LANGUAGE ASSESSMENT IN MULTILINGUAL CHILDREN

The Non-word Repetition Task

M.M.C. van der Linden
3035433

Master Thesis
'Linguistics: The Study of the Language Faculty'
Utrecht University
2008

Supervisors: dr. Elise de Bree, dr. Ellen Gerrits
Independent reader: prof. dr. Wim Zonneveld

TABLE OF CONTENTS

ABSTRACT

Language assessment in multilingual children is problematic, due to several biases and inaccuracies in traditional norm-referenced knowledge-dependent language measures. There is a need for a task that is able to distinguish between a language *difference* caused by second language learning, and a *specific language impairment* (SLI). It was previously shown that the non-word repetition task (NRT) may be able make this distinction (e.g., Archibald, 2008; Campbell et al., 1997; Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer et al., 2000; Girbau & Schwartz, 2008). The present study examines whether the NRT is suitable as a screening tool to detect language disorders in multilingual children from diverse linguistic backgrounds in The Netherlands. A non-word repetition task was adapted specifically for this purpose. In the first study, the *screening study*, multilingual children were screened for language problems with the NRT together with a traditional knowledge-based language screening tool. Multilingual children with language problems performed significantly worse on the NRT compared to typically developing multilinguals and monolinguals. A follow-up of the language status (two years after testing) of the multilingual children in the language impaired group revealed that the NRT performance on four-syllable non-words is a more accurate measure than the knowledge-based language screening tool. The second study, the *assessment study*, examined NRT performance in typically developing multilingual children with children diagnosed with SLI. The results of the present study confirmed the hypothesis that the NRT is able to discriminate between multilingual children with and without language impairment. In the present study, specifically the performance on four-syllable non-words was an accurate measure to identify language impairment. The results confirmed that the NRT is a culturally fair method to identify language disorders, and therefore, the NRT is a promising screening tool in the multilingual population. However, the findings indicate that the performance on the NRT *alone* is never sufficient to rule in or rule out language impairment. Future research is necessary in order to obtain solid reference data on NRT performance by multilinguals. It may be suggested that mainly four-syllable words are used.

ACKNOWLEDGEMENTS

My gratitude goes to my supervisors Ellen Gerrits and Elise de Bree. Ellen, thank you for inviting me into this interesting research project. Due to the long distance between Maastricht and Utrecht, the supervision mainly involved *e-coaching*. Your quick and to-the-point replies to my questions or comments were very valuable. Elise de Bree was the one who had put me into contact with Ellen. Elise, thanks for your guidance, involvement, and enthusiasm. It helped me through the sometimes very boring *laatste loodjes*. Professor Wim Zonneveld started a chain by introducing me to Elise one year ago, and was willing to take the role as independent reader of this thesis. Thank you so much for that.

Additionally, I want to thank my parents for supporting me in going to 'school' again. And a very big thank you to Tim for being there for me (and more).

1. GENERAL INTRODUCTION

Children with Specific Language Impairment (SLI) have a significant limitation in language abilities, in the absence of deficits that usually accompany language learning problems – such as hearing impairment, low non-verbal IQ, and neurological damage (Leonard, 2002). Research has shown that approximately 7% of all children has a Specific Language Impairment (Tomblin *et al.*, 1997). Suffering from SLI has major consequences for a child because it is not able to communicate adequately with the people in his environment, which may cause social-emotional and academic problems. Children with SLI need speech language therapy, and often need to go to a special school. Therefore, it is very important to identify SLI at a young age.

The group of children enrolled in special schools for children with severe speech and language problems is heterogeneous with respect to their cultural, socio-economic, and linguistic background. Around 50% of all primary school children in the large cities in The Netherlands are (children of) immigrants (i.e., '*allochtoon*') (Aarts *et al.*, 2004), a proportion that is also represented in the group of children with SLI. 'Allochtone' children are often bilingual, learning Dutch as a second language at an older age, often not before school age.[1] Identifying SLI in multilinguals is a substantial problem. The cultural, socio-economic, and linguistic factors can have major influences on the performance on traditional norm-referenced language tests, since these tasks rely to a great extent on previous experience and world knowledge (Campbell *et al.*, 1997). These language measures are designed to be used for Dutch monolinguals, and cannot simply be applied to children who were not exposed to Dutch language and/or culture as much as monolinguals.

Due to the increasing number of multilingual children in The Netherlands, there is a growing need for a task that can discriminate between children whose language problems are caused by a difference in experience and knowledge (a *second language problem*, i.e., a language *difference*), and children having SLI (i.e., a language *disorder*). It has already been shown that *non-word repetition* is able to make such a distinction (e.g., Archibald, 2008; Campbell *et al.,* 1997). Its clinical application in The Netherlands has not yet been addressed directly. Therefore, this thesis focuses on the question:

'Is non-word repetition suitable as a screening tool to detect language disorders in multilingual children from diverse linguistic backgrounds in The Netherlands?'

[1] In some cases, 'allochtone' children are *trilingual*, speaking three languages. There are also (rare) cases of children speaking four languages. Therefore, in this thesis I will refer to these children as 'multilinguals' – a term that covers all the possible instances.

This thesis is part of the research project *The Non-Word Repetition test (NWR-test)*, initiated in 2004 by Ellen Gerrits (*Maastricht University Medical Center*) in collaboration with *GGD Zuid-Limburg* in Maastricht, and the Speech and Hearing Centre in The Hague (Gerrits, 2004).[2] The aim of the project is to adapt, standardise, and validate a non-word repetition task specifically for young children from diverse language backgrounds. It contributes to the currently more and more debated topic of language assessment of multilingual children in The Netherlands. This thesis will discuss the experiments conducted in the project *The Non-Word Repetition test*.

The content of this thesis is as follows. Chapter 2 addresses the general background rationale of the project and this thesis. A concise description of Specific Language Impairment will be provided. Multilingualism and multiculturalism in the Dutch context will be discussed very briefly. Specifically, the language problems in multilingual children will be discussed, as well as the problems that can occur in identifying language disorders in multilingual children with traditional norm-referenced language measures. Solutions for these problems will be discussed, with the focus on one specific processing-dependent assessment procedure: *non-word repetition*. The characteristics of non-word repetition will be discussed, concentrating on non-word repetition in children with SLI, and in multilingual children. Chapter 2 will create the theoretical framework for the experiments that were conducted for this study. The design, method, and the results of two experiments will be discussed in chapters 3 and 4. Chapter 5 will deal with the general discussion, conclusion, and suggestions for future research.

[2] The project was funded by FENAC (Federatie Nederlandse Audiologische Centra, Stichting Methodiekontwikkeling en Deskundigheidsbevordering taal- en spraakdiagnostiek bij kinderen, '*Dutch Federation of Speech and Hearing Centres*').

2. THEORETICAL BACKGROUND

2.1 Introduction

This chapter addresses the general rationale of the current study. It comprises four sections. In section 2.2, characteristics and possible causes of Specific Language Impairment (SLI) will be discussed. Section 2.3 addresses multiculturalism and multilingualism in the Dutch context. Specifically, the problems with respect to identifying SLI in multilingual children will be discussed. The importance of the development of alternative measures for language assessment of multilingual children will be addressed, since traditional language tests typically seem to fail in identifying SLI in multilinguals accurately. Section 2.4 discusses one of the alternatives: a processing-dependent measure, namely, the *non-word repetition task*. The characteristics of non-word repetition in children with SLI and in multilinguals will be discussed. Specifically, the clinical utility of non-word repetition in identifying SLI in multilingual children will be considered. Once the theoretical framework has been created, section 2.5 will address the questions and aims of the present study.

2.2 Specific language impairment

Children with Specific Language Impairment (SLI) show a significant limitation in language abilities, despite developing normal in all other areas (Bishop, 2006; Leonard, 2002). The prevalence of SLI in kindergarten children is approximately 7% (Tomblin *et al.*, 1997). Children with SLI usually have problems in several domains of language. The mean length of uttered sentences is often shorter than in typically developing children, and sentences are often less complex. Children with SLI also often show phonological restrictions, problems with lexical storage and access, and pragmatic problems (De Jong, 1999).

Leonard (2002) describes characteristics of SLI in different types of languages and concludes that even *if* there is a universal feature of SLI – apart from generally slow and poor language learning – it is well hidden. However, there is a large amount of evidence that children with SLI have a larger deficit in morphology and syntax than they do in other domains of language. In Dutch children with SLI, the language output shows agreement errors, word order errors, and omission of articles (Leonard, 2002).

The consequences of childhood SLI are substantial. Due to the language problems, the child is not able to communicate adequately with the people in his or her environment. This can cause, for example, social-emotional and academic problems. Children with SLI need speech language intervention, and often go to special schools fitted for children with severe speech and language problems.

The cause of SLI is still a hotly debated topic. Several theories of SLI have been proposed. Roughly speaking, there are two approaches explaining the language problems in SLI: a linguistic approach, and a processing approach. The *linguistic approach* claims that a specific deficit in grammar underlies SLI. It is assumed that the language impairment is caused by incomplete knowledge of particular language rules, principles or constraints (Leonard, 2002). For a review of several linguistic theories, see De Jong (1999), and Leonard (2002). Currently, most researchers choose a *processing approach* when explaining the language problems in SLI (see Leonard, 2002, for a review of several processing theories). The processing approach claims that SLI is caused by a deficit in (speech)-processing capacity. The advantage of using a processing-based theory is that it can account for the broad range of deficits in linguistic domains, while specific *grammatical* theories cannot explain, for example, limitations in lexical access and storage, or pragmatic problems. It may be the case that children with SLI either have a limitation in general processing capacity or a processing deficit in specific mechanisms, e.g. phonological memory, or temporal processing, which leads to language learning problems. For instance, a deficit in phonological short-term memory can have major consequences for novel word learning, since word learning requires phonological storage (Gathercole & Baddeley, 1990). The presence of a processing deficit underlying SLI is supported by a substantial amount of evidence. It has, for example, previously been shown that children with SLI typically show poor performance on processing tasks such as *non-word repetition* (e.g., Archibald, 2008; Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000). This specific topic will further be addressed in section 2.4.

