Norwegian words are rare, even when including derivational and inflectional affixes. They are also rarely included when controlling for segmental complexity, which this test does. Segmental complexity is controlled for by including 2 items at each length with an initial cluster, and 2 items with medial clusters. This leaves 2 items without clusters and typical stress, and 2 items without clusters and atypical stress. Final position clusters are not included, not because of a low frequency of occurrence, but rather because it is not believed that they provide information that initial and medial clusters do not. The framework also explicitly states that segmental complexity should be controlled for with initial and medial clusters (Chiat, 2015). Segmental complexity is properly controlled for with an equal split of items containing clusters and without clusters. Prosody in the form of stress is controlled for by including 2 items with atypical stress at each syllable length. This ensures that 2 items with typical stress and no clusters and 2 items with atypical stress and no clusters can be compared at each length. Typical stress refers to, as previously mentioned, primary stress on the initial syllable. Prosody in the form of tonal accents is controlled, as the proceeding section on tonal accents will explain in-depth. The final factor this test controls for is wordlikeness. Participants were needed to gather wordlikeness ratings, and that procedure and the results are presented in the section on wordlikeness, which comes after the section on tonal accents.

Tonal accents

Most Norwegian dialects have distinct tonal accents (Kristoffersen, 2000). There are some which do not, for example a dialect around Bergen (note, not the city-dialect of Bergen, which has tonal accents), a small area north of Trondheim, and many dialects in the two most northern counties. The lack of tonal accent in the northernmost counties is often attributed to sub-stratal influences from the Finn-Ugric Sami language also spoken there (Kristoffersen, 2000). These tonal accents are by most divided into two main groups: East and West Norwegian. East Norwegian is comprised of the dialects of Eastern Norway and the central parts of Southern Norway, and West Norwegian is the rest. These two groups are referred to as low tone dialects and high tone dialects, respectively. Dialects belonging to these two groups produce words with either accent 1 or accent 2, but these two terms mean different things depending on the group. In low tone dialects, accent 1 is characterized by the fundamental frequency (f0), commonly referred to as pitch (pitch being purely a perceptual concept), starting with a low value on the stressed syllable and rising at the unstressed, post stress syllable. Accent 1 in high tone dialects is characterized by a high f0 value at the stressed syllable, and a decline towards the unstressed, post stress syllable. No matter which group is referred to, accent 1 is characterized by either a steady increase or a steady decrease, with the start anchored to the stressed syllable. Accent 2 has similar initial and final f0, increasing or decreasing in-between the two syllables. In low tone dialects, accent 2 is characterized by an initial high f0 value, which drops to a low f0 value before rising to a high final f0 value. In high tone dialects, this tonal contour is the complete opposite, beginning and ending with a low f0 value, but rising in the middle.

Accent 1 in low tone dialects: Low to high f0 value.

Accent 1 in high tone dialects: High to low f0 value.

Accent 2 in low tone dialects: High to low to high f0 value.

Accent 2 in high tone dialects: Low to high to low f0 value.

Tonal accents are an important characteristic of Norwegian phonology, but it is not feasible to control for it in an NWR test without either ignoring other, more pressing factors, or subjecting participants to an incredibly lengthy NWR test. Almost all the items in the LS NWR test are produced with low tone accent 2, except [hu'dɛl] and [tə'sʉːl], which are produced with low tone accent 1 because of their structure, as Norwegian words with final syllable stress can only be produced with accent 1. The participants in this pilot are all from areas of Norway with low tone dialect, so it was decided that presenting the LS test-set with a tonal accent most salient to them would be for the best, when controlling for low tone and high tone variations was deemed as ecologically invalid. Considering that the test set would not have to be changed to switch from low tone to high tone, recording the test-set with a high tone dialect in the future to end up with two variations would probably be more valid. This would only come to pass if the test-set ends up being trialled further.

Wordlikeness

Wordlikeness is the last controlled factor, but the subjective nature of it made it unpredictable. This factor was the catalyst for several iterations of alternatives, and the changing of parameters to properly control for it. According to the framework, controlling for wordlikeness means to include 1 item of high wordlikeness and 1 item of low wordlikeness in each category (See Appendix H). In this study, high and low wordlikeness refers to the mean score being higher or lower than the Likert scale median, which is 3. Two items at each length control for the same factors, and these are what is referred to as categories. Item 1 [ʻnʉ:sə] and 2 [ʻtʉ:mɪ] are both of the same length, include typical stress, and do not include clusters. Similarly, item 3 [ʻskɑ:nɑ] and 4 [ʻkly:nu] are the same length, include typical stress, and include initial clusters. Items 1 and 2 constitute one category, and items 3 and 4 constitute another. Alternatives were made for each category, where alternatives that were expected to be rated as wordlike included phoneme sequences that were deemed to be highly frequent. Alternatives which were expected to be rated as non-wordlike include less frequent but still phonotactically legal segments. The wordlikeness of the test items were rated by participants on a typical Likert scale, and the test items were randomized so as not to allow for the participants to recognize a pattern in the nonwords they were rating. There were 6 participants in total from the counties of Telemark, Oslo, and Viken. Participants were met online and asked to rate each NW on if they sounded Norwegian. The scale went from 1 to 5, where 1 was strongly disagree, and 5 was strongly agree. They were told in advance that it would not be helpful to attempt to provide answers in a manner that they think the researcher would prefer, but rather to give their honest opinion. The alternatives were not pre-recorded, rather spoken in real time, and the participants were free to hear each alternative as many times as they deemed necessary.

To ensure that every category was filled, there had to be 3 sessions of wordlikeness ratings. Alternatives were fashioned beforehand, to limit the number of sessions, but sometimes none of the alternatives attained means which were appropriate for the category. The next section provides information on how each group of items were created (group being items of the same length), and how some factors were manipulated to attain satisfactory wordlikeness ratings.

The process of creating NWs for the LS NWR test

This section covers how the previous considerations were used to create the LS test-items. Initially, the alternatives for all the test items included no real affixes, derivational or inflectional, but after two rounds of wordlikeness ratings it was deemed necessary to include some affixes to attain high wordlikeness ratings, when applicable (See Appendix I). Real Norwegian roots are almost never longer than 2 syllables without the inclusion of an affix, and this effect became apparent when attempting to fill the high wordlikeness requirement for longer items (Kristoffersen, 2000).

2 syllable items

The 2-syllable items were by far the easiest to create, as most free morphemes in Norwegian are disyllabic (Kristoffersen, 2000). This meant that there was no need to include bound morphemes to manipulate wordlikeness, and a varied phonemic inventory could be the focus. When creating the 2-syllable items, there was an initial bias to including NWs which had a _CVC structure. _CVC is used to show that the word includes phonological information which occurs prior to the consonant-vowel-consonant. An example of this is [hu'dɛl], because of an issue with defining what "typical" and "atypical" stress patterns were. Typically, when a Norwegian simplex word (without affixes) ends on a heavy syllable weight (_CVC; or long vowel), the final syllable is assigned primary stress (Kristoffersen, 2000). It was therefore initially thought that an NW of a CVCVC structure was not atypical if it contained final stress, because although final stress is more atypical, it would not be atypical for that structure. It was eventually decided that what is and is not typical regarding stress should be defined by pattern-frequencies in the language at different syllable lengths, not by the structure of the NW. In Norwegian, real words which are not loan words typically have initial primary stress. Originally, the atypical items selected for this test were a part of the initial wordlikeness rating. After 2 of the rating sessions the subject of what is typical and atypical were revisited, and the test settled on using these items. The atypical category in the 2-syllable part of the set went through several iterations because of this uncertainty in deciding on how to define typical and atypical, struggling to attain satisfactory wordlikeness ratings. When presenting wordlikeness-participants with NWs with atypical stress of a CVCV structure, they usually rated them as having low wordlikeness. There is no doubt that it would be possible to create atypical stress NWs with a CVCV structure, but none of the 8 alternatives ever passed the requirement. The 2-syllable items do not contain /r/, as it was initially decided that this phoneme was not to be included at all. As mentioned previously, there are many /r/ allophones in different Norwegian dialects, and the aim of creating this set was for it to be "universal" for Norwegian. It was eventually decided that it would be better to include /r/, and rather record several versions of the set using other allophones in the future if necessary, as it is one of only two liquids in Norwegian, the other being /l/. 5 of the 8 NWs at the 2-syllable length include /l/, and this is not because of an inherent bias in the creation. Liquids are very common in Norwegian (Kristoffersen, 2000), and it is hypothesised based on wordlikeness feedback that without them many of the alternative items did not have a sonority hierarchy that the participants deemed as particularly wordlike. It was therefore decided that /r/ should be included for the longer NWs, to aid in the manipulation of wordlikeness, although it is used sparingly.

3 syllable items

At the 3-syllable length, more patterns start to emerge. 6 of the 8 items end on a final /ə/, which is in many cases an inflectional affix, which changes a noun into an infinitive verb, but it does not have to, and in many cases it does not. Adding -e to hopp *a jump* (as in a place to jump from or a command to jump) creates hoppe *to jump*, but nouns such as kake *cake*, hake *chin*, rake *rake* are roots which include a final -e without it being an affix. The rationale is therefore that frequent phonemes which also happen to occasionally be morphemes should not be counted as a morpheme. It is irrelevant if an item includes a syllable which is only a real word orthographically, as the participants only hear the NWs. The English LS NWR test (Chiat, Polišenská & Szewczyk, 2012), for example, includes "tow" in <smishertow>, but the NW is pronounced as ['smɪʃə,tɑʊ]. A syllable being a real word is also very difficult to omit, where the English LS NWR includes 5 phonological sequences which could be regarded as real words in English. An example of this is /tɒs/ *toss* in /tɒskəlimə/ (Chiat & Polišenská, 2016). 1-syllable real words are very common in Norwegian (Kristoffersen, 2000), and it was decided that a varied phonemic inventory is more important than a potential overlap between a typical syllable in a longer test item and an actual real word. This stays consistent with the considerations made for the CL test. The inclusion of syllables constituting real Norwegian words is even more apparent in the CL test, as all the syllables are produced with even prosody, whereas real Norwegian words can to a larger degree be obfuscated by tone and stress in the LS test. <dipasse> [ʻdɪp.ɑsə] is an example where the prosody ensures that it does not contain a 2-syllable morpheme. <passe> *to fit* can indeed be pronounced as [ʻpɑsə], but the way that this NW is pronounced, there is a clear divide and pause between [ʻdɪp] and [ɑsə]. Participants were quick to point out constituents that were real Norwegian words, but this did not happen for this item, where two participants stated that it was a "weird" word. It should be noted, however, that the NW <befaning> does include a derivational prefix, be-, which was meant to be reserved for the 4-syllable items, and it also includes the inflectional suffix -ing. This is probably why it attained a mean average of 5. The stress pattern is also typical for its structure, and it might therefore be a little too similar to a real word.

4 syllable items

This syllable length was by far the most difficult category to fill. As has already been established, simplex Norwegian words are usually only 2-syllable or rarely 3-syllable long (Kristoffersen, 2000). Compound words and complex words (root + affix(es)) are however very common. There does not seem to be a good way to create compound NWs, because that intent would never come across to participants, unless previously presented NWs were used. It is not possible to present a 4-syllable nonword and ensure that the participants understand that it is a compound word, as they have no frame of reference for the constituents. This would have been easier if all the NWs were free to use as many clusters in different positions as needed, as compound words often contain at least 1 medial cluster at the conjunction between the roots. It was therefore decided that the focus should be on creating complex item alternatives with derivational and inflectional affixes for the items which were supposed to be of high wordlikeness. This is allowed in the framework, if the target language typically inflects words according to their grammatical category (Chiat, 2015). All the NWs of high wordlikeness include a derivational prefix (Appendix C), except for <dråttelige>, which was the only 4-syllable NW that attained high wordlikeness before the introduction of derivational prefixes. It does however contain two inflectional affixes, -lig which creates an adjective, and -e which denotes plurality. -e denotes plurality here because it follows -lig, making this connection clearer than the examples brought up regarding the 2-syllable items. Of the NWs that include a derivational prefix, <bigapeler> scored lowest, which might be attributed to how rare the bi- prefix is. The English

LS test also includes three derivational morphemes in /ˈsɑnəri/, /stɒfəli/, and /rɪˈvaɪk/, but it does not control derivational affixes in the way this LS test does. Derivational affixes were therefore specifically included to manipulate the wordlikeness of the words. This does increase the benefit of specific vocabulary.

Test presentation

NWs from both test-sets were randomized and presented to participants in the form of an animated video (See Appendix K for still images). This video, animated by Astrid Ellensdatter Mork-Knutsen, is a simple animation aiming to capture the participants' attention, without providing unnecessary stimulus. Polišenská and Kapalková (2014) also used an animated story in their NWR study, in which a bead necklace broke, and every time the participants repeated a NW the beads were put back onto the necklace one after another. Wishing to replicate this tool, we created an animation which portrays a wizard attempting to bring the stars back to the skies after they mysteriously disappeared. This allows the participants to see their progress visualised and to feel a sense of achievement after each stimulus item. The video has been engineered in a way as not to distract, but rather sharpen focus towards an easily visualised goal. For example, visual noise as simple as animating the wizard's mouth was circumvented by hiding it in his beard. Care was also put into making sure that there is enough of a pause between the appearance of new stars, so the video could be used in research on other languages if so wished. The start of NWs is spaced apart by 8 seconds, allowing the participants to prepare when seeing the wizard before he speaks, listening when he speaks, and replicate before a new star starts appearing. If another star appears before they have had time to replicate, they might feel as though they had no impact on the test. This issue would not have appeared if the test had been presented through a PowerPoint, as the tester could control when to continue the test. However, it does ensure that all participants have the same amount of time to replicate. The NWs are pre-recorded, so that there is no change in stimulus between participants.

There are definitive advantages and disadvantages to presenting the NWs in the form of a video:

The most obvious advantage is the consistency of the stimuli. Pre-recorded NWs that are presented at a set interval ensures that participants have an identical amount of time to interact with the NWs, and that the input the children receive is the exact same. The latter could also be accomplished in other ways, such as presenting the test sets with a PowerPoint presentation, as referenced earlier (Polišenská & Kapalková, 2014). Presenting the stimuli as a PowerPoint presentation does not however provide consistency in the timing between NWs, but it does ensure greater control for the researcher. It is possible to make it so a Powerpoint presentation continues at a certain speed, which at that point would be identical to presenting the test as a video.

Presenting in the form of a video limits the researcher's ability to pause or proceed after ensuring participant response. On one hand this can be beneficial in ensuring that all participants have the time they personally need to respond, but on the other hand participants can choose not to respond to stimuli if they so wish. This test is also structured as a call and response, and having the test set continue without a response can alleviate potential stress a participant might illicit in "having" to provide a response.

To summarize the choice of using a video; consistency of stimulus and response, combined with a more visually stimulating test environment was valued higher than the need for a natural pace. Presenting the NWs in this manner is an attempt to expound on the work done by Polišenská and Kapalková, (2014) creating a test which the participants find engaging.

Procedure

To test the validity of the test-set, the pilot was only performed on monolingual children with typical development (TD). There will not be any data showing how well children with DLD perform on the test. This also means that it will not be possible to gauge the accuracy and sensitivity of the test. Nevertheless, this work is meant to serve as the groundwork for future work within this field in Norway, and to contribute to LITMUS.

This pilot included 13 children (8 female, 5 male) between the ages of 5 to 7, the inclusion criterion was that they were either in their last year of kindergarten or their first year of school. All participants grew up surrounded by people who spoke in a low tone tonal accent. Participants were recruited through direct contact with primary schools and kindergartens, who reached out to parents on the pilot's behalf, as well as online postings.

Participants were not subject to screening of nonverbal IQ, socioeconomic status, or other instruments which might indicate that they were not TD. All parents reported that their children were typically developing and without any auditive or cognitive difficulties which might impact their performance, and that their language development was typical. One parent reported that their child had just recently started mastering [ɾ], and this was considered when scoring. Participants met individually with the researcher online together with their parents, and sound was recorded (on both sides of the interview when possible). After a short introduction, a couple of priming NWs (/bɪɾtə/ and /slɪɾp/) were presented to the participants, which they were told to reproduce. The video began, once understanding of the procedure was ascertained.

There were originally 15 participants (9 female, 6 male), but 2 participants did not feel comfortable when met online. In these cases, parents were sent the video with explicit instructions on how to administer it at home, but the participants were ultimately unwilling.

Scoring

The COST Action IS0804 allows for two levels of scoring. Whole-item scoring and segment-scoring. Whole-item scoring is used in the CNRep, the Preschool Repetition Test, and the CL NWR test (Seeff-Gabriel et al., 2008; Chiat & Polišenská, 2016). Segment-scoring is used in the CL NWR (Boerma et al., 2015), and some newer experimental CNRep versions. For this pilot it was seen as sufficient to only perform whole-item scoring, as it is clinically more appropriate, and is informative enough to provide the answers this pilot seeks (Roy & Chiat, 2004; Boerma et al., 2015). This rationale is reflected in the CL NWR test experiment by Chiat & Polišenská, (2016). Participant responses are also qualitatively analysed to gauge the validity of the LS test items, but this analysis does not impact their overall score. There is a divide between those who score additions as correct or incorrect, as some argue that additions do not equate to lost phonological information (Boerma et al., 2015; Dollaghan & Campbell, 1998). Others point to the evidence that children with DLD include significantly more additions in relation to their TD peers (Sorenson Duncan & Paradis, 2016; Wagner et al., 1999). This pilot scores additions as incorrect, which is in line with the framework (Chiat & Polišenská, 2016). Whole items are correct if they contain all segments in the target item in the correct order and only those segments

(Chiat, 2015). A segment is correct if it falls within the target segmental category, as outlined earlier. It is not incorrect to produce /ɭ/ instead of /l/, as they are both allophones of the same target phoneme. Phonetic distortion does not constitute as an error either, on the same principle, because it does not change the meaning of the production. For example, if a person produces the /s/ in /sɑ/ *said* with a lateral lisp as [ɬsɑ] it would not be the same as producing [lɑ] *laid*. Although [ɬs] might cause confusion as it is a deviant allophone of /s/, it still belongs to the target phoneme category, and should be treated as such. Allowances are made if the child consistently produces a segmental substitution which is consistent for their production if it is characteristic for immature speech.

Ethical aspects

In this pilot, 6 participants took part in measuring wordlikeness, and 13 participants took part in the NWR tests. As a scientist, it is imperative to follow guidelines on how to conduct research, particularly when the latter participant group were children. Care was taken to ensure that the guidelines in the Helsinki declaration regarding vulnerable groups was followed during the entire process of this study (Førde, 2014, 10.10). Although the study does not have any diagnostic implications, it does process personal data in the form of audio recordings (only for the NWR participants), and so an approval application was sent to Regional Committees for Medical and Health Research Ethics (REK) before the collection of data. The study was approved by REK (See Appendix P). An application was also sent to NSD. NSD stated that as it did not require extended Data Protection Impact Assessment (DPIA), it was not necessary for it to be reviewed by NSD. An important key point in ensuring good ethics is informed consent, which was reviewed and signed by all participants (Appendix M & N). As the NWR participants were children, they do not have consent capacity, and so consent was provided by their parent(s) or guardian(s). A particularly prudent point to get across in the informed consent form for the NWR test was that results would not be used as diagnostic measures. Although the children do not have consent capacity, it was important to ensure that they understood what partaking in the study involved and that they themselves would be interested in participating. Not only is this important from an ethical standpoint, but it is also imperative for the pilot results, as unwillingness to participate would greatly impact their performance. For this purpose, an age-appropriate information sheet was created (Appendix O), particularly focusing on the fact that they were just as free to withdraw from the study as they were to participate. Understanding of what the test entailed was further cemented in the introduction prior to the testing-phase, and two potential participants did indeed withdraw at that point. All results from the study were anonymized, with full confidentiality between the researcher and participants for privacy reasons. All audio recordings were stored and analysed in a safe research-space at the University of Bergen. As participation was not to result in any diagnostic or therapeutic value, there was no need to ensure alternative intervention. No financial payment or benefit was provided to participants, to further ensure voluntary participation. The testing-phase was only 6 minutes long and was therefore not very time consuming. The testing-phase continued whether participants responded to the stimulus, ensuring that they had further agency during this phase.

Validity

The degree to which inferences in a study are close to the truth in reality, is referred to as validity (Polit & Beck, 2012). It is not absolute, but rather a reflection of the appropriateness with which inferences are drawn in a given context. All aspects of validity pertaining not just to the construction of the test in itself, but also the validity of the research design and method are important to investigate. It is, however, prudent to point out that the main goal of this thesis is the development of the Norwegian NWR test itself, whilst piloting it was a sub-goal to gauge the feasibility of conducting the test and whether the assumptions the test was created on was present in a real testing scenario. The inclusion of the CL NWR test was also a part of this sub-goal, as it allows for investigation of the correlation in performance on the two tests.

Statistical conclusion validity

Statistical conclusion validity is the degree to which statistical methods used to support inferences on the relationship between independent and dependent variables are true statistical conclusions (Polit & Beck, 2012). There are many threats to statistical conclusion validity, but here the focus will be on three particularly important aspects and their threats, as outlined by Polit & Beck (2012). ***Statistical power*** is the ability to detect true relationships among variables and is tied to sample size. When the sample size is low, the statistical power tends to be low, and analyses can fail to show a true relationship between independent and dependent variables. This is a shortcoming in this pilot (N= 13), owing to the limitations of a masters' thesis and the current pandemic. Independent variables were, however, separated and differentiated to strongly define them. ***Range*** refers to the control of extraneous variations through homogeneity. In the case of this pilot, it would refer to the age of participants. Limiting the age of participants ensures that the results cannot be generalized past this range. This is often a necessary evil in small preliminary pilots such as this, as a certain degree of homogeneity is necessary to extrapolate results. This is also why only monolingual TD children were included. ***Intervention fidelity*** refers to the degree to which the implementation of the intervention is faithful to its original plan. In this pilot, a standard protocol was followed in all interaction with participants, to ensure constancy of conditions to the degree which this is possible. Participants were in their own homes, and therefore the degree of outside distractions could not be controlled. A manipulation check was performed prior to the testing-phase to ensure that they understood their role.

Internal validity

Internal validity measures the validity of an inference which states that it is indeed the independent variable that is causing an outcome. In the LS NWR test, difficulty of correct repetition is the independent variable, meaning that performance is the dependent variable. It is not only made up of linguistic complexity (in the form of segmental complexity), but also length, prosody, and wordlikeness. Segmental complexity, length, prosody, and wordlikeness were all controlled. However, there are other factors which impact this variable, namely memory capacity, developmental stage, cognitive functioning, and the psycholinguistic properties of the target phonemes. Memory capacity was not controlled. Developmental stage was controlled by including participants of a small age-range, but further differentiation could not be performed, as all participants barring 3 were 6-7 years of age. Cognitive functioning was not controlled. To improve the internal validity, tests of memory capacity and cognitive functioning could have been performed. This is a clear limitation of this study, owing to time limitations. Inclusion and exclusion criteria stated that participants were to be TD, but this was not ascertained through testing, rather the subjective opinion of parents.

Construct validity

Construct validity is the degree to which operationalized indicators measure what they are supposed to (Polit & Beck, 2012). It is therefore, similarly to external validity (brought up later), related to generalization. Simply put, the construct the LS NWR test attempts to measure is DLD. As has been established in the theoretical background, DLD is characterised by a deficiency in the phonological loop of the working memory and the short-term memory. Establishing construct validity therefore means to analyse the data provided by participants to support this theoretical structure. Important to note, this pilot does not include participants with DLD, and it is therefore impossible to directly establish a connection between the operationalizations and the construct. The results do, however, support the idea that the phonological short-term memory is measured, which is an aspect of the construct, as participant performance decreased in correlation with the complexity of test-items. Phonological short-term memory can be said to be a stand-in for the broader construct (Polit & Beck, 2012). Participants might behave differently in a test situation than they normally would as they are aware that they are participating in a study. This is a potential threat to construct validity, but it was mediated by attempting to encourage the participants. After all, the intention is that they perform to the best of their abilities. This was done by the stars appearing after repetition, and consistent positive body language over video. The way the researcher acts in relation to the participants' responses is also a threat to construct validity, where they might provide clues or hints in their communication with the participants. This could be mediated by using blinded raters, but this was not possible in this pilot. The researcher never verbally communicated with the participants during the testing-phase, and parents were instructed to refrain from aiding the participants. Parents would sometimes respond to their children with praise when they correctly replicated a word. It is expected that this latter threat is not particularly relevant to this study, as all the items are different, and so there is a limit to how clues or hints to proper reproduction could occur.

Content validity

Content validity is the degree to which the measure represents all facets of the given construct (Lawshe, 1975). It is believed that the LS NWR test has good content validity, as it is based on the LITMUS framework, which shows that the factors included in the test and the nature of the measure are applicable to measuring DLD.

Face validity

Face validity is the subjective opinion to whether the operationalization seems like a good translation of the construct on the surface (Holde, 2010). Albeit a weak form of construct validity, it is strengthened in this case by the reliance on the LITMUS framework. The test consists of factors deemed to measure DLD, as stated by the framework. On this foundation, it is expected that the LS NWR test and video has good face validity.

Criterion validity

Criterion validity, also called instrumental validity, is a measure of the quality of measurement methods (Shuttleworth, 2009, 12.01). Investigating the criterion validity of the LS NWR test was one of the research goals in this study, facilitated by the inclusion of the already established CL NWR test. The CL test has already shown to be valid through previous research (Boerma et al., 2015; Chiat & Polišenská, 2016). The Pearson's correlation coefficient was used to evaluate the concurrent validity of the LS test (part of criterion validity), by gauging the correlation between the two tests.

External validity

External validity concerns, as briefly pointed out earlier, how well empirical findings can be generalized across the population the study investigates (Polit & Beck, 2012). As pointed out in reference to statistical conclusion validity, the number of participants in this study is low, which affects how representative the sample is expected to be. In this case, the population the study wishes to generalise across are monolingual Norwegian speaking children. Participants are also recruited from a small geographical area (Oslo & Viken), and so it is right to question how representative for all first-language Norwegian speakers the sample is. Choosing a smaller geographical area was deemed as necessary, as the items in the LS NWR test were produced with a low-tone tonal accent, which would be more familiar to them.

Reliability

The stability of the LS NWR test could not be ascertained, as it would require a test-retest. This was not feasible to perform before the deadline. Investigating whether participants scored similarly on the tests on separate occasions would have provided information on this aspect (Polit & Beck, 2012). Unfortunately, as this thesis was a solo project it was not possible to investigate the inter-rater reliability of participant scoring. This is however to a degree mediated by the fact that audio recordings were performed, which allowed for rating of participant recordings several times. This was realized at a time when it was no longer feasible to include another rater, as they would be rating sensitive personal information, and this was not cleared in the application to REK. This is a clear limitation of this research and a gap which can be closed through future research.

The internal reliability was rather investigated using Cronbach's coefficient alpha. The Cronbach's alpha coefficient was .7 for the LS NWR test, but only .34 for the CL NWR test. An alpha of .7 is seen as good internal reliability (DeVillis, 2012), but .34 is not acceptable. This might be because the test is quite short, at only 16 items, as this reduces the value of alpha (Nunally & Bernstein, 1994). When calculating Cronbach's alpha coefficient on both test-sets, we end up with an alpha of .73. This is higher than what the LS NWR test achieved on its own, and in conjunction with the correlation analysis it supports the idea that the tests measure the same variable.

Discussion

A benefit of an NWR test is the speed and ease in which it can be performed. The administration of both tests was around 8 minutes in total, where the video itself is exactly 6 minutes long. All participants, barring one, responded to every single item present in the test administration. It can therefore be said that the conduction of the test is feasible, and that it is not too long as to tire out participants. There was only one case in which the video had to be restarted, as the participant did not respond to the first item. This issue was resolved once another round of priming had been performed. It can therefore be said that the guidelines for the test are sufficient. The wizard in the video also tells the participants that they need to repeat the magic words he is saying. The LS NWR test was developed according to the LITMUS framework and includes all the factors the framework deems as having the largest impact on participant performance. It is therefore a supplement which can measure phonological memory, without further increasing the load on both the speech therapist and the child who also usually perform several other diagnostic tests. Although it is based on the LITMUS framework there were many decisions on inclusions and exclusions which were up to my intuitions as a native speaker with a linguistic background. This, however, comes part and parcel with creating a test such as this. It includes the same number of items as the English LS NWR, and controls the same factors, with the addition of tonal accent (Chiat, Polišenská, & Szewczyk, 2012). In retrospect, it would be better served to present the CL- and LS NWR tests as individual units, but the results do not seem to indicate any contamination. It does not rely on specific vocabulary, although some items were repeated with a substitution to create a real word. Future testing is necessary to ascertain if these items need to be removed, creating new items to fill their category. The quantitative analyses which can be performed on the Norwegian LS NWR test can be used to compare performance, not just on different factors, but also across the languages the participant speaks. The test can also, through future work, be adapted to clinical needs, providing information on whether an individual has difficulties pertaining to phonological short-term memory. It should, however, be performed in conjunction with an established diagnostic test. As presented in the theoretical background, DLD is heterogenous, and an individual might still have DLD without having any apparent difficulties with phonological short-term memory. Just as it is impossible to create a truly universal CL test, it is also impossible to create a truly universal Norwegian LS NWR test. This limitation was recognised early on. The phoneme selection ensured that no realizations only occur in certain dialects, except the /r/ realization. The items were however, created to be flexible, so that items could be re-recorded with other phoneme realizations if future testing deems this necessary. This will not remove any existing factors from an item category.

References

Abrahamsen, J. E., & Morland, A. (2012). Fonem. In J. E. Abrahamsen & A. Morland (Eds.), Starthjelp i fonetikk og lingvistikk (3. utg., s. 8-26). Trondheim, Norway: Fagbokforlaget Vigmostad & Bjørke AS.

Archibald, L. M., & Gathercole, S. E. (2006). Nonword repetition: A comparison of tests. *Journal of Speech, Language, and Hearing Research, 49*(5), 970-983. doi:10.1044/1092-4388(2006/070)

Bishop, D. V., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2017). Phase 2 Of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry, 58*(10), 1068-1080. doi:10.1111/jcpp.12721

Bishop, D. V. M. (2004). Specific language impairment: Diagnostic dilemmas. In L. Verhoeven & H. v. Balkom (Eds.), *Classification of developmental language disorders: Theoretical issues and clinical implications* (pp. 309-326). Mahwah, N.J: Lawrence Erlbaum Associates.

Bjerkan, K. M. (2005). Fonologi. In K. E. Kristoffersen, H. G. Simonsen & A. Sveen (Eds.), *Språk. En grunnbok* (pp. 198-221). Oslo, Norway: Universitetsforlaget.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research, 58*, 1747–1760.

Carter, A. K., Dillon, C. M., & Pisoni, D. B. (2002). Imitation of nonwords by hearing impaired children with cochlear implants: supragemental analyses. *Clinical Linguistics and Phonetics, 16*(8), 619-638.

Carter, A. K., & Gerken, L. (2003). Similarities in weak syllable omissions between children with specific language impairment and normally developing language: a preliminary report. *Journal of Communication Disorders, 36*(2), 165-179.

Casalini, C, Brizzolara, D., Chilosi, A., Cipriani, P., Marcolini, S., Pecini, C., Roncoli, S. and Burani, C. (2007). Nonword repetition in children with specific language impairment: A deficit in phonological working memory or in long-term verbal knowledge? *Cortex, 43*,769- 776.