In the Netherlands, most kindergarten children are screened for a language delay with the TSI ('Taalscreeningsinstrument', *Language Screening Tool*). It is a norm-referenced language screening tool used to detect language problems in Dutch 5-year-olds. The application of the TSI is a public health activity provided by the GGD ('Gemeentelijke Gezondheidsdienst', *Public Health Service*), and it is administered by speech therapists visiting the schools.[3] The TSI includes elements such as general questions (e.g., "What is your name?"), language comprehension (e.g., "Name all things that can drive"), morphology (plurals), word and sentence repetition, and reasoning (e.g., "Why should you not play with matches?").

[3] In some parts of The Netherlands, this language screening is not offered or different tools are used.

The diagnosis of SLI is accomplished by administering norm-referenced language tests developed for Dutch, supplemented by tests that can exclude the presence of for example hearing impairment, or a low non-verbal IQ. The group of children with SLI is a heterogeneous group with respect to linguistic, socio-economic, and cultural background. The heterogeneity of the population makes it hard to develop broadly applicable language assessment instruments (Archibald, 2008). The problems concerning language assessment in multilinguals is currently becoming more and more prominent due to the growing multicultural, thus multilingual, population in The Netherlands. In the following sections, this issue will be addressed.

2.3 Multiculturalism and multilingualism

2.3.1 Introduction

According to CBS statistics, there are 16.405.399 citizens in The Netherlands in 2008. Of this population, 3.215.416 people are part of a minority group of (children of) immigrants (‘*allochtonen*’) (CBS, 2008).[4] The group of immigrants in The Netherlands is still growing. Multiculturalism is most common in the four largest cities of The Netherlands: Amsterdam, Den Haag, Rotterdam, and Utrecht. An overview of the frequency of the five largest languages spoken by ‘allochtone’ children in primary schools in one of the largest cities (The Hague), ranked in decreasing order, is provided below in table 1.

Table 1. *Languages spoken by children in primary schools in The Hague. Taken from Aarts, et al., 2004 (pg. 200).*

	Language	Frequency
1	Turkish	4.798
2	Hind(ustan)i	3.620
3	Berber	2.769
4	Standard Arabic	2.740
5	English	2.170

‘Allochtone’ children are often bilingual, learning Dutch as a second language at an older age, often not before they reach school age (i.e., age four). Section 2.3.2 discusses the language characteristics of multilingual children.

[4] Centraal Bureau voor de Statistiek, ‘*Statistics Netherlands*’ (http://www.cbs.nl , last checked 4 November, 2008).

2.3.2 Language in multilingual children

Two different types of multilingualism can be distinguished: *simultaneous* and *successive* second language learning. In the first case, the child is brought up with two (or more) languages from the very first moment, which means that it has two (or more) 'mother tongues' with balanced language proficiency in the languages. In the latter case, the child starts learning one language and at a later age another language is added. Usually, the proficiency in the two languages is not balanced. Language abilities in the second language can be influenced by several factors, for example: age, intelligence and language talent, personality and learning style, social-cultural factors, and language contact and language input (Appel & Vermeer, 2000).

Most minority children start learning Dutch as a second language at an older age in school, and in contact with their peers (Steenge, 2006), and are therefore *successive* language learners. Children learning Dutch as a second language typically show a lag in several language domains. There can be phonological problems, e.g. in pronunciation of phonemes that are not present in the first language. There can be absence of morphological marking on verbs and nouns. Specifically, irregular morphological marking can be problematic for second language learners. Produced sentences may be shorter, and less complex than in native speakers of Dutch. The second language learning child may have problems with word order. Also, comprehension of complex sentences may cause problems. Typically, second language learners have a small vocabulary. In some cases, pragmatic language problems are present (Appel & Vermeer, 2000). Although these language problems resemble the language problems present in children with SLI, they are not caused by language *impairment* per se. The language output of multilingual children may reflect either slow language learning, or first-language interference on second language learning, or an underlying (specific) language impairment, or a combination of these (Archibald, 2008).

When assessing language proficiency in a multilingual child, it is important to determine whether the child has a language *difference* caused by the fact that it is learning Dutch as a second language, or whether it suffers from SLI (i.e. a language *disorder*). However, traditional norm-referenced language tests are not able to make this distinction. Little is known about multilingual children with SLI, which makes diagnosis and intervention problematic in this population (Steenge, 2006).[5] The following section will discuss this issue.

[5] The Amsterdam University is currently running a project ('BISLI'), aiming at disentangling problems stemming from bilingualism and SLI. For more information on this project , I refer to Jan de Jong's homepage http://home.medewerker.uva.nl/j.dejong1/(last checked 4 November, 2008).

2.3.3 SLI in multilingual children

The group of multilingual children in special schools for children with SLI is substantial, and still growing. Assessment of SLI in multilingual children is problematic, since traditional norm-referenced tests for Dutch monolinguals cannot simply be applied to children who were not exposed to Dutch language as much as monolinguals.

In the group of multilingual children, biased and inaccurate language assessment may lead to *under-* or *overidentification* of SLI – respectively, 'failure to identify SLI, when it is in fact present', and 'identifying SLI in non-impaired children'. The problems of a multilingual child with SLI may be ignored due to the fact that his or her slow language development is regarded as a consequence of learning a second language. This causes late referral for language intervention (Salameh *et al.*, 2002). It is desirable for a test to overidentify rather than to underidentify, so that every child that needs intervention will receive intervention. However, prominent overidentification will result in an overrepresentation of non-impaired children in speech/language therapy or in special schools. Putting children into intervention programmes and/or special schools is costly, in financial, as well as in human terms. Therefore it can be assumed that large overidentification must be avoided (Washington & Craig, 2004).

Under- and overidentification of SLI in multilingual children can be caused by several factors. The three best known biases in norm-referenced tests are content bias, linguistic bias, and disproportionate representation in normative samples (Laing & Kamhi, 2003). C*ontent bias* occurs when the content and method of a language test is based on the assumption that every child has been exposed to the same concepts and vocabulary or had the same life experience. Dutch language tests are created with words and concepts that are familiar in Dutch culture, which multilingual/-cultural children may not be acquainted with. For example, a child from Japanese origin may not be acquainted with the concept 'windmill' – a very common object in The Netherlands. There may also be a cultural difference in interaction patterns. For example, in some cultures the procedure pointing to objects might be uncommon, although that typically is the expected interaction during a test (Laing & Kamhi, 2003). Unfamiliarity with this knowledge may lead to overidentification of SLI. A *linguistic bias* can be caused by differences between (a) the language or dialect used by the examiner, (b) the language or dialect used by the child, and (c) the language or dialect that is expected to be used by the child. This may lead to either overidentification when the examiner attributes 'errors' to dialect differences. Underidentification can occur when the examiner attributes differences in performance to dialect, while they are in fact errors. A third possible problem is *disproportionate representation in normative samples*. Traditionally, norm-referenced tests do not include multilingual or multicultural children in their normative samples. Some tests (e.g. the English 1981 version of the Peabody Picture Vocabulary Test-Revised) did include data from children from minority groups to their normative groups; however this may do nothing more than decreasing the mean distribution of the normative sample (Laing & Kamhi, 2003). For the

Dutch language test '*Taaltoets Alle Kinderen*' ('*Language Test for All Children*') (Verhoeven & Vermeer, 2006) normative data for (Dutch) Turkish, Moroccan and Surinam children are provided. It does, however, not account for the amount of time that the child spent in The Netherlands, while this factor may have major influences on the language proficiency.

In The Netherlands, the discussion on accurate and unbiased language assessment of multilingual children is becoming more and more prominent.[6] Traditional norm-referenced language tests are not clear-cut for multilinguals, and a solution for this problem is necessary. In the following section, a discussion of possible alternative assessment procedures will be provided.

2.3.4 Alternative assessment procedures of language impairment

There are several alternative procedures that can be used to assess language abilities in multilingual children, for example: sampling and analysing of spontaneous language data with an interpreter, ethnographic interviewing techniques, dynamic assessment, and processing-dependent measures. Spontaneous language samples of the child can be judged by an interpreter together with the clinician in order to determine whether the language development is also delayed or disordered in the first language. Ethnographic interviewing with the parents (with an interpreter) can provide the clinician with additional information on the development of, and proficiency in the first language. Laing and Kamhi (2003) claim that language sampling and ethnographic interviewing should always be part of the assessment procedure in a multilingual child. Ideally, this is also the case in The Netherlands – but unfortunately it is not possible in all cases due to time restrictions and financial limitations. Depending on the judgements of an interpreter has some other drawbacks. In most instances, this interpreter is not a skilled clinician or linguist, which makes his or her judgements possibly less reliable. Ideally, they should be able to understand linguistic terminology, and be able to translate language errors as accurately as possible. This is often not the case.