Chiat, S. (2015). Nonword repetition. In S. Armon-Lotem, J. D. Jong & N. Meir (Eds.), *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment* (pp. 125-150). Bristol: Multilingual Matters. doi:10.21832/9781783093137

Chiat, S., & Polišenská, K. (2016). A framework for crosslinguistic nonword repetition tests: Effects of bilingualism and socioeconomic status on children's performance. *Journal of Speech, Language, and Hearing Research, 59*(5), 1179-1189. doi:10.1044/2016_jslhr-l-15-0293

Chiat, S., Polišenská, K. and Szewczyk, J. (2012). Crosslinguistic Nonword Repetition Tasks: British English version. A part of LITMUS COST IS0804 Battery. Unpublished material

Chiat, S., & Roy, P. (2007). The Preschool Repetition Test: An evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research*, *50*, 429-443.

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of nonword repetition tasks in children with and without specific language impairments. *International Journal of Language & Communication Disorders, 43*(1), 1-40. doi:10.1080/13682820601116485

Conti-Ramsden, G., & Botting, N. (1999). Classification of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 42*(5), 1195-1204. doi:10.1044/jslhr.4205.1195

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment. *Journal of Child Psychology and Psychiatry, 42*(6), 741-748. doi:10.1017/S0021963001007600

Crystal, D. (1994) *An encyclopedic dictionary of language and languages.* London, U.K: Penguin Books Ltd.

Cutler A. (1998) Prosodic Structure and Word Recognition. In: *Language Comprehension: A Biological Perspective*. Springer, Berlin: Heidelberg. doi:10.1007/978-3-642-97734-3_2

Deanda, S., Arias-trejo, N., Poulin-dubois, D., Zesiger, P., & Friend, M. (2015). Minimal second language exposure, SES, and early word comprehension: New evidence from a direct assessment. *Bilingualism: Language and Cognition, 19*(1), 162-180. doi:10.1017/s1366728914000820

DeVellis, R. F. (2012). Scale development: Theory and applications (3 ed.). Thousand Oaks, California: Sage.

Dodd, B., Crosbie, S., McIntosh, B., Holm, A., Harvey, C., Liddy, M., Fontyne, K., Pinchin, B., & Rigby, H. (2008). The impact of selecting different contrasts in phonological therapy. *International Journal of Speech-Language Pathology, 10*(5), 334-345. doi:10.1080/14417040701732590

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech Language and Hearing Research, 41*(5), 1136-1146.

Egeberg, E. (2016). *Minoritetsspråk og flerspråklighet: En håndbok i utredning og vurdering* (2 ed.). Oslo, Norway: Cappelen Damm akademisk.

Fintoft, K. (1982). *Fonetikk: oppgaver med svar.* Trondheim, Norway: Tapir.

Fintoft, K., Bollingmo, M., Feilberg, J., Gjettum, B, & Mjaavatn, P. E. (1983). *4 år: En undersøkelse av normalspråket hos norske 4-åringer.* Trondheim, Norway: Universitetet i Trondheim -Norges Lærerhøgskole.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of Segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*(4), 481-496. doi:10.1006/jmla.1999.2692

Førde, R. (2014, 10.10). Helsinkideklarasjonen. Retrieved from https://www.etikkom.no/fbib/praktisk/lover-og-retningslinjer/helsinkideklarasjonen/

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*(1), 83-94. doi:10.3758/bf03210559

Gathercole, S E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27*, 513–543.

Gathercole, S., & Baddeley, A. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language*, *29*, 336-360.

Gerken, L. (1994). Metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of Child Language, 21*(3), 565-584.

Graf Estes, K., Evans, J. and Else-Quest, N. M. (2007). Differences in the Nonword Repetition: Performance of Children With and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language and Hearing Research, 5*, 177-195.

Hayes, B. (2009). *Introductory phonology.* Chichester, England: Wiley-Blackwell.

Holden, R. R. (2010). Face validity. *The Corsini Encyclopedia of Psychology*. doi:10.1002/9780470479216.corpsy0341

International Phonetic Association. (2015). The International Phonetic Alphabet (Revised to 2015). Retrieved from https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf

Janus, M., Labonté, C., Kirkpatrick, R., Davies, S., & Duku, E. (2017). The impact of speech and language problems in kindergarten on academic learning and special education status in grade three. *International Journal of Speech-Language Pathology, 21*(1), 75-88. doi:10.1080/17549507.2017.1381164

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and frequency in specific Language Impairment: Accuracy and error data from Two Nonword Repetition Tests. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1642-1655. doi:10.1044/1092-4388(2010/09-0222)

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory & Language*, *33*: 630–645.

Kohnert, K. (2010). Bilingual children with primary language impairment: issues, evidence and implication for clinical actions, *Journal of Communication Disorders, 43*(6): 456- 473.

Kristoffersen, G. (2000). *The Phonology of Norwegian*: United Kingdom: Oxford University Press.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x

Leonard, L. B. (2014). *Children with specific language impairment* (2. ed.). Cambridge, Massachussets: MIT press.

Meir, N., Walters, J., & Armon-Lotem, S. (2015). Disentangling SLI and bilingualism using sentence repetition tasks: The impact of L1 and L2 properties. *International Journal of Bilingualism*, 1 –32. doi:10.1177/1367006915609240

Morland, A. (2012). Artikulasjon og fonetisk transkripsjon. I J.E. Abrahamsen & A. Morland (Eds.), Starthjelp i fonetikk og lingvistikk (3.ed, pp. 8-26). Trondheim, Norway: Fagbokforlaget Vigmostad & Bjørke AS.

Nunnally, J., & Bernstein, L. (1994). *Psychometric theory*. New York, NY: McGraw-Hill Higher, Inc.

Papazian, E. (1994). «Om sje-lyden i norsk, og ombyttinga av den med kje-lyden», *Norskrift, 83*: 1-105. University of Oslo: Institutt for nordistikk og litteraturvitenskap.

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistic, 31*(2): 3-28.

Paradis, J., Crago, M., Genesee, F., & Rice, M. (2003). French-English bilingual children with SLI: How do they compare with their monolingual peers? *Journal of Speech, Language, and Hearing Research, 46*(1), 113. doi:1092- 4388/03/4601-0113

Paradis, J., Genesee, F., & Crago, M. B. (2011). *Dual language development and disorders: A handbook of bilingualism and second language learning* (2 ed.). Baltimore, MD: Brookes.

Polišenská, K., & Kapalková, S. (2014). Improving child compliance on a computer-administered nonword repetition task. *Journal of Speech, Language, and Hearing Research, 57*(3), 1060-1068. doi:10.1044/1092-4388(2013/13-0014)

Polit, D. F., & Beck, C. T. (2012). *Nursing research: Generating and assessing evidence for nursing practice* (9 ed.). Philadelphia: Wolters Kluwer Health.

Preus, A. (1982). Barn med artikulasjonsvansker. Oslo, Norway: Universitetsforlaget.

Reuterskiöld-Wagner, C., Sahlén, B., & Nyman, A. (2005). Non-word repetition and non-word discrimination in Swedish preschool children. *Clinical Linguistics & Phonetics, 19*(8), 681-699. doi:10.1080/02699200400000343

Roy, P., & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: Evidence from typically developing children. *Journal of Speech, Language, and Hearing Research, 47*, 223-234.

Sahlén, B., Reuterskiöld-Wagner, C., Nettelbladt, U., & Radeborg, K. (1999). Non-word repetition in children with language impairment pitfalls and possibilities. *International Journal of Language and Communication Disorders, 34*(3), 337-352.

Salameh, E., Nettelbladt, U., Håkansson, G., & Gullberg, B. (2002). Language impairment in Swedish bilingual children: A comparison between bilingual and monolingual children in Malmö. *Acta Pædiatrica, 91*(2), 229-234. doi:10.1111/j.1651-2227.2002.tb01700.x

Schwarz, I. E., & Nippold, M. A. (2002). The importance of early intervention. *Advances in Speech Language Pathology, 4*(1), 69-73. doi:10.1080/14417040210001669271

Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *Early repetition battery (ERB): Manual*. London, U.K: Pearson Assessment.

Shuttleworth, M. (2009, 12.01). Criterion Validity. Retrieved from https://explorable.com/criterion-validity

Snowling, M. J. (1981). Phonemic deficits in developmental dyslexia. *Psychological Research*, *43*(2), 219–234. doi: 10.1007/bf00309831

Snowling, M., Bishop, D. V. M., & Stothard, S. E. (2000). Is preschool language impairment a risk factor for dyslexia in adolescence? *Journal of Child Psychology and Psychiatry,41*,587-600.

Snowling, M., Goulandris, N., Bowlby, M., & Howell, P. (1986). Segmentation and speech perception in relation to reading skill: A developmental analysis. *Journal of Experimental Child Psychology*, *41*(3), 489–507. doi: 10.1016/0022-0965(86)90006-8

Sorenson Duncan, T., & Paradis, J. (2016). English language Learners' Nonword Repetition performance: The influence of Age, L2 vocabulary size, length of l2 exposure, and L1 phonology. *Journal of Speech, Language, and Hearing Research, 59*(1), 39-48. doi:10.1044/2015_jslhr-l-14-0020

Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword Repetition and Sentence Repetition as Clinical Markers of Specific Language Impairment: The Case of Cantonese. *Journal of Speech, Language, and Hearing Research*, *49*(2), 219–236. doi: 10.1044/1092-4388(2006/019)

Thordardottir, E. and Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders, 46*, 1–16.

Vanvik, A (1979). *Norsk fonetikk. Lydlæren i standard østnorsk, supplert med materiale fra dialektene*. Oslo, Norway: Universitetet i Oslo: Fonetisk institutt.

Van Dommelen, W. (2003). An acoustic analysis of Norwegian /ç/ and /ʃ/ as spoken by young people. *Journal of the International Phonetic Association, 33*(2), 131-141. doi:10.1017/S0025100303001245

Van Weerdenburg, M., Verhoeven, L., & Van Balkom, H. (2006). Towards a typology of specific language impairment. *Journal of Child Psychology and Psychiatry, 47*(2), 176-189. doi:10.1111/j.1469-7610.2005.01454.x

Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed.

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword Repetition Performance in School-Age Children With and Without Language Impairment. *Journal of Speech, Language, and Hearing Research*, *43*(4), 865–878. doi: 10.1044/jslhr.4304.865

Yuzawa, M., & Saito, S. (2006). The role of prosody and long-term phonological knowledge in Japanese children's nonword repetition performance. *Cognitive Development, 21*, 146-157. doi: 10.1016/j.cogdev.2006.01.003

The Development and Pilot of a Norwegian Nonword Repetition Test Following the Framework of Language Impairment Testing in a Multilingual Setting

Martin Alexander Gulbrandsen

Master's programme in Speech Therapy

Faculty of Psychology

Department of Biological and Medical Psychology

University of Bergen

Spring 2021

Sammendrag

Det er en mangel på verktøy for screening og diagnostisering av barn med utviklingsmessige språkvansker (DLD) i Norge, noe som er veldig tydelig i alderen før barn lærer seg å lese og skrive, og spesielt om de er tospråklige. Hovedformålet med denne oppgaven er å utvikle og pilotere en språk-spesifikk (LS) nonord repetisjonsoppgave (NWR) test som følger rammeverket laget av COST Action IS0804 for Language Impairment Testing in a Multilingual Setting (LITMUS). Et delmål er å undersøke om de forskjellige faktorene fremmet av rammeverket har en påvirkning på vanskelighetsgraden av nonordene i denne norske LS NWR'en. Den teoretiske bakgrunnen presenterer nåværende forskning og relevante funn, som skaper grunnlaget for begrunnelsen og metoden i denne oppgaven. LS NWR testen er presentert sammen med en kryss-lingvistisk (CL) NWR skapt gjennom COST. Nonordene er randomisert og testet på en populasjon av 13 barn, 5 til 7 år gamle, som er beskrevet som typisk utviklende. Resultatene viser at deltakernes skårer på LS NWR testen korrelerer med resultatene deres på CL NWR testen, noe som er lovende for videre forskning på test-settet. Lengde og prosodi var vist å ha en påvirkning på resultatene, hvor kun lengde var signifikant. Resultatene viser også at den Norske LS NWR testen har god intern reliabilitet.

Nøkkelord: DLD, nonord repetisjon, LITMUS, flerspråklighet, språktilegnelse

Abstract

Currently, there are few tools for screening and diagnosing children with developmental language disorder (DLD) in Norway, particularly at the age before children learn to read and write, and especially if they are bilingual. The main purpose of this thesis is to develop and pilot a language-specific (LS) nonword repetition (NWR) test following the framework made by the COST Action IS0804 for Language Impairment Testing in a Multilingual Setting (LITMUS). A secondary purpose is to investigate how the factors outlined in the framework for NWR creation impacts the difficulty of the LS NWR items. The theoretical background presents current research and relevant findings, providing the purpose and method of this thesis. The LS NWR is presented in tandem with the crosslinguistic (CL) NWR created through COST Action IS0804. Test-items were randomized and presented to a population of 13 children, 5 to 7 years old, described as typically developing. Results show that participant scores correlate with their results on the already established CL NWR, which is promising for further testing of the LS test-set. Length and prosody of test-items were found to have an impact on participant performance, where only the impact of length was significant. The results also indicate that the Norwegian LS NWR test shows good internal reliability.

Keywords: DLD, nonword repetition, LITMUS, bilingualism, language acquisition

Introduction

Screening of monolingual and bilingual children

An ongoing issue in speech therapy, particularly in Norway, is a lack of applicable screening tests which can aid in identifying developmental language disorder (DLD). The identification of language issues is imperative for the individual to receive support, and the longer it takes, the further they fall behind their peers in the acquisition of literacy skills and building social relations. The first years of school can be a difficult formative experience without any extra support. Screening bilingual children is another difficult task, as most screening tools rely on target language proficiency, and skills which a bilingual child might not have learned yet. The number of bilingual children is growing, based on demographic changes in Europe (Armon-Lotem et al., 2015). Bilingual children are overrepresented in relation to monolingual children four times over in special education classes in Oslo as early as 1998 (Nordahl & Overland, 1998), yet at the same time, young bilingual children are underrepresented internationally for DLD (Kapantzoglou et al., 2012). There are, however, no underlying reasons for why bilingual individuals would or would not be prone to DLD (Egeberg, 2016), and so it is expected that these phenomena are a result of misdiagnosis. The uncertainty about the diagnosis is further strengthened by research showing similarities in linguistic manifestations of acquiring a second language and those seen in individuals with DLD (Paradis et al., 2003). Researchers across Europe banded together through COST Action IS0804 (European Cooperation in Science and Technology) to cooperate in creating a framework and tools to properly assess bilingual children in a test battery named LITMUS (Language Impairment Testing in a Multilingual Setting), with the ideal of creating tests so that a bilingual child can be tested in both their languages. These tools are available for several languages. One of these tools is nonword repetition (NWR), which importantly does not rely on target language vocabulary or literacy skills. This tool can therefore both be a valuable tool in the assessment of bilinguals and monolinguals and can be utilized before the acquisition of literacy skills.

The main goal of this study is to develop a Norwegian language-specific (LS) NWR test following this framework, with a sub-goal of piloting it on monolingual children with typically developing (TD) language skills to investigate whether conduction of the test is feasible, and to what degree it correlates with results on the already established crosslinguistic (CL) NWR test created for LITMUS (Chiat, 2015). The main difference between these two tests is that an LS NWR test can consist of language-specific properties to measure target language development more accurately, whilst a CL NWR test focuses on measuring language development without being specific to the target language.

Nonword repetition as a diagnostic tool

NWR can be used to evaluate language acquisition in children and can be used in conjunction with other diagnostic tools to identify children with developmental language disorder (DLD) or dyslexia. This is because NWR is designed to closely match the phonological aspect of word learning, and correlates to the level of the user's phonological working memory. It is hypothesized that a deficit in phonological working memory comes part and parcel with DLD (Coady & Evans, 2008). Several studies have shown that children with DLD perform significantly poorer in relation to their TD peers on multisyllabic non-words (Conti-Ramsden & Hesketh, 2003; Gathercole & Baddeley, 1990; Montgomery, 1995). Poor NWR performance has even been found to potentially be a clinical marker for DLD, having the second highest sensitivity (78%) and a high specificity (87%) when compared to other psycholinguistic markers for identifying DLD in monolingual children (Conti-Ramsden et al., 2001). Only NWR and sentence repetition (which is also in LITMUS; and a Norwegian version was developed and piloted by Bome & Vangen, 2017) were shown to accurately identify a history of DLD (Conti-Ramsden et al., 2001). In Gathercole and Baddeley's (1990) study, they found that 8-year-old children with DLD performed at the same level as 6-year-old typically developing (TD) children on standardised testing of vocabulary, comprehension, and reading. Performance on NWR, however, corresponded with the scores of on average 4-year-old TD children when matched for vocabulary and language comprehension. This therefore represents a delay of 4 years in non-word repetition ability (Gathercole & Baddeley, 1993). As NWR simulates the acquisition of novel words, performance on such a test can be influenced by language-specific knowledge although an NWR test does not directly draw on the vocabulary of a language. Kohnert (2013) promotes NWR for assessment of bilinguals, specifically because it is a process-dependent task which does not rely on language input to the degree that experience-dependent tasks do. This aspect was the impetus for Dollaghan & Campbell (1998) creating a NWR test, as they argue that traditional, standardised tests are too reliant on a child's general language experience and specifically their vocabulary. Evidence shows, however, that children have an easier time repeating nonwords which share phonological characteristics with actual words in their language (Jones et al., 2010; Leclercq et al., 2013). There are several Norwegian NWR tests, but they are usually sub-tests which do not manipulate complexity, barring number of syllables, and there does not exist a Norwegian NWR test which follows the LITMUS rationale.

The relationship between storage and comprehension

Language comprehension requires a complex interaction of linguistic and cognitive processes (Sahlén et al., 1999). One of the prerequisites for comprehension is short-term memory, as it is imperative in forming semantic representations of words and sentences (Baddeley, 1986; Gathercole and Baddeley, 1990). Attempting to decipher how this impact is realized requires an understanding of how the working memory and its constituents work. According to the working memory model by Gathercole and Baddeley (1990, 1996) there are four parts which constitute the working memory: a central executive unit which directs attention, and the three slave systems; the visuospatial sketchpad, the phonological loop and the episodic buffer. The visuospatial sketchpad is specialized for processing visuospatial information/visual semantics. The phonological loop is specialized for processing of verbal material/language. The episodic buffer is specialized for integrating the functions of the visuospatial sketchpad and the phonological loop with a sense of time, so that the information they process occurs in a continuous sequence.

The focus here is on the phonological loop. Gathercole and Baddeley (1990) argue that the phonological loop is impaired in individuals with DLD, and that DLD is, more specifically, a consequence of a deficit in the phonological loop. Adams and Gathercole (1995) as well as Montgomery (1995) claim that it is crucial to have a well-functioning phonological loop to develop vocabulary and grammar. Tallal & Curtis (1991) claim, in a similar vein, that problems relating to sentence comprehension are due to restrictions in short-term memory, and not due to a lack of linguistic knowledge.

NWR and bilingualism

As earlier mentioned, a known issue in diagnosing language problems is assessing the language development of a bilingual child. It is typically difficult to determine whether perceived language problems stem from the naturally occurring deprivation of target stimulus, or if they are caused by deficient learning mechanisms (Kohnert, 2010; Paradis, 2010). Whether the child has parents who speak different languages, or if they speak one language at home and another outside the home, they will typically receive less exposure to one language in relation to their monolingual peers (Boerma et al., 2017). In addition to learning an extra language at the same time as their peers, they are required to manage and monitor these two languages. The most important environmental factor for language development is the linguistic input that a child is exposed to (Gathercole & Hoff, 2007), and so less exposure equals less development. Problems arise when a monolingual speech therapist attempts to evaluate a bilingual child who speaks a language the speech therapist is not familiar with, and there are few tests they can employ. In this scenario, nonword repetition can be advantageous, as it is based on, by definition, repeating words that the child has not heard before. If the children are unable to benefit from specific vocabulary, they should not be disadvantaged if their target language exposure or knowledge of target language words is limited. There is a limited amount of research which has gone into this area, which found that differences in language experience amongst monolingual children have less impact on NWR test performance in comparison to tests of vocabulary and grammar (Campbell et al., 1997; Engel et al., 2008). It is important to clarify the usage of the word "disadvantaged". As mentioned earlier, research has shown that target language vocabulary does impact the performance on a NWR test, and as such it is expected that a monolingual child will perform better on a language-specific (LS) test in comparison to their bilingual peers. Although children perform better on NWs which are influenced by their native language, a test such as this one which utilises a cross-linguistic (CL) test and an LS test is postulated to help uncover whether the bilingual child has DLD or not. If the child performs well according to normative data on the CL test, but struggles on the LS test, one could draw the connection that their difficulties with the target language stems from unfamiliarity with the language and not from DLD.

The LITMUS NWR framework

As earlier mentioned, COST Action IS0804 created a framework for developing NWR tests, which is presented by Chiat (2015). This framework comprises different factors deemed to impact the difficulty of NWR and the degree to which language-specific properties affect the test. In the following paragraphs, these factors will be described. In the Method section, these factors will be applied in a Norwegian context.

Length

The length of an NW, calculated by the number of syllables present, was one of the first factors observed to impact performance on NWR. As the number of syllables increase, so does the difficulty of repetition. Shorter nonwords tax the phonological short-term memory to a lesser degree, correlating to a higher accuracy among participants (Stokes et al., 2006, Weismer et al., 2000). Typically, NWR tests consist of NWs ranging from 2-4 syllables, although this depends on several factors. 5-syllable NWs are often included in LS NWR tests when 5-syllable real words are typical in a speaker's vocabulary, or for languages in which some other factors which increase the complexity of a NW are less relevant (for instance by a lack of consonant clusters, which increase the difficulty of NWR). On the other hand, monosyllabic items are rarely included, as they often produce ceiling effects (Graf Estes et al., 2007).

Segmental complexity

Segmental complexity in this case refers to the complexity of phonological sequencing in a word or NW. An example of this is how the Norwegian pronunciation of *troop* /tɾɔp/ is more complex than *top* /tɔp/, not simply because it includes an additional phoneme, but because those two initial consonants appear together in a cluster. Another example, which doesn't increase the number of phonemes, is the difference between /bɑl/ *ball* and /blɑ/ *leaf*, where the latter is seen as more complex, again because the speaker has to produce two different consonants in quick succession, a consonant cluster. The complexity of the segmental sequencing of NWs in an NWRT has recently emerged as a key factor, where it has been shown that children perform consistently better when repeating NWs containing only single consonants compared to NWs containing clusters (Archibald & Gathercole, 2006; Jones et al., 2010). Controlling for segmental complexity is therefore a valid approach to controlling the difficulty of an NWR test. This effect is expected to be dependent on familiarity with complex syllabic structures, such as consonant clusters, but to which extent this familiarity affects results an NWR including consonant clusters is not known (Chiat, 2015). Segmental complexity is controlled for, when applicable for an NWR test, by including NWs with clusters in different applicable positions in the test-set. Some NWs will in that scenario have no clusters, an initial cluster, a medial cluster, or a final cluster.

Prosody

Prosody refers specifically to the elements of speech which are properties of syllables or larger units, such as words, and include amongst other features, stress, and intonation (Cutler, 1998). Considering how wildly prosodic characteristics differ between languages, the test framework focuses solely on stress (Chiat, 2015). Prosodic structure has shown to have a significant effect on NWR (Sahlén et al., 1999; Chiat & Roy, 2007; Roy & Chiat, 2004; Williams et al., 2013; Yuzawa & Saito, 2006). This effect is particularly apparent in the study by Archibald and Gathercole (2006), where the same syllable sequence was repeated twice, with differing prosody, and produced different results in the same participants. According to the framework, prosody and segmental complexity should be controlled separately to allow for investigation without making the test excessively long (Chiat, 2015). The framework dictates that, in languages with variable word stress, 6 items should be assigned atypical stress patterns, meaning 2 at each NW length (which assumes that the LS NWR test controls length by including items of 2-4 syllable length). The remaining 18 will therefore bear typical stress patterns.

Phonotactics and phonotactic probability

Phonotactics concerns the possible sequential arrangements phonological units can have in a language and their position (Crystal, 1994). An example which is applicable for both English and Norwegian would be how /spr-/ can occur initially, but /spm-/ would be an “illegal” arrangement and therefore not allowed in the target language. Traditionally, phonotactic features have been regarded in the dichotomous terms of legal and illegal, but the sounds which can be found in the legal category of a language (i.e. the sound combinations that are possible in the language) certainly do not occur with the same frequency. The frequency of phonological segments and sequences of said segments occurring in a language is referred to as phonotactic probability (Jusczyk, Luce, & Charles-Luce, 1994). An NW which has high phonotactic probability consists of segments which are more likely to occur in the target language. To ascertain phonotactic probability, it is typical to analyse n-grams. N-grams show the probability of phonemes or syllables appearing in a contiguous sequence (Chiat, 2015). A sample of text or speech is used to create an n-gram. As it is required to have access to a corpus which provides the necessary information to create n-grams this factor is not applicable in all languages. There is, however, another factor that can serve as an alternative measure, namely wordlikeness.

Wordlikeness

Controlling for wordlikeness is recommended by the framework, in lieu of information on phonotactic probability. Wordlikeness is a measure of the degree to which NWs resemble real words in the target language according to a native speaker’s judgement. It is therefore not an inherent property of the NW, like phonotactic probability is, but rather a subjective rating. It is assumed, however, that this subjective rating is influenced by the inherent properties of the NW, impacted by length, prosody, and the frequency of phoneme sequences (phonotactics). NWs which have typical length, prosody, and higher frequency phoneme sequences will therefore be judged as more wordlike, which was found to have a correlation with objective phonotactics in a study by Frisch et al. (2000). Wordlikeness can also be influenced by whether items contain syllables or syllable sequences that are real morphemes (Casalini et al., 2007), and can also therefore be controlled by the inclusion or exclusion of these factors. High and low wordlikeness has shown significant effects on a child’s performance on NWR (Gathercole, 1995). When an NWR test is controlling length and prosody, it is hypothesized that wordlikeness can be utilized as a pseudo measurement of phonotactic probability. It should be noted that wordlikeness can only provide a broad measure of the phonological familiarity of NWs and is therefore never preferable to analysis of phonotactic probability. When wordlikeness is used, it is controlled in the same manner as phonotactic probability, where each unique item group consists of one item with high wordlikeness and one item with low wordlikeness. This doubles the size of the test-set but creates more ground for comparison of the other factors as well.

Aims of the study

The overarching aim of the pilot is, as briefly mentioned earlier, investigating the feasibility in conducting the LS NWR test, establishing its face validity. The goal is also to investigate how the factors that were implemented in the test impact the complexity of the test. The CL NWR test was also included to compare results on the two tests and thus to investigate a potential relationship in performance. Correlation does indeed not imply causation, but a potential correlation in scores on the CL and LS can still be treated as supporting evidence for the idea that the Norwegian LS NWR test can measure DLD. The existence of the CL test raises the question as to whether there is a point to making an LS NWR test, if the CL test is sufficient. The benefit of the CL test is that it is applicable with children regardless of language background. An LS NWR test however, is expected to be more discriminating, as it includes LS complexities (Chiat, 2015). The research questions are as follows:

- To what degree can the internal reliability of the LS NWR test be ascertained at this preliminary stage of LS NWR test development?
- How do the results indicate that the participants perform on the CL test in relation to the LS test?
- Can a correlation between results on the CL test and the LS test be established to support the hypothesis that the LS test can measure language skills?
- Will the different factors utilized to construct the LS NWR test have an impact on performance, and to what degree?
- Can unique error types inform the continued development of the LS NWR test?

Method

Creation of the crosslinguistic nonword repetition test

As previously mentioned, this study utilizes the CL test created by Chiat, Polišenská, & Szewczyk (2012) for the LITMUS test battery. This NWR test was created with the sole goal of being a test which children with different native languages could approach on even footing, using phonemes available in most if not all European languages, and attempting to avoid language-specific properties which cater to specific languages (Chiat, 2015). It is not possible to create a truly universal NWR test, as items will undeniably carry language-specific properties such as length and segmental complexity which are more reminiscent of one language than another. Prosody varies between languages, and it is therefore not possible to create a test set with truly language-neutral prosody, but the assumption this test set is based on is that the most neutral prosody is one which avoids particular prosodic patterns all together, stressing each sequence equally. The syllables in these items are produced with even length and pitch, but with final syllable lengthening which is characteristic of the end of an utterance, as according to the framework (Chiat, 2015). It consists of 16 items, ranging from 2-syllable length to 5-syllable length. To accommodate for languages of varying segmental complexity, consonant clusters are not included. As consonant clusters are not included, the difficulty of the test is secured by including 5-syllable length items. The CL NWR test consists of several alternatives for each item with the aim of ensuring that a set can be assembled without using items which include a real morpheme or word from the target language. The narrow phonological inventory and the goal of creating a set which compares almost identical items across languages, culminates in most of the alternatives including segments which are monosyllabic Norwegian words. This also is the case for English, and many alternatives contain at least one component syllable which are real monosyllabic English words (Chiat & Polišenská, 2016). This is further complicated by the neutral prosody of the test-set and can not be avoided. Alternatives including segments which constitute real monosyllabic Norwegian words are therefore included in this pilot, but alternatives including components which constitute as a disyllabic Norwegian word are not included. Alternatives including disyllabic words like <mu'li> *possible* and <li'ta> *small* were not chosen. The full phonological inventory of the CL test is included in the chosen alternatives, except alternatives containing /z/ which does not appear in Norwegian. All alternatives in the CL NWR test can be found in Appendix A**.** The chosen alternatives for this pilot can be found in Appendix

B. The CL NWR test does not control for phonotactic probability, nor wordlikeness, so these factors do not need to be considered for the CL test in this pilot. All item alternatives consist of a simple CV structure.

Creation of the Norwegian Language-Specific nonword repetition

The creation of this LS NWR test follows the COST Action IS0804 framework by controlling for length, segmental complexity, prosody, and wordlikeness (Chiat, 2015). There is no available corpus to provide information on phonotactic probability. It includes 24 items influenced by Norwegian (See Appendix C). Two factors, which are not specifically mentioned in the framework are the phoneme inventory (perhaps because it is implicit) and tonal accents (because they only appear in some languages, such as Norwegian).

An important factor in LS NWR creation, as just mentioned, is the ***phoneme inventory*** of the target language. The creation of this test focused on including a wide set of Norwegian phonemes, excluding phones exclusive to certain dialects. Excluding rarer phonemes was done on the premise of making the test “universal” for Norwegian, so that the phonemes used are common, or at least, semi-common to participants regardless of dialectal background.

To decide on which sounds to include in the LS test-set it was necessary to create an overview of which phones exist in Norwegian, and to create an overview of which phones were to be used. Appendix D & E show all appearing consonants in Norwegian, and which consonants were to be included in the LS NWR test. They show the place of articulation, manner of articulation, and voicing needed to produce these phones. Appendix F & G show all appearing vowels, and which were to be included in the LS NWR test. They show the height of the tongue, the portion of the tongue, which is raised or lowered, and the rounding of the lips. These were based on the overview in Morland (2012) and Kristoffersen (2000). Although the sounds in Appendix D & F appear in Norwegian on a holistic level, similarly to many other languages, they do not appear in all dialects. The goal in making a test specific to a language is for it to be representative for that language as a whole, limiting dialectal impact.

The primary reason these selections were made is that these are the most common realizations of Norwegian target phonemes and occur in most dialects. There are, however, one exceptions. The alveolar tap /ɾ/, which is either voiced or unvoiced depending on position and phonetic environment was chosen as the realization of the target phoneme /r/. /r, ɾ, ʀ, χ, ɹ, ɣ, ɻ, ɰ/ all have similar functions in Norwegian, but although they are allophones, their usage depends on dialect. It is therefore not a “universal” representation to the degree which the other included phones are. It was however chosen because /r/ is a liquid, and liquids are very common in Norwegian. The only other liquid is /l/, and so it was decided that it was necessary to include a representation of /r/. The test-set has been constructed in a way that is flexible to this phoneme realization being changed, without changing segmental complexity. Should future testing on participants with other dialects necessitate a re-recording it would not interfere with the overall construction of the test-set.