In their paper, Laing and Kamhi (2003) focus on the two other techniques: dynamic assessment and processing dependent measures. *Dynamic assessment* may be characterised as 'diagnostic learning': it involves the assessment of the child's ability to learn from guided assistance.[7] There are several different approaches to dynamic assessment. One is *test-teach-retest*, which involves assessment with a teaching step in between, and a check whether the child benefited from the support, and made sufficient progress in language or not. Low responsiveness to language training might indicate an underlying language deficit. Another variant of dynamic assessment is *task/stimulus variability*,

[6] For more information on this topic in the Dutch context, I refer to Manuela Julien's book 'Taalstoornissen bij meertalige kinderen' ('*Language impairments in multilingual children*').

[7] Traditional norm-referenced tests are *static assessment* measures, where the clinician is not allowed to provide assistance to the child during testing.

which involves modifying the task so that it relates more to the child's culture and previous life experience. For example, assessment in more naturalistic environments might provide a more appropriate measure of language proficiency, in stead of pointing to pictures or completing given sentences. With *graduated prompting*, the child receives systematic verbal cues during the task (ranging from least supportive to most supportive). For example, the clinician can check whether the child can be stimulated to use sentence structures not currently being produced. During the task, the clinician observes whether, and how, the child benefits from the prompts. This may indicate which linguistic forms and structures to target, and how much improvement can be expected in the child (Laing & Kamhi, 2003). It is evident that dynamic assessment is relatively time-consuming, and requires preparation.

Another important, relatively recent, alternative to the traditional knowledge-based measures are processing-dependent measures, such as digit span, working memory, and *non-word repetition*. Processing-dependent measures evaluate *language processing* abilities, rather than (language) *knowledge*. Children with SLI typically show poor performance on those tasks. Non-word repetition proved to be an accurate means to distinguish typically developing children from children with SLI (as will be discussed in section 2.4). Also, unlike dynamic assessment, language sampling, and ethnographic interviewing, it is a fast and easy method of language assessment. This thesis will focus on the non-word repetition task as a screening tool to identify SLI in multilingual children. The task will be discussed in the following section, with a focus on non-word repetition performance in children with SLI. More importantly, non-word repetition in multilinguals will be addressed, too.

2.4 Non-word repetition and SLI

2.4.1 Introduction

Non-word repetition requires the immediate recall of auditorily presented novel words (e.g., /ji'fot/ or /nu'pifat/).[8] It reflects an important mechanism in language-learning: learning the phonological form of a new word, which is one of the aspects of vocabulary acquisition (Archibald, 2008). In the (English) literature, several variants of non-word repetition tasks are mentioned. The two most familiar examples are the Children's Test of Non-word Repetition (CNRep) designed by Gathercole and Baddeley in 1996, and the Non-word Repetition Test (NRT) designed by Dollaghan and Campbell (1998).[9] There are several differences between these non-word repetition tasks. For

[8] Performance on non-word repetition is usually expressed in the number of words repeated correctly (raw scores) or the percentage phonemes repeated correctly. Most studies investigating non-word repetition performance have reported phoneme level scores because it is considered to be a 'richer' method of scoring.

[9] These two tasks were also translated for the Dutch language area, by De Jong and Van der Leij (1999) who created a Dutch version of the CNRep, and De Bree *et al.* (2007) who made a Dutch NRT.

example, with respect to wordlikeness: (syllables in) the non-words in the English CNRep contain words and affixes (e.g., ‘blonterstaping’), which makes these items *high-wordlike*. The wordlikeness of the items in the English NRT, however, is *low* (e.g., ‘naichoitauvub’).[10] Also, the articulatory complexity of the items is higher in the CNRep than in the NRT, caused by the presence of consonant clusters, such as /bl/ and /st/, and phonemes that are acquired relatively late in development, such as /r/. The NRT only includes early acquired phonemes, and no consonant clusters (Graf Estes *et al*., 2007).[11]

2.4.2 Non-word repetition performance in children with SLI

Several international studies showed that non-word repetition is a useful measure in distinguishing children with SLI from typically developing children, independent of IQ, prior knowledge of events, or language structures (e.g., Archibald, 2008; Campbell *et al.*, 1997; Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000). Individuals with SLI typically perform poorly on non-word repetition, even when they can produce all individual phonemes of the task (Bishop, 2006).

Gathercole and Baddeley (1990) assessed verbal memory skills with a non-word repetition task in typically developing children and children with SLI. The authors found that children with language impairment scored significantly more poorly on non-word repetition. There was an effect of non-word length: the longer the non-word, the more errors were made. This effect was largest in the children with language impairment. Dollaghan and Campbell (1998) showed that on the basis of non-word repetition performance, language impairment could be ruled in or out with a high level of accuracy. A knowledge-based language test on the other hand, was much less accurate, and failed in ruling out language impairment. Ellis-Weismer *et al.* (2000) tested non-word repetition performance (with Dollaghan & Campbell’s NRT) in a population based sample, and did not find as high an accuracy of non-word repetition as Dollaghan and Campbell (1998) did. A difference in the composition of the sample seems to play a large role in the difference between Dollaghan and Campbell’s findings and the results from Ellis-Weismer *et al.* (i.e., a larger group difference with respect to language proficiency may have boosted the differences on the performance on the NRT). From their analysis, the authors conclude that non-word repetition scores may provide assistance in

[10] The use of low-wordlike non-words allows the child to rely more on the phonological short-term working memory, while when repeating high-wordlike non-words there is support from the lexical knowledge. This implies that children with SLI (with a small lexicon) are not able to take advantage of high wordlikeness, while non-impaired children *can*. High word-like items are not equally unfamiliar to both groups. Ideally, a task is designed such that the influence of prior language knowledge is minimised.

[11] A low articulatory complexity of the items in a non-word repetition task is desired, so that poor performance cannot be attributed to an articulation deficit (which is a very common characteristic in children with SLI) (Dollaghan & Campbell, 1998). A high articulatory complexity could undesirably boost the difference between performance of typically developing children and children with SLI (Graf Estes *et al*., 2007), and should therefore be avoided.

ruling in and ruling out language impairment. However, NRT scores *alone* are not sufficient for ruling in and ruling out language impairment. Conti-Ramsden (2003) found that, although the ability of the CNRep to identify all *non-impaired* children correctly was high, the ability to identify all *impaired* children was relatively low. In general, the task was sufficiently accurate. However, when used together with a test of past tense marking, its predictive ability was higher.

To understand why children with SLI show poor non-word repetition performance, it is necessary to understand which underlying cognitive mechanisms are involved in successful non-word repetition. This issue will be further addressed in section 2.4.3.

2.4.3 What does non-word repetition measure?

At this moment there is still no definite answer to the question what a non-word repetition task exactly measures. It is difficult to pinpoint the underlying cognitive mechanisms necessary for non-word repetition, since non-word repetition involves a number of different steps (Archibald, 2008; Gathercole, 2006) (see figure 1):

- Hearing, perceiving, and segmenting the phonological form (*auditory processing*);
- Encoding and retaining the phonological representation (*phonological analysis* and *storage*);
- Planning, programming, and executing the output (*speech-motor planning and output*).

The *phonological learning* (i.e., not only *repeating*) of a new phonological form will take place after multiple exposures to the items (Gathercole, 2006).

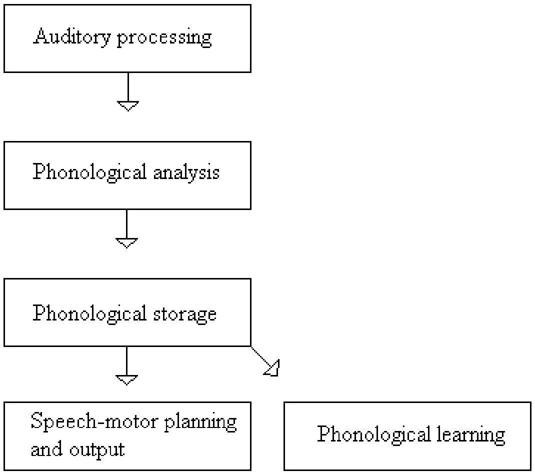

Figure 1*. Processes involved in non-word repetition, after Gathercole, 2006 (page 533).*

Children with SLI may be impaired in one or several of these areas, causing them to fail on non-word repetition tasks (Archibald, 2008). One important, and often used, interpretation may be that non-word repetition reflects the capacity of *phonological short-term memor*y (i.e., children with

SLI have a deficit in the *phonological storage stage*) (Gathercole, 2006). This may be concluded from the fact that longer non-words are more difficult to imitate correctly. Longer non-words take more time to perceive, analyse and imitate, leading to possible decay of their mental phonological representations before they can be imitated. This length effect is largest for children with SLI. Also, real words are easier to imitate than non-words, which implies that lexical knowledge supports retention (Archibald, 2008). In addition, performance on non-word repetition correlates to a great extent with scores on standard measures of short-term memory (such as *digit span* and *word list recall*). Children with SLI typically show poor performance on those immediate memory tasks (Archibald, 2008; Conti-Ramsden, 2001).

There are several other factors that may be the cause of the relationship between SLI and poor non-word repetition performance, such as *phonological awareness*, or a more *general processing factor* (Archibald, 2008). Archibald (2008) states that it may be that non-word repetition is such a clear-cut marker of SLI disregarding the heterogeneity of the population because it is constrained by *multiple processes*. No systematic research on the underlying mechanisms in non-word repetition has been carried out yet. Work remains to be done in order to identify the area of breakdown responsible for poor non-word repetition, and its implications for (the cause of) SLI.