Another important factor not mentioned in the framework, is ***tonal accents*** (often referred to as tonemes). It is a prosodic phenomenon of the Norwegian language one cannot ignore. Most Norwegian dialects have distinct tonal accents (Kristoffersen, 2000). Tonal accents are defined by a change in fundamental frequency (f0), commonly referred to as pitch (which is purely a perceptual concept) which is bound to the stressed syllable and can bear meaning. <bønder> (farmers) and <bønner> (beans) are both pronounced as [‘bœnːəɾ] in East-Norwegian, but the first word has a low f0 value on the initial syllable which is stressed, increasing in value towards the unstressed syllable, which is annotated as [^{1}bœnːəɾ], whilst the latter word has a higher initial f0 value, decreasing towards the middle of the word, increasing to a higher f0 value than the initial syllable had, annotated as [^{2}bœnːəɾ] . Norwegian dialects are divided into two types of tonal accents, low tone (East-Norwegian & central-southern Norwegian) and high tone (all other dialects, except for a choice few which have no tonal accent; Kristoffersen, 2000). There are two types of tonal accent for each of the two tone dialects, called accent 1 and accent 2. In low tone dialects, accent 1 is, as seen in the example, characterised by low-high f0, and accent 2 is characterised by high-low-high f0. High tone dialect is the polar opposite, where accent 1 is characterised by high-low f0, and accent 2 is low-high-low f0. Tonal accent is not controlled for in this test set because it is not a part of the framework, and would double the size of the test-set, so it is simply controlled. All items in the LS NWR test are produced with a low tone accent 2, except <hodell> and <tesul>, which are produced with a low tone accent 1, enforced by their structure (Norwegian words with final syllable stress can only be accent 1). This type of tonal accent correlates with the tonal accent of the participants.

The test controls for ***length*** by dividing the 24 items equally into the lengths of 2-4 syllables. As can be seen in the framework, this is common when segmental complexity is also controlled for (Chiat, 2015). The test controls for ***segmental complexity*** in accordance with the framework, which calls for the inclusion of initial and medial clusters when clusters are common in the target language (Chiat, 2015). ***Prosody*** is controlled for by assigning 2 items at each length atypical stress, and the remaining 6 items bear typical stress. Norwegian is like English in that it is typically trochaic, and so typical stress items bear initial stress, whilst atypical (iambic) stress items bear stress on the second syllable. Prosody and segmental complexity, are thus as recommended by the framework, controlled separately. Wordlikeness, in the absence of information on phonotactic probability, had to be measured by creating alternatives to each item and presenting them to native Norwegian speakers. Controlling for wordlikeness substitutes controlling for phonotactic probability but is structured in the same manner. This means that the test-set alternates between items of high and low wordlikeness. The next paragraphs discuss how wordlikeness was measured.

Wordlikeness

Participants and procedure

As previously mentioned, there is no available corpus which could inform the construction of the LS test regarding phonotactic probability. Measuring wordlikeness is more complex than controlling for the other factors, as it requires alternatives which already include the aforementioned factors, and it must be done with the help of participants. Wordlikeness is, after all, a subjective measure, and participant responses did not always align with the creator's intent for an item. 6 native Norwegian speakers took part in measuring wordlikeness, recruited through the creator's personal circle, between the ages of 24-27 from the counties of Telemark, Oslo, and Viken. Participants were selected from these areas, as the items were produced with a low-tone tonal accent in this pilot, and the target phoneme /r/ was realized as [ɾ], which correlates with the participants' dialects. Measurements of wordlikeness were performed several times to finalize the test-set, following the structure presented in Appendix H. In advance of sessions, alternatives were fashioned for each unique item, controlling for the other factors. An example of this is when attempting to attain satisfactory wordlikeness measurements for item 3 at the 2-syllable length, three alternatives were created; [‘skɑpɑ], [‘skɑlɔ], and [‘skɑːnɑ]. They were produced with typical initial stress, an initial cluster, and were subjectively deemed by the creator as sounding Norwegian. Prior to testing, it was also assured that all alternatives were phonotactically legal. The creator's language experience in conjunction with a dictionary was used in determining often, or rarely, occurring segments, combining these in creating alternatives. In this example, all three alternatives were measured as being highly wordlike, but [ˈskɑpɑ], and [ˈskɑlɔ] were not included, based on their similarities to real Norwegian words /kɑpɑ/ *the cape*, /skɑl/ *shall*. Realizing in retrospect that items included real words rarely happened, but when it did, new alternatives were created and tested.

The wordlikeness of the test items were rated by participants on a typical Likert scale, where the participants were asked how much an alternative sounded like a real Norwegian word. 1 being strongly disagree, and 5 being strongly agree. The test items were randomized so as not to allow the participants to recognize a pattern in the nonwords they were rating. As can be seen in Appendix H, the test set alternates between NWs of high and low wordlikeness. Alternatives were created for each category, and if they were not randomised, the participants could potentially realize which category the alternatives were attempting to fill. An item was categorized as having high or low wordlikeness, depending on whether the mean rating was above or below the Likert scale median, which was 3. If more than one alternative attained a mean higher or lower (depending on category) than the Likert scale median (3), an alternative was selected based on holistic phonological variance in the overall test-set. Participants were met online, and after a brief introduction, alternatives were produced by the researcher. NWs were not pre-recorded, and participants could hear each alternative as many times as they wished. Items were repeated if a mistake pertaining to pronunciation occurred. To attempt to avoid biases and distortion which can occur when using a Likert scale, the participants were informed prior to testing that the NWs presented were not necessarily supposed to sound extremely Norwegian or extremely non-Norwegian, and that the value they could provide was their own genuine impression of the NWs presented. The ratings participants provided were only used to calculate a mean rating for each NW, and no qualitative analysis was performed.

Measurement of wordlikeness

As the syllable length increased, some item categories were more difficult to fill than others. Initially, the alternatives contained no segments which were real Norwegian morphemes, but after several attempts at measuring wordlikeness, this criterion was discarded. Chiat (2015) points out that it is possible to manipulate the inclusion of recognisable affixes if it is common to inflect words according to their grammatical category in the target language. Derivational and inflectional affixes were therefore included in the test-set for items of longer syllable length when attempting to fill a high wordlikeness category (See Appendix I). A combination of derivational and inflectional affixes was used, so that a combination of prefixes and suffixes could be included, ensuring that the distribution of affixes was not lopsided on either foot of an item. Real Norwegian words without recognisable affixes are simply too rare for the item alternatives to attain high wordlikeness measurements without them (Kristoffersen, 2000). Items which included a typical Norwegian affix and was attempting to fill a high wordlikeness requirement rarely failed.

Test-set presentation

The 40 NWs from the two test-sets were pre-recorded and presented together in a randomized order (See Appendix J). There is no inherent reason for why the two sets were randomised as a single unit, other than that it is one of the two options of presentation available in the LITMUS framework (Chiat, 2015). This ensures that the stimuli are identical for each participant, and that the length of NWs does not procedurally increase, nor that they appear with alternating wordlikeness ratings. The items were recorded with a professional Røde NT1 microphone. The NWs were presented in an animated video, where a wizard needs the participants' help in saying magic words (See Appendix K). They are told by the character that the stars have disappeared from the night sky, and that they need to use magic words to bring them back. The animation was made by Astrid Ellensdatter Mork-Knutsen, and was purposely kept quite simple, so as not to overstimulate participants. The timeframe of each NW presentation is static, ensuring that the window in which participants can respond does not change. The start of each NW is 8 seconds apart from the next, and the character is on the screen before they start saying an NW, and afterwards, before the skies appear and a star fades in. This timeframe was chosen as it was assumed to provide the participants with ample time to repeat the stimuli before a star starts appearing, and it allows for the character to be on screen before an NW is uttered, preparing the participants. The animated test in its entirety is exactly 6 minutes long.

Pilot study

NWR test participants

This pilot included 13 participants (8 female, 5 male) monolingual children with TD 5-7 years of age, living in East-Norway. The children were recruited through elementary schools, kindergartens, and online postings. No tests of vocabulary or cognitive functioning were performed, as these results are not meant to serve as normative data. The exclusion criteria were the presence of a known hearing impairment, intellectual disability, and DLD.

Procedure and data collection

Because of COVID-19, a choice was made for the data collection to be online. This posed some challenges. The procedure was kept as static as possible for each participant. Participants met with the researcher over the internet on Zoom, with a parent present. The meeting began with an introduction between the researcher and the participant, where they were told that a wizard needed their help in bringing the stars back to the skies, and that they could contribute to that by repeating magical words they have never heard before. After this short introduction, a couple of priming NWs (/bɪɾtə/ and /slɪɾp/) were presented to the participants, which they were told to reproduce. The video began, once understanding of the procedure was ascertained. Audio was recorded on both sides of the interview when possible, preparing for the possible event of audio getting lost. The video was shared with the participants through sharing the researcher's screen. This choice was made to ensure that the researcher could pause the video if necessary. Pausing the video only occurred if an outside distraction occurred, which happened only once. The video was never played twice, and the nonwords were not presented twice if participants asked to hear them again. In one case, the video had to be played from the beginning, when it was apparent that the participant was not properly primed and would not respond to the first two items.

Scoring

The LITMUS framework allows for scoring on two levels: 1) percentage of whole-item correct (PIC) and 2) percentage of phonemes correct (PPC). PIC is used in the Children's nonword repetition task (CNRep) and in the Preschool Repetition Test (Seeff-Gabriel, Chiat, & Roy, 2008), whilst PPC is used in some newer CNRep studies. PIC scores NW repetition on the participant's ability to correctly reproduce all and only the segments of the target NW in the correct order. An occurrence of omission, addition, or substitution therefore constitutes a score of 0 for that item. PPC scores NW repetition on the segmental level, meaning that an error does not equal a score of 0, but rather that each NW has a degree of variance regarding analysis. One phonemic error does therefore not constitute as an overall 0 score but is better than performing two errors. A segment in this scenario refers to all individual phonemes. PPC scoring is more time-consuming but it therefore provides a narrower measure of performance, as it differentiates between the number of errors on one item. Even if segments are phonetically distorted, like if a participant produced an /s/ variant in which the airflow is over the lateral sides of the tongue (sometimes referred to as slurred), the segment would still be scored as correct (Chiat, 2015). This is the case for both types of scoring. Allowances are made if a child produces a substitution that is typical for them or for immature speech (Chiat, 2015). For Norwegian, immature speech includes phenomena such as fronting of velars (/k/, /g/), backing of dentals and alveolars (/t/, /d/), the stopping of fricatives, or the substitution of /ɾ/. This is usually informed by the parents beforehand but is often ascertained after listening to all the child's responses (K. Polišenská, personal communication, April 14, 2021). PIC is sufficient for this pilot, considering how it was expected for participants to make few errors in general, and was also found to be equally discriminating to PPC (See comparison of the two scoring methods in Boerma et al., 2015). PIC is also favoured in the CL NWR test experiment by Chiat & Polišenská, (2016) as it is clinically more appropriate, although it is recognised that this pilot is not clinical.

Results

All results are calculated using the PIC scoring method. To engage with the goal of investigating the impact of the different factors on performance, descriptive analyses were performed for the individual NWR tests. In many cases, it is difficult to perform descriptive analysis, such as Pearson's (*r*), as it cannot be expected that the data is normally distributed. When applicable, statistical analyses were performed, which is explicitly stated when relevant.

Internal reliability

The internal reliability of both tests was calculated using Cronbach's alpha coefficient. The alpha of the LS NWR test-items was .7, whilst the alpha of the CL NWR test-items was .34. The alpha of a scale should ideally be above .7 to have good internal reliability (DeVillis, 2012), which the LS alpha is. The CL alpha, however, is unacceptably low, which is perhaps owed to the length of the test, as it only consists of 16 items. This can in many cases reduce the alpha (Nunally & Bernstein, 1994). The Cronbach's alpha coefficient of both tests as one unit, however, was .73, which supports the idea that the tests measure the same variable.

LS and CL comparison and correlation

Table 1 presents the means and SDs of the PIC performance of all participants. Based on this, the participants did perform better on the CL test (M=88.9) than the LS test (M=82.7). However, performing an independent samples T-test on the scores of each individual participant did not reveal a significant difference in performance ($p=.120$). To investigate whether there was a correlation between individual PIC scores on the CL and LS test, Pearson's correlation coefficient (*r*) was calculated. These results show a strong positive correlation between performance on the two tests, $r = .67$, $n= 13$, $p <.05$. This shows that there is a linear correlation between the two tests, supporting the idea that both tests measure the same skill.

[Insert Table 1 here]

The impact of the different factors

Table 1 also reveals a significant effect of length on the difficulty of NWR, but this effect only appears when increasing from 3- to 4-syllable length. This effect is apparent in both the CL and the LS test. The first choice when comparing these variables would be a paired samples t-test. Considering the low sample size ($N=13$), it was decided to perform a Wilcoxon Signed-

Ranks Test, as it cannot be assumed that the different scores are normally distributed. The null hypothesis is that the population distributions of the different syllable lengths are identical. For the LS NWR test, a Wilcoxon Signed-Ranks test indicated that participants performed significantly better on the “2 syllable” length (mean rank = 5.5) in relation to the “4 syllable” length (mean rank = 0), p = .002. It also indicated that the participants performed significantly better on the “3 syllable” length (mean rank = 6) in relation to the “4 syllable” length (mean rank 0), p = .001. The difference between “2 syllable” (mean rank = 2) and “3 syllable” (mean rank = 2.67) was not significant, p = .5. Performing a Wilcoxon Signed-Ranks test on the CL test cannot be done at each syllable length, as the range in which participants can score is too low for such a test to be valid. The test was therefore done by combining the distribution of scores on syllable lengths 2-3 and 4-5, respectively. The Wilcoxon Signed-Ranks test indicated that participants performed significantly better on the “2-3 syllable” length (mean rank = 5.5) in relation to the “4-5 syllable” length (mean rank = 0), p = .002. These results indicate that the length of items has a strong effect on participant performance once the syllable length exceeds 3 syllables. The difference between the means shown in Table 1 at 2 and 3 syllables on both test-sets is negligible, but performance decreases at the 4- and 5-syllable length.

The results for the other factors included in the pilot only refer to performance on the LS NWR test, as the CL NWR test was only included in this study to investigate the relationship between the two tests, and was not adapted similarly for the factors concerned. These results are in some cases limited to descriptive statistics as some values are too low to perform a Wilcoxon Signed-Ranks test.

As can be seen in the descriptive analysis in Table 2, there is a difference between items with no cluster (M= 85.9), items with an initial cluster (M= 93.6), and items with a medial cluster (M= 82). These results seem to indicate that the items with initial clusters were easier than items with no clusters, which again were easier than items with medial clusters. There is a small impact of segmental complexity on items at the 2- and 3-syllable length. This impact is reversed at the 4-syllable length, and it is also not apparent in the total means for the LS test. According to this data, items with initial clusters were easier to correctly replicate than items with no clusters or items with medial clusters.

[Insert Table 2 here]

According to the descriptive analysis in Table 3, there is a difference between typical and atypical stress when comparing means. Items with typical stress scored a mean of 85.9. Items with atypical stress scored a mean of 73.42. Performing a Wilcoxon Signed-Ranks test however, does not reveal a significant difference in performance on "typical" prosody (mean rank = 4) and "atypical" prosody (mean rank = 6.75), $p = .053$, $p > .05$.

[Insert Table 3 here]

Based on the theoretical background, it was expected that wordlikeness would have an impact on performance. Surprisingly, no significant difference was found between performance on all items with high wordlikeness and low wordlikeness (Table 4), as the total PIC for high wordlikeness items was 85.3% and the total PIC for low wordlikeness items was 82.7%. A Wilcoxon Signed-Ranks test indicated that participants performed similarly on the "Low" wordlikeness items (mean rank = 4.17) in relation to the "High" wordlikeness items (mean rank = 5.5), $p = .438$. It is important to note, that the term wordlikeness is not dichotomous in the sense that an item scoring a 3.1 mean in the wordlikeness measure is counted as high, and an item scoring a 2.9 mean is counted as low. Participants performed better on low wordlikeness items at the 2-syllable length in relation to high wordlikeness items, but this effect was reversed and increased at the 4-syllable length. According to these results it seems as though the effect of wordlikeness is evident at the 4-syllable length.

[Insert Table 4 here]

Appendix L shows each unique incorrect response to all the items in the LS NWR test. Error types were synthesized to investigate whether the validity of the test-items is at risk. The validity is at risk if a child wrongly identifies an NW as a real word. Those of particular interest are the ones where a substitution or addition has taken place to convert an item or an item-segment into a real word/affix. There were 11 unique responses (out of 312 unique elicitations; 3.53%) in which this occurred, which can be seen as a negligible amount. On the subject of phoneme selection in the item-set, item 7 [hu'dɛl], item 18 ['lʉŋɔkɪɾɑ], and item 23 [bə'ti:ʋɛsə], are of particular interest. 3 out of 13 participants made the same error in their repetition of item 7 [hu'dɛl], where the /d/ was replaced by /t/. Not only are these phonemes separated by whether they are voiced or not, this substitution creates the real word [hu'tɛl] *hotel.* Item 18 ['lʉŋɔkɪɾɑ] was incorrectly elicited 8 out of 13 times, where /ŋ/ was replaced by /m/ 7 times, in which it was the only error in 5 cases. This is a considerable number of incorrect repetitions, but the cause is difficult to establish. Lastly, the addition of /l/ occurred in 3 repetitions to form a real affix in Item 23 [bə’tiːʋɛsə], which became [bə’tiːʋɛlsə]. This suffix is used to create verbal nouns and is quite commonly found in Norwegian.

Discussion

Presentation and face validity

A disadvantage of how the two NWR tests were randomized as one unit is that it is difficult to ascertain the face validity of the LS test in isolation. At face value, descriptive analysis seems to indicate that the face validity is high. Participants were in general interested in performing the test, remarking some items as “weird” or “funny” and pointing to appearing stars, guessing how big the next star was going to be. Some participants would laugh or giggle at some of the items they thought were odd. 2 participants asked at several points if the test was done soon, 1 of these stating mid-test that the items were difficult. These participants had the lowest scores and there might be parallels between their attention or interest and their results. There were potential participants who backed out during the interview-phase but considering how they did not enter the test administration, this is unrelated to any issues with the way the tests were presented.

In retrospect, the randomized NWs presented in the video should have gone through a second inspection to ensure that similar NWs did not get presented in sequence. Many of the NWs from the CL set are very similar, and there are circumstances where these similar NWs occur sequentially (See Appendix J).

The timeframe between the presentation of items was found to be long enough; all participants, except for one who repeated one item as it happened, answered before the scene swapped to a star appearing. There was only one case of an item not being repeated. This timeframe also allows for other researchers to use the animation in the creation of other LS NWR tests which follow the LITMUS framework.

Participants were asked about their thoughts on the “magic words” and if they had fun after the test administration. They reported that they thought they sounded funny, and that they had fun. Even the 2 participants who asked mid-test if they were done soon reported this. Some observations may seem anecdotal, but for the test to succeed the children must interpret it correctly, not be distracted by features that are not central to the test and they must demonstrate enough attention to fulfil the task and respond to each individual item. Most participants replicated items with the same prosody as the recording, but this was not done by everyone. This was particularly prominent when a participant was unsure of their repetition.

One initial goal of this pilot was to gauge whether participants scores increased with age. Of the 13 participants, only 3 were pre-school aged, and so it is therefore not valid to investigate this.

Online procedure

Performing the testing online was seen as a necessary choice, but not without its problems. Having the test presented over a zoom conference meant that audio quality for the participants relied heavily on whether they had access to headphones, a strong and consistent internet connection, and outside noise. Audio quality was also an issue in the transcription of participant-responses, as the quality of their microphones varied. This was in most cases circumvented by parents also recording locally on newer mobile phones, which often provided a better recording. The responses were only rated by one person. In retrospect, it would have been fortunate to include a second rater. A single rater means that the reliability of a rating cannot be ascertained. There were few ambiguous responses, as most participants provided a local recording.

Presenting the CL and LS items as a single randomized test did run the risk of contaminating results, particularly because of prosodic differences in the two tests. The results do not seem to indicate this, considering the high mean performances on both test-sets, and the difference in performance on typical and atypical LS items. The atypical prosody of certain LS items still had the expected effect of increasing the difficulty of the items.

The results on several factors were difficult to ascertain, considering the number of identical mistakes on some items (Appendix L, column 2-3). It is likely that it is influenced by the participants' audio environment. /ŋ/ is not an infrequent phoneme in Norwegian, and so there is no reason as to why this phoneme was substituted as often as it was. [hu'dɛl] and[bə'tiːʋɛsə], on the other hand, are more questionable inclusions in the test in future work, particularly [hu'dɛl] based on it being so similar to a real Norwegian word. Future testing is needed to ascertain whether these items need to be replaced. With the inclusion of /ɾ/ and the usage of a consistent low-tone tonal accent, it is also necessary to perform future testing on participants from other parts of Norway to verify if this recording of the LS NWR test truly is "Norwegian", or if the bias to a certain region of Norway is large enough to warrant a re-recording. As earlier stated, the test-set is specifically constructed in a way where new versions with differing allophones can be recorded for other regions, without changing target phonemes.

Conclusion

This thesis was mainly purported to develop a Norwegian NWR test, keeping with the guidelines and principles of the LITMUS framework. These guidelines and principles were helpful and well-funded, even when not specific to Norwegian. Importantly, inferences and assumptions had to be made based on my intuitions as a native Norwegian speaker with a background in linguistics, in addition to relevant literature and research being scarce. The LS NWR test created here was designed to be a part of the LITMUS test battery, and it accounts for relevant research and literature to the best of the creator's abilities. It was piloted on a small selection of participants to investigate several research questions. The internal reliability of the LS NWR test was deemed as good. The results also indicated that the participants performed slightly better on the CL test in relation to the LS test, and correlation analysis on the two tests revealed that there is indeed a strong linear correlation between the results on the two tests. This fits the idea that, in a different way, both tests measure the same skill. The impact of different factors on the difficulty of NWR was also investigated, where length had a significant effect on NWR proficiency. The segmental complexity and wordlikeness did not increase the difficulty when independent of length in this pilot of the LS NWR test. The results seem to indicate that items with medial clusters were more difficult to repeat in relation to items with initial clusters.

The face validity and interactivity of the test and testing procedure was investigated by gauging the participants' engagement with the test, and their responses when asked about their thoughts post-test. The results indicated that the test was not too taxing on participants, as all participants engaged well during the test administration. There were potential items which can threaten the validity of the test-set, needing further research. It should not be understated that this LS NWR test is in its preliminary stages and needs future piloting on not only monolingual children with TD, but also monolingual children with DLD, bilingual children with TD, and bilingual children with DLD to ascertain the effect of factors and the sensitivity and specificity of the test (which can only be established when children with DLD as well as children with TD are included). I would recommend future work to further pilot the test by including a larger group, monolingual and bilingual children with DLD, bilingual children with TD, and to include groups from other regions of Norway.

References

Archibald, L. M., & Gathercole, S. E. (2006). Nonword repetition: A comparison of tests. *Journal of Speech, Language, and Hearing Research, 49*(5), 970-983. doi:10.1044/1092-4388(2006/070)

Armon-Lotem, S., de Jong, J., & Meir, N. (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment.* Bristol, U.K: Multilingual Matters. doi:10.21832/9781783093137

Baddeley, A. (1986). *Working Memory*. Oxford, U.K: Oxford University Press.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research, 58*, 1747–1760.

Boerma, I. E., Mol, S. E., & Jolles, J. (2017). The role of home literacy environment, mentalizing, expressive verbal ability, and print exposure in third and fourth graders' reading comprehension. *Scientific Studies of Reading, 21*(3), 179-193. doi:10.1080/10888438.2016.1277727

Bome, C., & Vangen, I. C. (2017). The development and pilot of a Norwegian adaptation of the LITMUS Sentence Repetition task. *Master thesis*, University of Bergen.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment. *Journal of Speech, Language, and Hearing Research, 40*(3), 519-525. doi:10.1044/jslhr.4003.519

Casalini, C, Brizzolara, D., Chilosi, A., Cipriani, P., Marcolini, S., Pecini, C., Roncoli, S. and Burani, C. (2007). Nonword repetition in children with specific language impairment: A deficit in phonological working memory or in long-term verbal knowledge? *Cortex, 43*,769- 776.

Chiat, S. (2015). Nonword repetition. In S. Armon-Lotem, J. D. Jong & N. Meir (Authors), *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment* (pp. 125-150). Bristol, U.K: Multilingual Matters. doi: https://doi.org/10.21832/9781783093137

Chiat, S., & Polišenská, K. (2016). A framework for crosslinguistic nonword repetition tests: Effects of bilingualism and socioeconomic status on children's performance. *Journal of Speech, Language, and Hearing Research, 59*(5), 1179-1189. doi:10.1044/2016_jslhr-l-15-0293

Chiat, S., Polišenská, K. and Szewczyk, J. (2012). Crosslinguistic Nonword Repetition Tasks: British English version. A part of LITMUS COST IS0804 Battery. Unpublished material.

Chiat, S., & Roy, P. (2007). The Preschool Repetition Test: An evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research, 50*, 429-443.

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of nonword repetition tasks in children with and without specific language impairments. *International Journal of Language & Communication Disorders, 43*(1), 1-40. doi:10.1080/13682820601116485

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (sli). *Journal of Child Psychology and Psychiatry, 42*(6), 741-748. doi:10.1111/1469-7610.00770

Conti-Ramsden, G., & Hesketh, A. (2003). Risk markers for SLI: A study of young language-learning children. *International Journal of Language and Communication Disorders, 38*, 251-263.

Crystal, D. (1994). *An encyclopedic dictionary of language and languages*. London, U.K: Penguin Books Ltd.

Cutler A. (1998) Prosodic Structure and Word Recognition. In A. D. Friederici (Ed.), *Language Comprehension: A Biological Perspective*. Springer, Berlin: Heidelberg. doi:10.1007/978-3-642-97734-3_2

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech Language and Hearing Research, 41*(5), 1136-1146.

Egeberg, E. (2016). *Minoritetsspråk og flerspråklighet: En håndbok i utredning og vurdering* (2 ed.). Oslo, Norway: Cappelen Damm akademisk.

Engel, P. M., Santos, F. H., & Gathercole, S. E. (2008). Are working memory measures free of socioeconomic influence? *Journal of Speech, Language, and Hearing Research, 51*(6), 1580-1587. doi:10.1044/1092-4388(2008/07-0210)

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of Segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*(4), 481-496. doi:10.1006/jmla.1999.2692

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition, 23*(1), 83-94. doi:10.3758/bf03210559

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied psycholinguistics*, *27*(4), 513.

Gathercole, S., & Baddeley, A. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language*, *29*, 336-360.

Gathercole, S. E., & Baddeley, A. D. (1993). *Essays in cognitive psychology. Working memory and language.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gathercole, S. E., & Baddeley, A. D. (1996). *The Children's Test of Nonword Repetition*. London, U.K: Psychological Corporation

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory, 2*(2), 103-127. doi:10.1080/09658219408258940

Gathercole, V. C., & Hoff, E. (2009). Input and the Acquisition of Language: Three Questions. In E. Hoff & M. Shatz (Authors), *Blackwell handbook of language development* (pp. 107-127). Chichester, U.K: Wiley-Blackwell. doi:10.1002/9780470757833.ch6

Graf Estes, K., Evans, J. and Else-Quest, N. M. (2007). Differences in the Nonword Repetition: Performance of Children With and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language and Hearing Research, 5,* 177-195.

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and frequency in specific Language Impairment: Accuracy and error data from Two Nonword Repetition Tests. *Journal of Speech, Language, and Hearing Research, 53*(6), 1642-1655. doi:10.1044/1092-4388(2010/09-0222)

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory & Language, 33*, 630–645.

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S. (2012). Dynamic assessment of word learning skills: identifying language impairment in bilingual children. (Report). *Language, Speech, & Hearing Services in Schools, 43*, 81.

Kohnert, K. (2010). Bilingual children with primary language impairment: issues, evidence and implication for clinical actions, *Journal of Communication Disorders, 43*(6): 456- 473.

Kohnert, K. (2013). *Language disorders in bilingual children and adults* (2nd ed.). San Diego, CA: Plural Publisher.

Kristoffersen, G. (2000). *The Phonology of Norwegian*: Oxford, U.K: Oxford University Press.

Leclercq, A., Maillart, C., & Majerus, S. (2013). Nonword Repetition Problems in Children With Specific Language Impairment. *Topics in Language Disorders*, *33*(3), 238-254. doi:10.1097/tld.0b013e31829dd8c9

Montgomery, J. (1995). Sentence comprehension in children with specific language impairment: The role of phonological working memory. *Journal of Speech and Hearing Research, 38,* 187 199.

Morland, A. (2012). Artikulasjon og fonetisk transkripsjon. In J.E. Abrahamsen & A. Morland (Eds.), *Starthjelp i fonetikk og lingvistikk (3. ed)*. Trondheim, Norway: Fagbokforlaget Vigmostad & Bjørke AS.

Nordahl, T., & Overland, T. (1998). *Idealer og realiteter: evaluering av spesialundervisningen i Oslo kommune* (Vol. 20/98). Oslo, Norway: Norsk institutt for forskning om oppvekst, velferd og aldring.

Nunnally, J., & Bernstein, L. (1994). *Psychometric theory*. New York, NY: McGraw-Hill Higher, Inc.

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistic, 31*(2): 3-28.

Paradis, J., Crago, M., Genesee, F., & Rice, M. (2003). French-English bilingual children with SLI: How do they compare with their monolingual peers? *Journal of Speech, Language, and Hearing Research, 46*(1), 113-127. doi:1092- 4388/03/4601-0113

Roy, P., & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: Evidence from typically developing children. *Journal of Speech, Language, and Hearing Research, 47*, 223-234.

Sahlén, B., Reuterskiöld-Wagner, C., Nettelbladt, U., & Radeborg, K. (1999). Non-word repetition in children with language impairment pitfalls and possibilities. *International Journal of Language and Communication Disorders, 34*(3), 337-352.

Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *Early repetition battery (ERB): Manual*. London, U.K: Pearson Assessment.

Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword Repetition and Sentence Repetition as Clinical Markers of Specific Language Impairment: The Case of Cantonese. *Journal of Speech, Language, and Hearing Research, 49*(2), 219–236. doi:10.1044/1092-4388(2006/019)

Tallal P., Curtiss S. (1990) Neurological Basis of Developmental Language Disorders. In: Rothenberger A. (Ed.) *Brain and Behavior in Child Psychiatry*. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-75342-8_13

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword Repetition Performance in School-Age Children With and Without

Language Impairment. *Journal of Speech, Language, and Hearing Research*, *43*(4), 865–878. doi: 10.1044/jslhr.4304.865

Williams, D., Payne, H. and Marshall, C. R. (2013). Non-word repetition impairment in autism and SLI: Evidence for distinct underlying cognitive causes. *Journal of Autism and Developmental Disorders*, *43*, 404-417.

Yuzawa, M., & Saito, S. (2006). The role of prosody and long-term phonological knowledge in Japanese children's nonword repetition performance. *Cognitive Development*, *21*, 146-157. doi: 10.1016/j.cogdev.2006.01.003

Tables

Table 1

Percentage of Items Correct on the two NWR tests.