2.4.4 Non-word repetition and multilingualism/-culturalism

Non-word repetition might be an effective (screening) tool to discriminate between typically developing children and children with SLI, not only in monolinguals but also in multilingual children (e.g., Archibald, 2008; Campbell *et al.*, 1997; Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000; Girbau & Schwartz, 2008). Campbell *et al.* (1997) assessed the performance on three processing-dependent tasks and one knowledge-dependent task in ‘majority’ (White) and ‘minority’ (African American; Asian; Native American) subjects. Compared to the majority subjects, minority subjects performed significantly more poorly than the majority subjects on the knowledge-dependent task (an oral language scale). However, the two groups did not differ significantly on the processing-dependent tests, which included a non-word repetition task (the NRT). These findings suggest that the non-word repetition task is a culturally fair method for screening children for SLI, since the task minimally depends on prior knowledge or experience (but see Oller *et al.* (2006) for a different view on the cultural fairness of processing-dependent tasks).

It should be noted that there seems to be a strong relationship between familiarity with a language and verbal short-term memory. This has been shown by Thorn and Gathercole (1999), who investigated measures of phonological short-term memory (Digit Span; CNRep) and vocabulary in English speaking and two groups English-French speaking children (simultaneous learners of English and French, and successive second language learners of French). Subjects were tested with an

English and a French version of the CNRep. The simultaneous bilinguals performed equally well on the English and French non-word repetition tasks, while the monolinguals performed significantly better than the successive second language learners. This made the authors suggest that performance on non-word repetition is closely related to the level of language-specific phonotactic knowledge.

Kohnert *et al.* (2006) examined the performance of language impaired English speaking, typical English speaking, and typical simultaneous bilingual English/Spanish speaking children on two language-based processing tasks, including the NRT, as created by Dollaghan and Campbell in 1998. The children in the typical English group performed best on both tasks, the bilingual children performed more poorly, and the language impaired children performed worst. The results suggest that performance on non-word repetition *is* dependent on previous experience, since there is a difference between mono- and multilingual's performance. Typically developing bilinguals may have shown poor non-word repetition because there may have been an influence from language-learning experience. Poor knowledge of the language-specific phonotactic patterns of the non-words could have affected the performance.

Similar to the issue that Thorn and Gathercole (1999) had already addressed, Kohnert *et al.* (2006) stated that there is a need for a non-word repetition task that can be used with linguistically diverse children. This idea was picked up by Girbau and Schwartz in 2008, who investigated non-word repetition performance in language impaired, and typical bilinguals (with Spanish as the first language, and speaking English as a second language), and language impaired, and typical monolingual Spanish speaking children. To avoid influence from poor phonotactic knowledge, a non-word repetition task was constructed following Spanish phonotactic patterns. Non-word repetition performance on this *language-specific* task appeared to be an accurate identifier of SLI in both the monolingual and the multilingual groups.

Several researchers (e.g. Archibald, 2008; Campbell *et al.,* 1997; De Bree *et al.*, 2007; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.,* 2000; Gray, 2003; Girbau & Schwartz, 2008; Washington and Craig, 2004) mention the possibility of using non-word repetition as a clinical tool that can assist in detecting SLI. Specifically, the application of non-word repetition in the multilingual population is mentioned (e.g. Campbell *et al*., 1997; Gerrits, 2005; Girbau & Schwartz, 2008). Although non-word repetition has been used in many studies showing that the task is a reliable identifier of SLI, its clinical application has not been addressed very often, and not in the Dutch context. In the following sections, the clinical utility of non-word repetition in discriminating multilingual children with SLI from typically developing multilingual children will be considered.

2.4.5 Clinical use of non-word repetition

2.4.5.1 Advantages of non-word repetition over traditional language tests

In her overview paper of NRT, Archibald (2008) provides several advantages of non-word repetition over traditional knowledge-dependent language tests. First, non-word repetition performance is less biased by experience and cultural environment. It does not rely on prior knowledge of events, or language structures, but tests the ability to process new information (Campbell *et al.*, 1997). This implies that non-word repetition can be used for assessing language in multilingual children.[12] Non-word repetition is also *less* dependent of vocabulary size than knowledge-based language measures, however, vocabulary size was shown to have some influence on non-word repetition performance (and/or vice versa) by, for example, Munson *et al.* (2005).[13] Second, Ellis-Weismer *et al.* (2000) found that there was only a weak correlation between the non-verbal IQ and non-word repetition performance, which implies that non-word repetition performance is largely independent of IQ. Third, it was found that the non-word repetition task can be completed by children as young as 2 years of age (Archibald, 2008). This may be interesting considering the importance of early diagnosis and intervention of children at risk for SLI.[14] Fourth, non-word repetition is a simple task to administer and score. It can be administered in less than 10 minutes. Scoring (during or after testing) is rather easy. Additionally, it was shown that the diagnostic efficiency of non-word repetition is rather good, as will be discussed in the following paragraph.

2.4.5.2 Diagnostic efficiency of the task

Any test for language impairment (actually, any impairment/disorder/disease) roughly speaking has two possible outcomes: 'positive' (*impaired*) or 'negative' (*non-impaired*). A perfect test, is able to identify all impaired children (it has a *sensitivity* of 100%). This perfect test should also label all non-impaired children as having no impairment (it is 100% *specific*). The levels of sensitivity and specificity taken together indicate the level of *accuracy* of the test. A test with a low accuracy identifies too many 'false positives' (i.e., overidentifications) and/or 'false negatives' (i.e., underidentifications). Previous studies have shown that non-word repetition tasks show a high level of accuracy in discriminating between children with and without SLI (e.g., Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000; Girbau & Schwartz, 2008).

[12] Possibly, not all variants of non-word repetition tasks are able to do so. See Dollaghan and Campbell (1998) for the requirements that such a task has to meet, with respect to, for instance, wordlikeness, and articulatory complexity.

[13] Two questions that may arise are: Is the vocabulary small because of the *poor phonological short-term memory*, or does the child with a small lexicon perform poorly on a non-word repetition task because of the *small vocabulary*? The two seem to be intertwined, and more research in this field is necessary in order to identify the cause of the relationship.

[14] In practice, it may be difficult to assess children this young.

Several methods are available for determining the diagnostic efficiency of a task, for example *likelihood ratio calculation* or *Receiver Operating Characteristic* (ROC) analysis. Most studies on non-word repetition reported likelihood ratios to indicate the diagnostic accuracy of the task (e.g., Dollaghan & Campbell, 1998; Ellis-Weismer et al., 2000; Girbau & Schwartz, 2008; Kohnert et al., 2006). In order to be able to compare the results of the present study with those of previous studies, likelihood ratios will be calculated. The method will be discussed briefly in the following section.

Likelihood ratios do not provide sensitivity, sensitivity, and accuracy levels directly, but do give an indication of the diagnostic accuracy of a task. A likelihood ratio expresses the odds that a certain score would be expected in a child with language impairment, versus the probability that it would be expected in a typically developing child (Kohnert et al., 2006). The likelihood ratio for a positive test result, defined as a specific threshold score ('cut-off point'), is calculated by dividing the true positive rate (the proportion of impaired individuals with a score of this specific level or lower) by the false positive rate (the proportion of non-impaired individuals with this specific score) (Dollaghan & Campbell, 1998). The aim is to identify the threshold scores that are associated with optimal levels of likelihood ratios (i.e., high for a positive test result, and low for a negative test result). A likelihood ratio of 20 or more is considered high for ruling in impairment, i.e., language impairment can be ruled in with a very small likelihood of error. The likelihood for a negative test result sufficient to rule out the presence of impairment is calculated for a specific score cut point by dividing the false negative rate (the proportion of impaired individuals with a score at or above this cut point) by the true negative rate (the number of non-impaired individuals with a score above this level). A likelihood ratio of 0.08 or less is considered sufficient for ruling out impairment (Dollaghan & Campbell, 1998). For more information on determining and interpreting likelihood ratios see Sackett *et al.* (1985).

In the literature on non-word repetition and language impairment, different cut-points are used for defining a positive or negative test result (i.e., respectively ruling in, and ruling out language impairment). An overview is provided in table 2.

Table 2*. Overview of optimal cut-points with likelihood ratios for ruling in and ruling out language impairment in different studies. LI = language impairment, SLI = specific language impairment, TLD = typical language development.*

Authors	**Task**	**Experimental groups**	***Positive test result***		***Negative test result***	
			cut-point	**likelihood ratio**	**cut-point**	**likelihood ratio**
Dollaghan & Campbell (1998)	NRT	English, LI vs. TLD	PPC ≤70%	25.15	PPC ≥81%	0.03
Ellis-Weismer *et al.* (2000)	NRT	English, SLI vs. TLD	PPC ≤70%	2.78	PPC ≥81%	0.66
		English, SLI vs. TLD Using 'extreme cut-points'.	PPC ≤60%	4.50	PPC ≥90%	0.43
Kohnert *et al.* (2006)	NRT	English LI vs. English TLD & Spanish-English TLD	PPC ≤72%	5.07	PPC ≥93%	0.08
Girbau & Schwartz (2008)	Spanish non-word repetition task	Spanish, SLI vs. TLD	Raw score ≤50%	11.00	Raw score >50%	0
		Spanish-English, LI vs. TLD	Raw score <33%	9.00	Raw score ≥33%	0.20

Although the first three studies mentioned in table 2 used the same task (the English NRT, created by Dollaghan and Campbell in 1998), the studies found quite different results. The differences between the studies may, for example, be caused by differences between the subjects (e.g., a smaller or larger deficit in language in the impaired groups). Using identical cut-points did not always lead to similar likelihood ratios, as is shown by the results of Dollaghan and Campbell (1998) and Ellis-Weismer *et al.* (2000). Given the desired level of a likelihood ratio to rule *in* language impairment (>20), only Dollaghan and Campbell (1998) obtained sufficient results. Dollaghan and Campbell (1998), Kohnert et al. (2006), and Girbau and Schwartz (2008), were able to rule *out* language impairment with a high degree of confidence (<0.08). Ellis-Weismer *et al.*'s likelihood ratios were respectively 'intermediate high' and 'intermediate low'. It can be concluded that poor NRT performance by itself is not sufficient to identify (S)LI, and further testing is *always* necessary to rule it out – as is the case for all language screening tools.