		Participants	
NWR test	**Syllables**	Mean	SD
Crosslinguistic	**All**	**88.9**	**8.1**
	2	98.1	7.1
	3	96	11.6
	4	82.7	19.1
	5	76.9	28
Language-Specific	**All**	**82.7**	**11.4**
	2	90.3	15.1
	3	93.3	8.8
	4	62	29.5

Note: SD = standard deviation

Table 2

Descriptive analysis of segmental complexity

Length	Cluster	Mean	SD
2-syllables	No cluster	96.2	13.9
	Initial cluster	92.3	18.8
	Medial cluster	88.6	31.5
3-syllables	No cluster	100	0
	Initial cluster	100	0
	Medial cluster	80.8	25.3
4-syllables	No cluster	61.5	36.3
	Initial cluster	88.5	21.9
	Medial cluster	76.9	33
Total			
	No cluster	85.9	15
	Initial cluster	93.6	10.9
	Medial cluster	82	18.6

Note: SD = standard deviation,

Table 3

Descriptive analysis of prosody

Length	Stress	Mean	SD
2-syllables	Typical	96.2	13.9
	Atypical	84.6	24
3-syllables	Typical	100	0
	Atypical	92.3	18.8
4-syllables	Typical	61.5	36.3
	Atypical	46.2	38
Total			
	Typical	85.9	15
	Atypical	73.4	16

Note: SD = standard deviation

Table 4

PIC-differences between words of high and low wordlikeness

Syllables	*Wordlikeness*	*Total elicitations*	*Whole-item correct*	**PIC**
2	High	52	45	86.5%
	Low	52	49	94.2%
3	High	52	49	94.2%
	Low	52	48	92.3%
4	High	52	39	75%
	Low	52	32	61.5%
Total	High	156	133	85.3%
	Low	156	129	82.7

Note: PIC = Percentage of items correct.

Appendix A - Orthographic Realizations of the Alternatives Present in the CL NWR Test.

(Adapted from Chiat, 2015)

Length	Orthography					
	zibu	sibu	sipu	zipu		
2	lita	lida	dula	tula		
	maki	naki	magi	nagi		
	luni	lumi	nuli	muli		
	sipula	zipula	sibula	zibula		
3	bamudi	banudi	pamudi	panudi	pamuti	panuti
	malitu	malidu	nalitu	nalidu	malitu	malidu
	lumika	lunika	lumiga	luniga		
	zipalita	sipalita	zibalita	sibalita	zipalida	sipalida
4	mukitala	nukitala	mugitala	nugitala	mukidala	mugidala
	kasulumi	gasulumi	kasulumi	gazulumi	kasuluni	gasuluni
	litisaku	lidisaku	litisagu	lidisagu	litizaku	lidizaku
	sipumakila	sibumakila	sipunakila	sibunakila	sipumagila	sibumagila
5	tulikasumu	dulikasumu	tuligasumu	duligasumu	tulikazumu	dulikasumu
	malusikuba	maluzikuba	malusiguba	maluziguba	malusikupa	maluzikupa
	litapimuti	lidapimuti	litabimuti	lidabimuti	litapimudi	lidabimudi

Note: Length refers to syllable length.

Appendix B - Chosen Alternatives to Represent the CL NWR Test

(Adapted from Chiat, 2015)

Syllable length	Orthography	Transcription
2	sipu	[ˈsɪˌpʉ]
	lida	[ˈlɪˌdɑ]
	naki	[ˈnɑˌkɪ]
	nuli	[ˈnʉˌlɪ]
3	sibula	[ˈsɪbʉˌlɑ]
	banudi	[‘bɑnʉ,dɪ]
	malitu	[ˈmɑlɪˌtʉ]
	lumiga	[ˈlʉmɪˌgɑ]
4	sipalita	[ˈsɪpɑlɪˌtɑ]
	nukitala	[ˈnʉkɪtɑˌlɑ]
	gasulumi	[ˈgɑsʉlʉˌmɪ]
	llitisagu	[ˈlɪtɪsɑˌgʉ]
5	sibumakila	[´sɪbʉmɑkɪ,lɑ]
	tuligasumu	[ˈtʉlɪgɑsʉˌmʉ]
	malusikupa	[ˈmɑlʉsɪkʉˌpɑ]
	lidapimuti	[ˈlɪdɑpɪmʉˌtɪ]

Appendix C - The Norwegian Language-Specific Nonword Repetition Test

Stress	Cluster	Nr	Wordlikeness	Syllables	Phonetic transcription	Mean wordlikeness
				2 syllables		
Typical	None	1	High	Nuse	[‘nʉ:sə]	4.17
Typical	None	2	Low	Tumi	[‘tʉ:mɪ]	2.5
Typical	Initial	3	High	Skana	[‘skɑ:nɑ]	4.33
Typical	Initial	4	Low	Klyno	[‘kly:nu]	2.67
Typical	Medial	5	High	Pylte	[‘pʏltə]	4.17
Typical	Medial	6	Low	Bulpi	[‘bʉlpɪ]	2.33
Atypical	None	7	High	Hodell	[hu’dɛl]	3.67
Atypical	None	8	Low	Tesul	[tə’sʉ:l]	2.33
				3 syllables		
Typical	None	9	High	Tarane	[‘tɑ:ɾɑnə]	4.5
Typical	None	10	Low	Dipasse	[‘dɪpɑsə]	1.83
Typical	Initial	11	High	Kvasine	[‘kʋɑ:sɪnə]	4
Typical	Initial	12	Low	Plutade	[‘plʉ:tɑdə]	2.33
Typical	Medial	13	High	Velsjennig	[‘ʋɛlʃənɪ]	4.67
Typical	Medial	14	Low	Sukleme	[‘sʉkləmə]	2.83
Atypical	None	15	High	Befaning	[bə’fɑ:nɪŋ]	5
Atypical	None	16	Low	Sinupe	[sɪ’nʉ:pə]	1,83
				4 syllables		
Typical	None	17	High	Bigapeler	[‘bi:gɑpələɾ]	3.5
Typical	None	18	Low	Lungåkira	[‘lʉŋɔkɪɾɑ]	2.33
Typical	Initial	19	High	Dråtelige	[‘dɾɔtəli:ə]	4.33
Typical	Initial	20	Low	Blisunire	[‘bli:sʉnɪɾə]	2.67
Typical	Medial	21	High	Forhaviget	[‘fɔɾhɑʋi:gət]	4.17
Typical	Medial	22	Low	Konilsedi	[‘ku:nɪlsədɪ]	2.17
Atypical	None	23	High	Betivese	[bə’ti:ʋəsə]	4
Atypical	None	24	Low	Låtterite	[lo’tɛɾɪtə]	2.67

Appendix D - Consonants Present in Norwegian

(derived from Kristoffersen, 2000; Morland, 2012)

Consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g			ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ			
Trill				r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative		f		s	ʃ		ç	x ɣ			h
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral Approximant				l		ɭ	ʎ				
Other sounds							c͡ç ɟ͡ʝ				

Note: Symbols to the right in a cell are voiced, whereas the ones on the left are voiceless. Shaded areas denote articulations which are judged as impossible.

Appendix E - Consonants Used in the Language-Specific Nonword Repetition Test

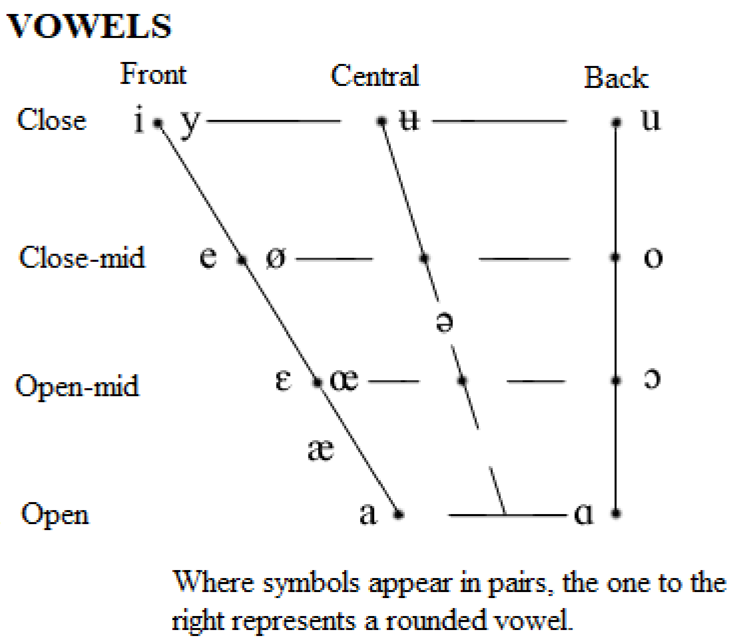
Consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d				k g			
Nasal	m			n				ŋ			
Trill											
Tap or Flap				ɾ							
Fricative		f		s	ʃ						h
Approximant		ʋ									
Lateral Approximant				l							
Other sounds											

Note: Symbols to the right in a cell are voiced, whereas the ones on the left are voiceless. Shaded areas denote articulations which are judged as impossible.

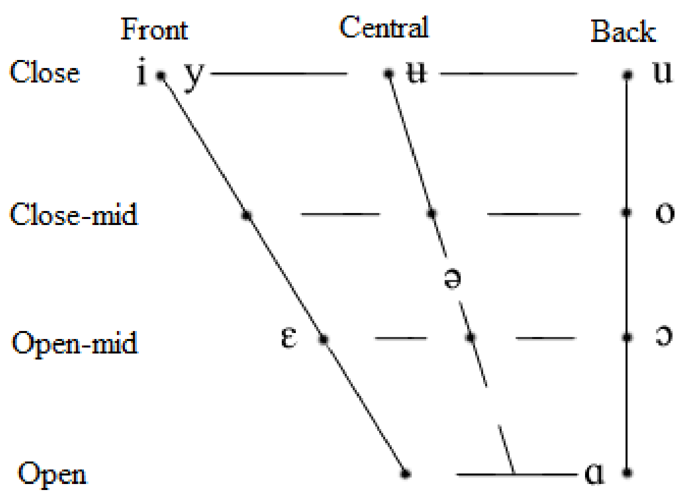
Appendix F - Vowels Present in Norwegian

(Adapted from Kristoffersen, 2000; Morland, 2012)



Appendix G - Vowels Present in the Norwegian LS NWR Test

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

Appendix H - Structure in Creating and Testing Items for Wordlikeness.

This Structure is Repeated at Each Syllable Length.

Nr	*Stress*	*Complexity*	*Wordlikeness*	*Alternative*	*Alternative*	*Alternative*
1	*Typical*	*No cluster*	*High*	*X*	*Y*	*Z*
2	*Typical*	*No cluster*	*Low*	*X*	*Y*	*Z*
3	*Typical*	*Initial cluster*	*High*	*X*	*Y*	*Z*
4	*Typical*	*Initial cluster*	*Low*	*X*	*Y*	*Z*
5	*Typical*	*Medial cluster*	*High*	*X*	*Y*	*Z*
6	*Typical*	*Medial cluster*	*Low*	*X*	*Y*	*Z*
7	*Atypical*	*No cluster*	*High*	*X*	*Y*	*Z*
8	*Atypical*	*No cluster*	*Low*	*X*	*Y*	*Z*

Note: X, Y, Z refer to any three different alternatives for each item.

Appendix I - Affixes Present in the LS NWR Test

Affix	*Example usage*	*Affix*	*Example usage*
Bi-	*Both/ Double*	*-er*	*Plurality*
For-	*Used to derive verbs from other words*	*-ning*	*Create verbal nouns from verbs.* *Create nouns from adjectives.*
Be-	*Connect verb to object without preposition.* *Used in conjunction with noun to create verb.*	*-a**	*Definite article female singular.* *Definite article neuter plural.*
-lig	*Like* (ex: childlike)	*-e**	Plurality

Note: *Constituents which can be used as affixes in Norwegian, but do not constitute as affixes in the creation of this test

Appendix J - Items from the CL test and the LS test as they Appear in the Procedure

1	hodell	[huˈdɛl]	21	tesul	[təˈsʉ:l]
2	banudi	[ˈbɑnʉˌdɪ]	22	lida	[ˈlɪˌdɑ]
3	konilsedi	[ˈku:nɪlsədɪ]	23	klyno	[ˈkly:nu]
4	låtterite	[lɔˈtɛɾɪtə]	24	nuli	[ˈnʉˌlɪ]
5	dråttelige	[ˈdɾɔtəli:ə]	25	lungåkira	[ˈlʉŋɔkɪɾɑ]
6	befaning	[bəˈfɑ:nɪŋ]	26	kvasine	[ˈkʋɑ:sɪnə]
7	tuligasumu	[ˈtʉlɪgɑsʉˌmʉ]	27	malitu	[ˈmɑlɪˌtʉ]
8	tarane	[ˈtɑ:ɾɑnə]	28	sukleme	[ˈsʉkləmə]
9	sinupe	[sɪˈnʉ:pə]	29	nuse	[ˈnʉ:ˌsə]
10	malusikupa	[ˈmɑlʉsɪkʉˌpɑ]	30	lidapimuti	[ˈlɪdɑpɪmʉˌtɪ]
11	nukitala	[ˈnʉkɪtɑˌlɑ]	31	velsjennig	[ˈʋɛlʃənɪ]
12	dippase	[ˈdɪpɑsə]	32	plutade	[ˈplʉ:tɑdə]
13	blisunire	[ˈbli:sʉnɪɾə]	33	litisagu	[ˈlɪtɪsɑˌgʉ]
14	sibumakila	[ˈsɪbʉmɑkɪˌlɑ]	34	betivese	[bəˈti:ʋəsə]
15	pylte	[ˈpʏltə]	35	bigapeller	[ˈbi:gɑpələɾ]
16	forhaviget	[ˈfɔɾhɑʋi:gət]	36	tumi	[ˈtʉ:mɪ]
17	lumiga	[ˈlʉmɪˌgɑ]	37	naki	[ˈnɑˌkɪ]
18	sipalita	[ˈsɪpɑlɪˌtɑ]	38	gasulumi	[ˈgɑsʉlʉˌmɪ]
19	sibula	[ˈsɪbʉˌlɑ]	39	bulpi	[ˈbʉlpɪ]
20	sipu	[ˈsɪˌpʉ]	40	skana	[ˈskɑ:nɑ]

Appendix K - Pictures from the Animated Test

Appendix L – The Error Type for each Incorrect Item in the LS NWR Test.

The error type for each incorrect item in the LS.

Item	Phonological change	Error type	*N*
Tumi	[mɪ] -> [lɪ]	Place & Manner	1
Skana	[nɑ] -> [mɑ]	Place	1
Klyno	[kly:] -> [kli:]	Vowel	1
Pylte	[ˈpʏltə] -> [ˈpʏlpə]	Repetition/ Place	1
	[pʏ] -> [pɪ]	Vowel	1
	[pʏ] -> [pʉ]	Vowel	1
Hodell	[dɛl] -> [tɛl]	Voicing (real word)	**3**
Tesul	[sʉ:l] -> [ʃʉ:l]	Place	1
Velsjennig	[nɪ] -> [nɪŋ]	Addition (real affix)	1
Sukleme	[klə] -> [plə]	Place	1
	[klə] -> [klɪ]	Vowel	1
	[sʉ] -> [sʏ]	Vowel	2
Befaning	[nɪŋ] -> [sɪŋ]	Place	1
	[bə] -> [də]	Place	1
Bigapeler	[bi:gɑ] -> [bi:l]	Reduction/ Place & Manner (real word)	1
	[bi:] -> [pi:]	Voicing	1
Lungåkira	[lʉŋɔkɪ] -> [lʉmɔʃɪ]	Place, Place & Manner	1
	[lʉŋɔ] -> [lʉŋɔə]	Vowel (real word)	1
	[ŋɔ] -> [mɔ]	Place	**5**
	[lʉŋɔ] -> [lʉmɔ]	Vowel, Place, Vowel	1
Dråttelige	[li:ə] -> [li:]	Syllable deletion	1
	No response	N/A	1
Blisunire	[bli:sʉnɪ] -> [bli:sʉmɪ]	Vowel, Place	1
Forhaviget	[ʋi:] -> [li:]	Place & Manner	1
	[ˈfɔɾhɑʋi:gət] -> [ˈfɑɾgəlɛgət]	Breakdown (Compound real word + affix)	1
	[ˈfɔɾhɑʋi:gət]> [ˈfɔɾgɑʋi:jət]	Place & Manner x2	1
	[fɔɾhɑ] -> [fɔɾɑ]	Cluster reduction	1
Konilsedi	[ˈku:nɪlsədɪ] -> [ˈku:nəsɛdəl]	Cluster reduction, Vowel + Reduction (real word x2)	1
	[ˈku:nɪlsədɪ] -> [ˈkusədɛlɪ]	Syllable substitution (real word), breakdown	1
Betivese	[ʋəsə] -> [ʋəlsə]	Addition (cluster creation; real affix)	**3**
	[bəˈti:] -> [bɑˈti:]	Vowel	1
	[bəˈti:] -> [bəˈtʏ]	Vowel	1
Låtterite	[tɛɾɪtə] -> [tɛɾmə]	Reduction x2, Addition	1
	[lɔˈtɛɾɪtə] -> [ɾɔˈtɛɾɪtə]	Manner	1
	[lɔˈtɛɾɪtə] -> [ɾɔˈtɛli:təɾ]	Manner x2, Additionx2	2
	[tɛɾɪ] -> [tɛlə]	Syllable deletion, Place & Manner	1
	[tɛɾɪ] -> [tɛlɪ]	Manner	1
	[lɔ] -> [nɔ]	Manner	1
	[lɔ] -> [hɔ]	Place & Manner	1
	[lɔˈtɛ] -> [lɔɾˈtɛ]	Addition(cluster creation)	1

Note: N refers to the number of participants making the exact same error.