This thesis focuses on the application of non-word repetition in a multilingual population. The following sections will discuss the questions and aims of the current study.

2.5 Present study

2.5.1 Introduction

Traditional norm-referenced language tests to a great extent rely on previous experience and world knowledge. This is problematic since cultural, socio-economic, and linguistic background factors can have major influences on the performance on these tasks (Campbell *et al.,* 1997). The knowledge-based language measures are typically designed for Dutch monolinguals, and cannot simply be applied to children who were not exposed to Dutch language and/or culture as much as monolinguals. Biased and inaccurate language assessment may lead to under- or overidentification of SLI in the group of multilingual children. Thus, there is need for a task that can discriminate between children with a *second language learning problem* (i.e., a language *difference*) and children having SLI (i.e., a language *disorder*). Several studies have already shown that the *non-word repetition task* (NRT) is able to make this distinction (e.g., Archibald, 2008; Campbell *et al.*, 1997; Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000; Girbau & Schwartz, 2008). Compared to traditional language measures, the NRT relies less on prior knowledge of events, vocabulary, or language structures. Rather, it tests the ability to process new (linguistic) information. This implies that non-word repetition can be used for assessing language specifically for multilingual children.

The clinical application of the NRT in The Netherlands, however, has not yet been addressed directly. Gerrits (2004) initiated the project *The Non-Word Repetition Test*. The aim of this project is to adapt, standardise, and validate a non-word repetition task as a screening tool for language disorders, specifically for young children from diverse language backgrounds. I joined this project when data had already been collected, and transcriptions and error analyses were already made. Statistical analyses and review of the results still remained to be done.

2.5.2 Questions and aims of the present study

This thesis addresses the following central question:

'Is non-word repetition suitable as a screening tool to detect language disorders in multilingual children from diverse linguistic backgrounds in The Netherlands?'

This thesis examines the performance of multilingual and monolingual children with and without (specific) language impairments on non-word repetition. Specifically, the following research questions will be addressed.

a. Does the non-word repetition task (NRT) discriminate between monolingual and multilingual children without SLI?
b. Is the NRT a reliable clinical tool for discriminating between multilingual children with and without (S)LI?
c. Is the NRT's reliability greater than the 'Taalscreeningsinstrument' (TSI)?[15]

To answer these research questions, two experiments were conducted.
In the *Screening study*, the NRT will be used for 5-year-old multilingual children with and without language problems, and monolingual children without language problems, in combination with a knowledge-dependent, norm-referenced screening tool (the TSI).[16] The purpose of the study is to compare the clinical utility of the NRT with the TSI in distinguishing between multilingual kindergarten children with and without language impairment. This study aims at answering questions a, b, and c.

Results from previous studies imply that there will be a significant difference between multilinguals without language impairment and multilinguals with language impairment on the NRT, while multilinguals and monolinguals without language impairment are assumed to show similar performance on this task. It is expected that there will be an effect of non-word length, i.e., the longer the non-words, the worse the performance - due to processing demands (see e.g. Bishop, 1996). This effect is expected to be larger in the language impaired children. It is expected that there will be no significant difference between non-impaired multilinguals and monolinguals on the NRT, while there will be a significant difference between multilinguals with and without language impairment on the TSI. In order to determine the accuracy of the tasks, likelihood ratios will be calculated.

The *Screening study* consists of two parts. In the first part of the study ('*Screening study part A*'), children were divided into a language impaired and a non-impaired group based on the performance on the TSI. However, the TSI has an expected cultural and linguistic bias since this screening instrument has been developed for Dutch monolingual children. Also, the TSI is a screening tool, and is thus expected to identify a large number of *true* positives, but also a large number of *false* positives – further assessment will have to determine whether the children failing on the task indeed have a language impairment. It is expected that more children will fail on the TSI, but turn out not to have a severe language disorder. The *Screening study part A* aims at answering questions a and b. A follow-up study, '*Screening study part B: Follow-up study*', of the children in the

[15] 'Language Screening Tool', a Dutch measure used for detecting possible language problems in kindergarten children currently used in The Netherlands for screening for language impairments in 5-year-olds (see also section 2.2).

[16] An existing non-word repetition task was adapted for the present studies, see section 3.1.2.

'impaired' group was conducted by judging information on the language status of the children two years after testing. This was done in order to determine how many children in this group in fact had a language impairment, and whether the NRT or the language screening tool TSI had more accurately predicted language impairment in this group. The expectation is that the TSI does not identify language impairment as well as the NRT, i.e., the TSI is expected to identify too many false positives in the multilingual group caused by the linguistic bias. The follow-up study aims at answering question c, as addition to questions a and b. See chapter 3 for a discussion of the *Screening study.*

The *Assessment study* aims at providing an more solid answer to questions a and b. In the *Screening study,* the children with language impairment were not diagnosed by application of a 'gold standard' of diagnosis.[17] Sackett *et al.* (1985) mention that in determining the clinical usefulness of a diagnostic test, a comparison has to be made with a 'gold standard' of diagnosis. In the *Assessment Study,* a comparison is made between multilingual children without language problems and multilingual children diagnosed with SLI conforming to the 'gold standard' by multidisciplinary tests with an interpreter. In this study, the NRT is used as an assessment tool for SLI. It is expected that the NRT is proficient in ruling in and ruling out SLI in these groups. This study will be addressed in chapter 4.

[17] For SLI, the 'gold standard' of diagnosis includes: language assessment with accepted norm-referenced language tests used to identify deficits in different language domains (syntax, lexicon, pragmatics, and phonology) administered together with an interpreter, accompanied by other multidisciplinary tests to exclude the presence of other factors such as cognitive disabilities or hearing impairment.

3. SCREENING STUDY

3.1 Screening study part A

3.1.1 Participants

Sixty-five five-year-old children participated in this study. The children included were tested with a language screening tool (‘*Taalscreeningsinstrument*’, TSI, Gerritsen, 1994), which is used by the GGD (‘*Gemeentelijke Gezondheidsdienst*’, ‘*Public Health Service*’) in Maastricht. This task is administered by a speech therapist visiting the school. The score on the screening tool determined inclusion in one of the groups in this study. If children obtained a score lower than 34 on TSI, they were labelled as having language problems. ‘Language problems’ should not be confused with ‘SLI’. None of the children in the ‘impaired’ group were diagnosed with SLI, as this requires more than language screening alone. (Some) children in the impaired group may have secondary language problems, caused by, for instance, second language learning, or general cognitive deficits. The children with language problems will be referred to as having ‘language impairment’ - not otherwise specified.

Children were also distributed on the basis of multilinguality or monolinguality. The children were divided into three groups:

1. Multilingual children without language problems (MuNI) (i.e., score ≥34 on the TSI) (*n* = 25). This group included 15 male and 10 female subjects.
2. Multilingual children with language problems (MuLI) (i.e., score TSI <34 on the TSI) (*n* = 25). This group included 15 male and 10 female subjects.
3. Monolingual (Dutch-speaking) children without language problems (MoNI) (control group, score ≥34 on the TSI) (*n* = 15).[18] This group included 4 male and 11 female subjects.

Because it was decided to include multilingual five-year-olds speaking *any* minority home language, the multilingual groups were highly heterogeneous with respect to the home languages of the children.[19] Mean age and TSI scores per group are provided in table 4.

[18] The group size of the MoNI group was smaller because less within-group variation was expected since these children all spoke only one and the same language.

[19] The home languages spoken by the multilingual participants included Afrikaans, Angolan dialect, Berber, (Serbo-) Croatian, English, French, Farsi, German, Kurdisch language, Polish, Portuguese, Mandarin, Somali, Spanish, Tamil language, and Turkish

Table 4*. Mean age and TSI scores with SD's per group. MuNI =multilingual non-impaired, MuLI = multilingual language impaired, MoNI = monolingual non-impaired.*

	MuLI **(*n* = 25)**		**MuNI** **(*n* = 25)**		**MoNI** **(*n* = 15)**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age (months)	65.80	2.97	65.40	3.14	67.47	2.67
TSI score	25.44	6.23	38.96	4.28	43.93	4.59

The children in the Monolingual NI group are a few months older than the children in the to other groups. A One-Way ANOVA showed that there was no significant difference between the three groups with respect to age (F(2,64)=2.39, p=0.100).

3.1.2 Stimuli

The Non-word Repetition Task used for this study resembled the task created by De Bree *et al.* (2007; De Bree, 2007), adapted by Gerrits (2004). The NRT was designed such that the wordlikeness and the articulatory complexity of the non-words were low. The non-words reported in De Bree (2007) met the following criteria, taken from Dollaghan and Campbell (1998):

1. The non-words and their individual syllables do not correspond to lexical items in Dutch. High wordlike test items may provide children with a larger lexicon to take advantage of their word knowledge.
2. The non-words contain only phonemes that are acquired early in language development to minimise the articulatory complexity of the task (i.e., no /r/ or consonant clusters).
3. Only phonemes that are acoustically salient were included, i.e., only tense vowels were included, no lax vowels.
4. No phoneme occurs more than once in each non-word.
5. The stimuli are pre-recorded in order to standardise presentation with respect to consistent rate, accuracy, and intonation).

As was already pointed out in section 2.4.4, there is a relationship between previous language experience (i.e., the amount of phonotactic knowledge), and performance on a language specific non-word repetition task (e.g., Girbau & Schwartz, 2008; Kohnert *et al.,* 2006). In this study, the non-word repetition task will be used with children from diverse linguistic backgrounds. This implies that the task has to be as *language neutral* as possible with respect to the languages spoken by the children to be tested. The composition of the task was based on the finding that Turkish, Standard Arabic, and Berber are the three minority languages that are most common in children enrolled for language assessment in The Netherlands (see also sections 2.3 and 3.2.1). The adaptations that were made by Gerrits (2004) were aimed specifically at these three languages.

The following adaptations were made:

1. Using only phonemes that appear in Dutch, as well as Turkish, Berber, and Standard Arabic.
2. Adding the consonant [k] to allow for more variation in the non-words.[20]
3. Adding more variation in the final consonant (De Bree's task had many nasals in non-word final position).

For more information, see Gerrits (2004).

The new test contained two-to-five-syllable non-words, six of each non-word length. Three practice items were included. An overview of the non-words is provided in table 5.

Table 5*. Non-words used for this study. ‘ marks stressed syllable.*

Practice items	**Test items**			
	2 syllables	**3 syllables**	**4 syllables**	**5 syllables**
’pef	fi’pas	do’finep	be’pytasuf	bawo’vylizen
’tus	ji’fot	ju’sewaun	ka’jemytɛip	fonɛi’wusetaf
mi’fap	ke’fys	ki’lumef	le’tomifus	kepi’sufytaus
	pe’lum	nu’pifat	py’sudajin	leda’nokimuf
	wo’kan	wa’fɛilin	si’batonek	peta’sɛikonif
	su’fep	sy’timof	ti’kepofam	tiwa’kesufaup

The 27 non-words were produced by an adult female native speaker of Dutch, recorded on a CD. Each word was preceded by a beep. The non-words were presented pseudo-randomly, i.e., the length of two consecutive non-words was never the same. To minimise order effects, two different item lists were made with different non-word order.

The TSI (‘Taalscreeningsinstrument’, *Language Screening Tool*) is a norm-referenced language screening tool used by speech therapists for detecting language problems in 5-year-olds as part of usual care. It includes components like general questions, language comprehension, morphology, word and sentence repetition, and reasoning. Failing on the TSI (i.e., obtaining a score of 34 or lower), leads to enrolment into further language assessment.

3.1.3 Procedure and data analysis

The TSI and NRT were administered by a speech therapist working for the *GGD Zuid-Limburg* in Maastricht in the period between December 2004 and April 2005. A CD with the stimuli was played through a Discman with loudspeakers. Each trial started with showing the child a picture of a fantasy animal (i.e., Pokémon animations). To make the task appealing to the young children, a

[20] Recall the criterion for the non-words to comprise only phonemes that are acquired early in language development. The consonant [k] is typically not part of the earlier acquired phonemes. However, more variation in the non-words was considered important, therefore a concession was made. Additionally, [k] is not described as one of the ‘Late Eight’ acquired phonemes for English (Dollaghan & Campbell, 1998), as well as for Dutch (Beers, 1995).

Bingo game with these pictures was created. The pictures of the fantasy animals are drawn from a pile, and the child matches the picture with the same picture on 3x3 matrices printed on a card. Instructions were as following: *"Through the speaker you will hear a beep. After that beep, you will hear a funny word. That's the name of a funny animal. Look, this is the first animal. You may repeat its name. You will hear the word only once."* After each item, the CD is paused, allowing the child to imitate, and to put the picture of the animal on the matching picture on the Bingo card. The three practice items were included to allow the child to get acquainted with the procedure of the task. All items were presented only once, except when the non-word could not be perceived due to background noise. The children's utterances were audio-recorded with a Sony mini-disk recorder with a microphone for later phoneme-to-phoneme transcription. Test lists 1 and 2 were counterbalanced between subjects. The non-word repetition task took each participant about 10 minutes.

3.1.3.1 Scoring

All utterances of the non-words were transcribed, and scored by a trained transcriber. On the basis of the transcriptions, two types of scores were calculated per non-word, based on De Bree (2007).[21]

Raw score. An item was awarded 1 point when the non-word was repeated completely correct. If one or more phonemes of the non-word were repeated incorrectly (i.e., substituted or omitted), 0 points were awarded. Mutual substitutions of the vowels [o], [y], and [u] were not scored as incorrect because (small) vowel inaccuracies (assumed to be present in second language learners, specifically in Turkish participants) are not desired to influence the score. But, for example, [a] – [i] substitutions were scored as incorrect. Phoneme additions were counted as correct, because they are not regarded as a reflection of loss of information about the target phonemes. The maximum total raw score was 24.

Phoneme score. The percentage of correctly repeated phonemes (PPC) per repeated non-word was calculated by dividing the number of correctly repeated phonemes by the total number of phonemes in the target non-word, and then multiplied by 100. For each child, the mean percentage of correctly repeated phonemes for all non-words was calculated. Voicing and devoicing errors (e.g., substitution of [b] by [p], or [t] by [d]), and some vowel substitutions (interchange of [o], [y], and [u]) were treated as correct. The percentages of substitutions and omissions of phonemes per non-word were calculated. Phoneme additions were counted as correct.

[21] De Bree (2007) also calculated a syllable score and a phoneme percentage addition score which were not considered to be relevant for this study, considering the clinical application of the task.

***Table 6**. Examples of scoring (raw score and phoneme score) for different imitations of the non-word /fiˈpas/ (“fipaas”) (5 phonemes). Phonemes correct, substitutions, and omissions provided in percentages.*

fi’pas	Raw score	Phonemes		
		Correct	Substituted	Omitted
fi’pas	1	100	0	0
mi’pas	0	80	20	0
i’pas	0	80	0	20
si’ka	0	40	40	20
ba’nas	0	40	60	0
’nas	0	40	20	40
sfi’pas	1	100	0	0
fi’past	1	100	0	0

3.1.3.2 Reliability

The transcription of the responses was checked for reliability. Recordings from 10% percent randomly selected subjects (in the total sample of the screening study and the assessment study together) were transcribed by a second transcriber (the author) to determine inter-rater reliability. The level of agreement for judgement of correctness was 93%, which was considered sufficient. Also, the reliability of the scoring was checked. Another sample of ten percent was scored by a second analyst (the author). Inter-rater reliability for the raw score was 99%, and 93% for the phoneme scores. These inter-rater reliability levels were considered sufficient.

3.1.4 Results

Eight multilinguals with language problems (32%), three multilinguals without language problems (12%), and seven monolinguals (47%) repeated fewer than 24 targets. This was possibly due to shyness or fatigue. The mean number of repeated target non-words for these children was 19.4 (SD=3.97), ranging from 12 to 23. Further inspection of the missing cases showed that the four- and five-syllable non-words were avoided most. Of the missing cases, 28% were four-syllable non-words, and 65% were five-syllable non-words (only 1% two-syllable non-words, and 6% three-syllable non-words). Children with missing cases were not included in the analyses of the raw scores, but were included in the analysis of the PPC scores.

3.1.4.1 Analyses

In table 7 below, mean age, the TSI score, the raw score on the NRT, and the mean percentage phonemes correct (PPC) on the NRT are presented per group.

Table 7*. Mean age (in months), and scores with SD's for TSI, raw score and PPC per group. For MuLI, n = 25, for MuNI, n = 25, for MoNI, n = 15, except when otherwise specified (for raw score).*

	MuLI **TSI < 34**		**MuNI** **TSI ≥ 34**		**MoNI** **TSI ≥ 34**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age	65.80	2.97	65.40	3.14	67.50	2.67
TSI score	25.44	6.23	38.96	4.28	43.93	4.59
Raw score	10.29 (*n* = 17)	4.50	15.41 (*n* = 22)	3.53	18.50 (*n* = 8)	3.89
PPC[22]	84.14	9.34	92.45	4.58	94.10	4.38

The raw scores of the NRT and the PPC scores are plotted in figures 3 and 4.