Appendix M – Participant Consent Form for Wordlikeness Measure Participants

UNIVERSITETET I BERGEN
Det psykologiske fakultet

Forespørsel om deltakelse i forskningsprosjektet

«Pilotstudie av en norsk språkspesifikk nonord-repetisjonsoppgave»

Bakgrunn og formål

Formålet med denne pilotstudien er å utvikle et verktøy som kan i fremtiden brukes innenfor kartleggingen av både enspråklige og flerspråklige barn med språkvansker. Testen består av nonord (ord som ikke har noen iboende betydning) som barnet skal repetere. Dette prosjektet er en del av et mastergradsstudium i logopedi. Mastergraden gjennomføres ved institutt for biologisk og medisinsk psykologi ved Universitetet i Bergen. For å forsikre at disse nonordene kan brukes slik det er ønsket, så behøves det voksne deltakere med norsk morsmål som kan hjelpe med å skåre disse ordene. Kunnskap om hvor norske eller «unorske» nonordene høres ut vil hjelpe med å forstå resultatene som vil komme frem når testsettene blir testet med barn senere.

Hva innebærer deltakelse i studien?

Deltakelse i studien vil foregå ved at du og forskeren møtes over et internettintervju, hvor nonordene blir presentert for deg en etter en. Du kan bruke så lang tid du vil på skåringen, og nonordene kan også bli repetert så mange ganger som ønskelig. Din oppgave er da å skåre ordene på en skal fra 1 til 5. 1 betyr at ordet ikke høres norsk ut i det hele tatt, og 5 betyr at ordet høres ekstremt norsk ut. Det vil ikke bli gjort noen lydinnspilling, ettersom forskeren noterer underveis. Det er usikkert hvor mange nonord det vil være, ettersom det i noen tilfeller vil kreve flere variasjoner. Selve intervjuet vil ikke vare lenger enn ca. 10 minutter.

Hva skjer med informasjon om deg?

Alle opplysninger vil bli behandlet konfidensielt, og alle involverte vil være underlagt taushetsplikt. Alle deltakere blir identifisert med en koblingsnøkkel, nettopp for å opprettholde anonymitet, men også for å kunne slette data på individuelle deltakere om nødvendig. Det vil ikke registreres noen personopplysninger utenom hvor i Norge du er fra. Prosjektet skal avsluttes i juni 2021. Data lagres på en trygg forskningsserver ved Universitet i Bergen, for å forsikre at ingen som ikke jobber direkte med prosjektet har tilgang til data som angår deltakere. Prosjektet har blitt meldt til Norsk senter for forskningsdata (NSD) og Regionale komiteer for medisinsk og helsefaglig forskningsetikk (REK).

Frivillig deltakelse

Det er frivillig å delta i studien, og du kan når som helst trekke ditt samtykke angående din deltakelse i studien uten å oppgi noen grunn. Som tidligere nevnt, vil hver deltaker bli identifisert internt med en koblingsnøkkel, så om dere ønsker å trekke dere, kan alt datamateriale slettes.

Kontakt:

Dersom du har spørsmål til studien, vennligst ta kontakt med:

Student

Martin Alexander Gulbrandsen Mail: martin.gulbrandsen@student.uin.no, Tlf: 45381885

Prosjektveileder

Jan de Jong Mail: jan.jong@uib.no

Samtykke til deltakelse i studien «Pilotstudie av en norsk språkspesifikk nonord-repetisjonsoppgave»

Jeg har mottatt informasjon om studien, og ønsker å delta.

Prosjektdeltakers navn, dato/sted

Appendix N – Participant Consent Form for Parents of NWR Participants

UNIVERSITETET I BERGEN
Det psykologiske fakultet

REKnr. 204308 dato: 22.02.2021

Forespørsel om deltakelse i forskningsprosjektet

«Pilotstudie av en norsk språkspesifikk nonord-repetisjonsoppgave»

Bakgrunn og formål

Formålet med denne pilotstudien er å utvikle et verktøy som i fremtiden kan brukes innenfor kartleggingen av både enspråklige og flerspråklige barn med språkvansker. Testen består av nonord (ord som ikke har noen iboende betydning) som barnet skal repetere. Dette prosjektet er en del av et mastergradsstudium i logopedi. Mastergraden gjennomføres ved institutt for biologisk og medisinsk psykologi ved Universitetet i Bergen. Ettersom denne studien er det første steget etter konstruksjonen av nonordene så er det førskolebarn og førsteklassebarn i 5-7 alderen med norsk som morsmål og *ingen kjente språkvansker* som kan bidra til å videreutvikle testen. Deltakelse i studien betyr ikke at barnet har vansker med språket, og fokuset her er på om nonordene i seg selv er gode nok til at barn uten språkvansker sjeldent gjør feil. Formålet er ikke å «diagnostisere» barna som tar del i dette studiet, ettersom det er kvaliteten av nonordene som er fokuset. Vi håper dere ønsker å gi samtykke til at deres barn deltar i dette prosjektet.

Hva innebærer deltakelse i studien?

Vi vil undersøke barns evne til å repetere ord de aldri har hørt før, men som har flere egenskaper som man kan kjenne igjen i det norske språket. Testen er utformet som en video, hvor en trollmann ber barna om hjelp med å trylle stjernene tilbake på himmelen. Han ber de da repetere etter seg, og etter en kort pause så vil stjernene komme frem på himmelen. Den består av 40 ord, og vil ta ca. 10 minutter å gjennomføre. Det er nødvendig med lydopptak av deltakerne, slik at vi kan forsikre oss om at datamaterialet stemmer i ettertid. På denne måten slipper også barnet å bli distrahert av at noen sitter og skriver notater. Ettersom testen går ut på at barnet repeterer det de hører, så er det selvfølgelig også viktig at de ikke har noen kjente hørselsproblemer. Alle barn er forskjellige, og det forventes at noen kan bli distraherte eller ikke ønsker å repetere noen av ordene han sier, og dette er helt greit. Datamaterialet som vi sitter igjen med etter testperioden vil kunne opplyse oss om kvaliteten av nonordene i testen, og om det er mulig å bruke den videre for da å teste barn *med* utviklingsmessige språkvansker. Det er igjen viktig å understreke akkurat dette at utprøvingen av testen ikke vil føre til noen diagnostiske konklusjoner som kan knyttes til det enkelte barnet, ettersom testen ikke har noen diagnostisk verdi på dette stadiet. Det er også viktig at barnet ikke får ekstern hjelp under testingen, men de må gjerne oppmuntres om det er behov for det.

Hva skjer med informasjon om barnet?

Alle opplysninger vil bli behandlet konfidensielt, og alle involverte vil være underlagt taushetsplikt. Alle deltakende barn blir identifisert med en koblingsnøkkel i studien, nettopp for å opprettholde anonymitet, men også for å kunne slette data på individuelle deltakere om nødvendig. Personopplysninger utenom alder og kjønn vil ikke bli registrert, slik at lydopptakene forholder seg anonymiserte. Prosjektet skal avsluttes i juni 2021. Etter lydopptak har blitt brukt for å forsikre seg om at datamaterialet er komplett, så vil de destrueres. Ettersom barna blir en del av en gruppe, og ingen informasjon utenom alder og kjønn er kjent, så vil de heller ikke kunne gjenkjennes i en eventuell publikasjon. Data lagres på en trygg forskningsserver ved Universitet i Bergen, for å forsikre at ingen som ikke jobber direkte med prosjektet har tilgang til data som angår deltakere. Prosjektet har blitt meldt til Norsk senter for forskningsdata (NSD) og Regionale komiteer for medisinsk og helsefaglig forskningsetikk (REK). NSD bedømte at det ikke var nødvendig med godkjennelse, og REK godkjente prosjektet (REKnr. 204308). Skrivet om godkjennelse kan deles om det er ønsket.

Frivillig deltakelse

Det er frivillig å delta i studien, og dere kan når som helst trekke deres samtykke angående barnets deltakelse i studien uten å oppgi noen grunn. Som tidligere nevnt, vil hver deltaker bli identifisert internt med en koblingsnøkkel, så om dere ønsker å trekke dere, kan alt datamateriale fra barnet bli slettet.

Kontakt:

Dersom dere har spørsmål til studien, vennligst ta kontakt med:

Student

Martin Alexander Gulbrandsen Mail: martin.gulbrandsen@student.uib.no, Tlf:453 81 885

Prosjektveileder

Jan de Jong Mail: jan.jong@uib.no

Samtykke til deltakelse i studien «Pilotstudie av en norsk språkspesifikk nonord-repetisjonsoppgave»

Jeg har mottatt informasjon om studien, og ønsker at mitt barn skal delta.

Prosjektdeltakers navn og fødselsdato

Signert av prosjektdeltakers foresatte, dato/sted

Appendix O – Information Sheet Adapted for Children

UNIVERSITETET I BERGEN

Herme etter tulleord

HVORFOR BLIR DU SPURT OM Å VÆRE MED?

Det er mange barn som syns at å si og skjønne ord er vanskelig. Kanskje du kjenner noen som syns det er vanskelig? Denne leken som jeg lurer på om du vil være med på er laget for å hjelpe akkurat de som syns det er vanskelig. Du blir derfor spurt om du vil være med på å prøve ut denne leken.

HVA VIL SKJE DERSOM DU DELTAR?

Det du skal gjøre er å få se på en video av en trollmann som gjør tryllekunster. Da kommer han til å si magiske ord, og da må du herme etter han for at magien skal virke. Videoen er ikke lang, og du må gjerne ha med deg en voksen. Enten så møtes vi på et rom sånn at du kan se på videoen på en storskjerm, eller så møtes vi over nettet og du får se videoen på PCen eller iPaden. Det er mange andre barn som også skal være med å hjelpe, men de skal ikke være med når vi ser på videoen.

HVA VIL SKJE DERSOM DU IKKE DELTAR

Det er du som bestemmer om du har lyst til å være med eller ikke, og selv om du har sagt ja så kan du si nei etterpå. Det er ingen andre som får høre på opptaket av deg, og det blir slettet senere.

Appendix P – REK Approval

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK vest	Camilla Gjerstad		22.02.2021	204308
			Deres referanse:	

Jan de Jong

204308 En norsk test for gjentakelse av nonord

Forskningsansvarlig: Universitetet i Bergen

Søker: Jan de Jong

Søkers beskrivelse av formål:

Dette prosjektet har som mål å skape en norsk nonord repetisjonstest (Non-word repetition test, NWRT), samt pilotere testadministrasjon. Det endelige målet vil være at denne testen skal brukes sammen med andre tester for å vurdere potensielle språkvansker (developmental language disorder, DLD) blant små barn, eller hjelpe til med å vurdere dysleksi hos voksne. Å lage en annen måte å vurdere tale- og språkvansker på, som er avhengig av andre aspekter av språket vårt, er ment å være et hjelpemiddel i en situasjon der det er best egnet. Det er for tiden ingen norske NWRT-er som tar for seg denne testmetoden på en slik dybde. Flere testbatterier har en kort deltest som omhandler nonord, men disse er ofte forholdsvis rudimentære eller for å kartlegge evne til å identifisere minimale par.
Dette prosjektet har som mål å konstruere men også å pilotere testen, Prosjektet handler likevel ikke så mye om deltakerne som den gjør selve testsettet. Prosjektet er en pilot som forsøker å verifisere om nonordrepetisjonsoppgavene forårsaker en takeffekt blant deltakerne. Målet og fokuset i dette studiet er derfor ikke resultatene til de enkelte deltakerne, men derimot å finne ut av om enkelt stimuli har en uforventet vanskelighetsgrad.

REKs vurdering

Vi viser til din tilbakemelding mottatt 16.02.21 for ovennevnte forskningsprosjekt. Tilbakemeldingen er behandlet av komiteleder for REK vest på delegert fullmakt fra komiteen, med hjemmel i forskningsetikkforskriften § 7, første ledd, tredje punktum. Søknaden er vurdert med hjemmel i helseforskningsloven § 10.

REK vest ba om tilbakemelding (brev av 02.02.21)
En ba om at revidert forskningsprotokoll og informasjonsskriv ble sendt til REK vest.

Tilbakemelding fra prosjektleder
Det er vedlagt revidert versjon av forskningsprotokoll, samt informasjon/samtykkebrev til voksne. Det er også vedlagt et samtykkebrev som er ment for voksne deltakere som skal vurdere ordlikhet. Det forventes ikke at aspektet som inkluderer voksne deltakere har behov for godkjenning av dere, men det er valgt å inkludere denne informasjonen uansett. Det originale skrivet til barna er også vedlagt, da det ikke ble bedt om endring.

Vurdering av tilbakemeldingen
REK vest ved komitéleder har vurdert tilbakemeldingen. Vi ber om at informasjonsskrivet til de foresatte merkes med logoen til den forskningsansvarlige (Universitetet i Bergen), REKnr. 204308 og dato.

Det fremgår av tilbakemeldingen at studien også vil inkludere voksne deltakere som skal vurdere ordlikhet. Det er uklart for REK vest hva denne undersøkelsen går ut på og hva søker mener med at "*Det forventes ikke at aspektet som inkluderer voksne deltakere har behov for godkjenning av dere*." REK vest gjør oppmerksom på at forskning på voksne deltakere skal også vurderes av REK, dersom undersøkelsen kan frembringe ny kunnskap om helse og sykdom. Dersom det skal gjøres vesentlige endringer i studien, må det sendes en endringsmelding til REK vest.

Vedtak

Godkjent

REK vest har gjort en helhetlig forskningsetisk vurdering av alle prosjektets sider. Prosjektet godkjennes med hjemmel i helseforskningsloven § 10.

Med vennlig hilsen

Marit Grønning
Professor dr.med.
komiteleder REK vest

Camilla Gjerstad
rådgiver

Sluttmelding
Søker skal sende sluttmelding til REK vest på eget skjema senest seks måneder etter godkjenningsperioden er utløpt, jf. hfl. § 12. Dersom prosjektet ikke igangsettes eller gjennomføres skal prosjektleder også sende melding om dette via sluttmeldingsskjemaet.

Søknad om å foreta vesentlige endringer
Dersom man ønsker å foreta vesentlige endringer i forhold til formål, metode, tidsløp eller organisering, skal søknad sendes til den regionale komiteen for medisinsk og helsefaglig forskningsetikk som har gitt forhåndsgodkjenning. Søknaden skal beskrive hvilke endringer som ønskes foretatt og begrunnelsen for disse, jf. hfl. § 11.