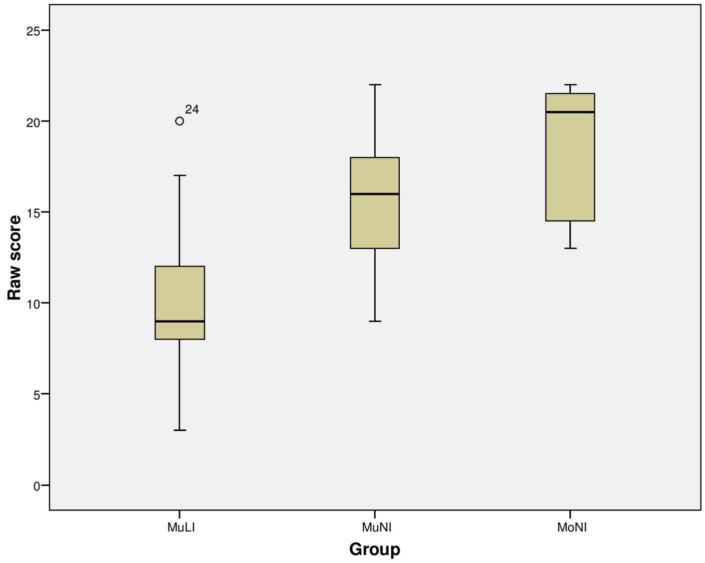

Figure 3*. Raw scores per group, only for children with completed NRTs. The boxes contain 50% of all cases. The line in the box indicates median value. The whiskers indicate the maximum and minimum values. The open circle refers to a mild outlier.*

[22] In the PPC analysis, missing items (i.e., non-words that were not repeated) were excluded from the calculations. It was considered 'unfair' to award a score of zero to missed items, since the child would most probably have obtained a higher score when it would in fact have repeated the non-word. When awarding a 0% score to missed repetitions, the mean (SD) PPC scores are: MuLI: 78.26 (15.60), MuNI: 88.29 (12.60), MoNI: 91.23 (5.90). This results in a larger group difference.

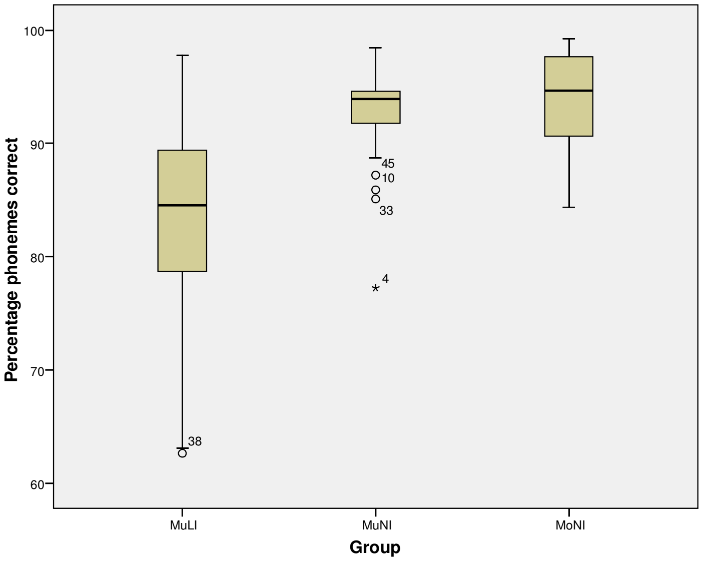

***Figure 4**. Percentage phonemes correct per group. The boxes contain 50% of all cases. The line in the box indicates median value. The whiskers indicate the maximum and minimum values. Open circles refer to mild outliers; the asterisk refers to an extreme outlier.*

A One-way ANOVA showed that the three groups differ significantly on the TSI score (F(2,64)=72.16 , p<0.001). Games-Howell post-hocs confirmed that all three groups differed significantly from each other (p<0.05) (Monolinguals NI > Multilinguals NI > Multilinguals LI). The three groups differ significantly on the raw score of the NRT (F(2,46)=13.97, p<0.001).[23] Tukey HSD post-hocs showed that the difference between Multilinguals NI and Monolinguals NI was not significant (p=0.154), while Multilinguals LI performed significantly worse than Multilinguals NI (p<0.001) and Monolinguals NI (p<0.001). A One-way ANOVA also showed a significant group effect on the PPC score (F(2,64)=13.53, p<0.001). Games-Howell post-hocs showed that there was no significant difference between Multilinguals NI and Monolinguals NI (p=0.502), while the Multilinguals LI performed significantly worse than the two other groups (p<0.001).

It can be concluded from the boxplots that the range of distribution for the PPC score is much larger in the Multilinguals LI group than for the two other groups. This difference is possibly caused by the large heterogeneity of the group of children with (S)LI. [24]

Correlations between the different types of scores were computed. All groups taken together, there was a strong correlation between the TSI score and the PPC score (*r*(64)=0.86, p<0.001). For the children with completed NRTs, there was a moderate correlation between the raw score and the TSI score (*r*(46)=0.55, p<0.001), and a strong correlation between the raw score and the PPC score

[23] Games-Howell post-hocs were only applied when variances were not homogeneous (as determined by *Levene's Test of Homogeneity of Variances*). In other cases, Tukey-HSD post-hocs were used.

[24] The interpretation that the PPC score is possibly influenced by the home language (type) of the children was not confirmed in an analysis with language type as dependent variable. This implies that there was no effect of the home language on the performance on the task. However, the limited number of children per language (type) group could have influenced the results of this analysis.

(*r*(46)=0.90, p<0.001). Correlations per group were also computed, and revealed that there was no significant correlation between the TSI score and the PPC score for either of the three groups (MuLI (*r*(24)=0.13, p=0.55; MuNi (*r*(24)=0.23, p=0.28); MoNi (*r*(14)= -0.13, p=0.64)). The same was true for the correlation between raw score and TSI (MuLi (*r*(16)=0.19, p=0.460); MuNI (*r*(21)=0.26, p=0.25); MoNI (*r*(7)=-0.43, p=0.29)). In all three groups, a strong correlation between the raw score and the PPC score was found (MuLi (*r*(16)=0.90, p<0.001); MuNI (*r*(21)=0.79, p<0.001); MoNI (*r*(7)=0.96, p<0.001)). The correlation between the raw score and the PPC score suggests that both scores are equally good at predicting language impairment. The presence of a (moderate) correlation between the TSI score and the PPC score and raw score (found when the groups were taken together) was unexpected, because it was assumed that the TSI contains a linguistic bias which is not present in the NRT (which implies that there would be no correlation, or only a weak correlation, between the two tasks).

Non-word length

Mean PPC values were calculated per non-word length group. In all three groups, the PPC score decreases when target length increases. This effect is largest for the group Multilinguals LI (see figure 5). For Monolinguals NI, the decrease is smallest.

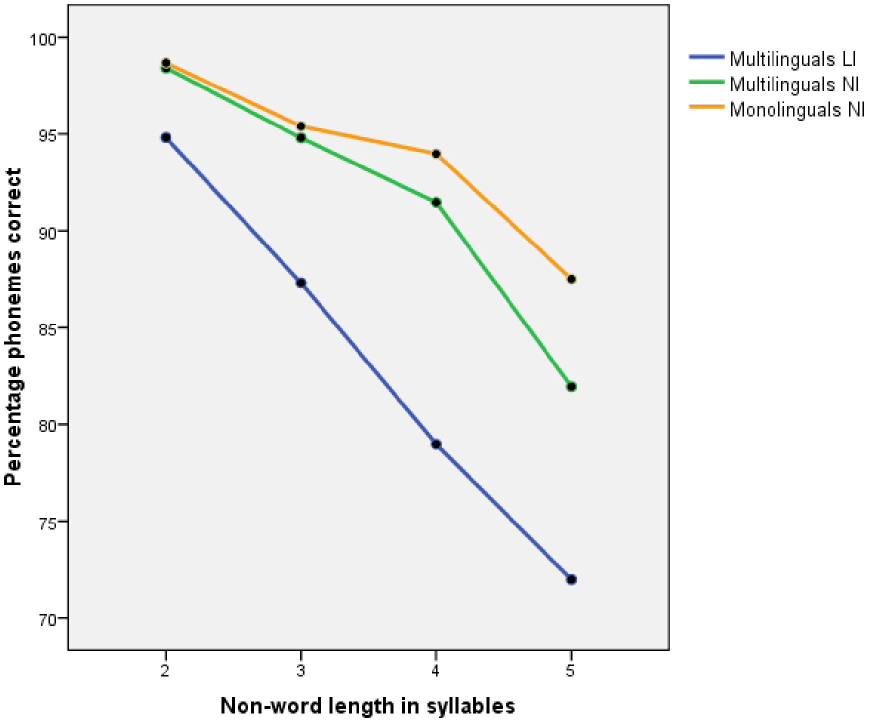

Figure 5. *Mean PPC per non-word length per group.*

The target length effect can further be examined by means of a Two-way Repeated Measures ANOVA. A Two-way Repeated Measures ANOVA, with the PPC score as dependent variable, Group (Multilinguals LI, Multilinguals NI, Monolinguals NI) as between-subjects factor, and Non-word Length (2, 3, 4, 5 syllables) as the repeated within-subjects factor showed a significant main effect for Group (F(2,58)=10.28, p<0.001) and Non-word Length (F(2.16,125.01)=56.62, p<0.001). There was a significant interaction between Group and Non-word Length (F(4.31,125.01)=3.20, p=0.013).[25] This implies that the steepness of the decrease in PPC with non-word length is associated with the group. Games-Howell post-hocs further showed that the Multilinguals LI performed significantly worse than the other two groups (p<0.05), but that the Multilinguals NI and Monolinguals NI did not differ (p=0.42).

The interaction between Group and Non-word Length seems due to the relatively steep decrease of the PPC with increasing non-word length especially for the Multilinguals LI compared with a more gradual decrease of the Monolingual NI group. To determine the location of the interaction, a One-way ANOVA with post-hoc analysis was performed. A series of One-way ANOVA's with Group as between-subjects factor and Non-word length (2, 3, 4, and 5 syllables) as dependent variables, showed that the difference between the groups is significant for all non-word lengths (two-syllable non-words F(2,64)=9.52, p<0.001; three-syllable non-words F(2,64)=8.08, p=0.001; four-syllable non-words F(2,64)=12.26, p<0.001; five-syllable non-words F(2,60)=5.20, p=0.008). Post-hoc analysis shows the location of the interaction between group and non-word length.

Table 8*. Post-hoc multiple comparisons. For the 2- to 4-syllable non-words, Games-Howell post-hoc analysis was used, for 5-syllable non-words, Tukey HSD post-hocs were used.*

Non-word length (in syllables)	Multilingual LI - Multilingual NI	Multilingual LI - Monolingual NI	Multilingual NI - Monolingual NI
2	p=0.002 (*)	p=0.006 (*)	p=0.993
3	p=0.009 (*)	p=0.005 (*)	p=0.819
4	p=0.003 (*)	p<0.001 (*)	p=0.267
5 [26]	p=0.075	p=0.009 (*)	p=0.516

(*) = the mean difference is significant.

Summarising, the Multilinguals LI performed significantly worse than the Multilinguals NI on the two-, three-, and four-syllable non-words, but not on the five-syllable non-words. The PPC scores of Multilinguals LI were significantly lower than Monolinguals NI on all non-word lengths. There was no significant difference between the PPC scores of Multilinguals NI and Monolinguals NI on any of the non-word lengths.

[25] Huynh-Feldt corrections for asphericity lead to corrected degrees of freedom.

[26] A test of Homogeneity of Variances was not significant for 5-syllable non-words (p=0.463), therefore Tukey HSD post-hocs were used.

Phoneme errors

Different error scores were calculated for the two possible error types: phoneme percentage substitutions (PPS), and phoneme percentage omissions (PPO).[27] The results are presented in table 9.

***Table 9**. Phoneme percentage substitution and phoneme percentage omission per group.*

	MuLI		**MuNI**		**MoNI**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
PPS	21.60	5.97	18.13	5.19	16.93	4.73
PPO	18.12	9.20	12.45	5.91	18.70	20.01

In general, substitution is a more frequently applied strategy than omission, except for the Monolinguals. The analysis of substitution errors showed different levels for the three groups. A One-way ANOVA showed that there was a significant difference between the three groups with respect to the *percentage phoneme substitutions* (F(2,64)=4.28, p=0.018). Tukey HSD post-hoc analysis showed that the Multilinguals LI made significantly more substitution errors than the Monolinguals NI (p=0.028). There was no significant difference between Multilinguals NI and Multilinguals LI (p=0.067), and Monolinguals NI (p=0.779). The presence of phoneme substitution implies that length and syllable structure of the non-words were often preserved in the child's response. There was no significant difference between the three groups with respect to *percentage phoneme omission* (F(2,60)=2.06, p=0.137).

3.1.4.2 Ruling in/ruling out language impairment with the NRT

As mentioned before, both the raw score and the PPC score can be used for indicating performance on a non-word repetition task. However, the PPC score is used more often than the raw score, and it is considered to be more accurate in investigating phonological memory. More importantly, the PPC scores of *all* children can be included in this analysis, also children who did not repeat all items (children with missing cases were not included in the analyses of the raw scores). Taking these arguments into account, in this study only the accuracy of the PPC scores (i.e., total PPC scores, and PPC per non-word length) will further be examined. In the following paragraphs, the utility of the non-word repetition task as a diagnostic measure will be examined by means of *likelihood calculations*.[28] A likelihood ratio expresses the odds that a certain level of percentage phonemes repeated correctly (PPC) would be expected in a child with (as opposed to one without) language impairment. Likelihood ratios were calculated following the method described in section 2.4.5.2.

[27] Analysis of the phoneme percentage substitution and omission per non-word length did not yield informative data, and is therefore not reported here.

[28] Another method that was used to determine the accuracy of the task, ROC analysis, will be discussed in the Appendix.

Likelihood ratios for the PPC scores

Likelihood ratios were determined in order to find the most sensitive cut-points for ruling in and ruling out the presence of language impairment. Optimal likelihood ratios for the PPC scores will be provided in table 10.

***Table 10**. Likelihood ratios for the percentage phonemes repeated correctly (PPC), for language impaired and non-impaired multilingual, with optimal cut point for this method: PPC ≤80.*

PPC	**MuLI (*n* = 25)**		**MuNI (*n* = 25)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤ 80	7	0.28	1	0.04	Positive test result: 7.00
81-90	12	0.48	4	0.16	3.00
≥ 91	6	0.24	20	0.80	Negative test result: 0.30

The likelihood ratio for ruling a child into the LI group based on a PPC at or below 80% is (0.28/0.04=) 7.00. This means that a PPC of 80% or lower is approximately 7 times more likely to come from a child with LI than from a child without LI in the present sample. The likelihood ratio for a negative test result (to rule out the presence of a language impairment, defined as a PPC of 91% or greater is (0.24/0.8=) 0.30, which means that a PPC this high was approximately one-third as likely to come from a child with LI as from a child without LI. Although the likelihood ratios seems quite good, the task still identifies false negatives (i.e., impaired children scoring *above* the cut-off score for a positive test result).

Likelihood ratios for the PPC per non-word length

There was an effect of non-word length on the PPC performance (see figure 5). Therefore, it may be interesting to examine the likelihood ratios of the PPC scores per non-word length. The levels of performance that discriminated best between the Multilingual LI group and the Multilingual NI group were determined, and are provided per non-word length in tables 11 to 14 below.

***Table 11**. Likelihood ratios for percentage phonemes correct for two-syllable non-words (PPC2).*

PPC2	**MuLI (*n* = 25)**		**MuNI (*n* = 25)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤93	11	0.44	2	0.08	5.50
≥94	14	0.56	23	0.92	0.61

***Table 12**. Likelihood ratios for percentage phonemes correct for the three-syllable non-words (PPC3).*

PPC3	**MuLI (*n* = 25)**		**MuNI (*n* = 25)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤86	9	0.36	1	0.04	9.00
87-91	7	0.28	6	0.24	1.17
≥92	9	0.36	18	0.72	0.50

Table 13*. Likelihood ratios for percentage phonemes correct for the four-syllable non-words (PPC4).*

PPC4	**MuLI (*n* = 25)**		**MuNI (*n* =25)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤78	11	0.44	1	0.04	11.00
79-90	9	0.36	9	0.36	1.00
≥91	5	0.20	15	0.60	0.33

Table 14*. Likelihood ratios for percentage phonemes correct for the five-syllable non-words (PPC5).*

PPC5	**MuLI (*n* = 25)**		**MuNI (*n* = 25)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤70	10	0.43	2	0.09	5.00
71-80	5	0.22	4	0.17	1.25
≥81	8	0.35	17	0.74	0.47

The levels of performance on each PPC yielded informative likelihood ratios. The obtained likelihood ratios for the performance on the three- and four-syllable non-words and scores are higher than those obtained with the total PPC score. Particularly, the performance on the four-syllable non-words distinguished between LI and NI with a relatively high degree of confidence.

3.1.5 Discussion

In the Screening study part A, it was shown that the non-word repetition task is able to discriminate between multilingual children with and without language problems. The multilingual children with language problems had more difficulty with repeating non-words than (multilingual and monolingual) children without language impairment. There were no significant differences between the non-word performance of children without language problems, whether they were multilingual or not. These results were as expected, and they suggest that the non-word repetition task is indeed a culturally fair method to identify language problems.

Further analysis revealed that the longer a non-word, the more errors are made. This was true for all three groups. A more severe non-word length effect was found in the group of children with language problems, compared to the two non-impaired groups. These results are similar to the findings of other studies (e.g., Dollaghan & Campbell, 1998; Girbau & Schwartz, 2008). The most frequent errors for all three groups were consonant substitutions, which indicates that the syllable stucture and the item length were often maintained.

The percentage phonemes repeated correctly (PPC) and the raw score were equally good at identifying children with language impairment, however, the PPC score is considered to be the more precise measure. For this reason, the PPC was further examined with respect to its diagnostic efficiency. In order to determine the accuracy of the task, likelihood ratios were calculated for the total PPC scores and the PPC scores per non-word length. The likelihood ratios obtained in this study suggest that the PPC score has the potential to be a valuable screening tool to detect language impairment in multilingual children. Specifically, a total PPC score of 80 or less indicates that a child is likely to have a language problems. If a child obtains a PPC score of 91 or greater, it is likely to have no language impairment. It should be noted that the task does identify a considerable level of false negatives (children with a language impairment scoring *above* the cut-off score for a positive test result). This indicates that the results should be interpreted and treated carefully. In order to use a task as a screening tool, it may be suggested to use cut-off points that lie above the 'optimal' cut-off point, in order to obtain more (false) positives.

Of the PPC scores per non-word length (ranging from two to five syllable-items), the *PPC4 score* (i.e., the percentage phonemes repeated correctly in four-syllable non-words) was shown to be the most accurate measure in discriminating between children with and without language impairment, at the threshold PPC4 scores of ≤78% for a positive test result, and ≥91 for a negative test result. The likelihood ratios for ruling in and ruling out language impairment exceed those of the total PPC. This implies that the performance on non-words of four syllables is the most exact measure for discriminating between children with and without language impairment.

Compared to previous studies (Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000; Kohnert *et al.,* 2006; Girbau & Schwartz, 2008), the PPC cut-off points for a positive test result are relatively high in the present study. This implies that the children in the present study performed better than the groups in studies of others, but differences between the tasks, and the composition of the experimental groups make it impossible to make a clear-cut comparison between the different studies. The levels of likelihood ratios are in line with previous studies by Ellis-Weismer *et al.* (2000), Kohnert *et al. (*2006), and Girbau and Schwartz (2008), although the findings of this study do not replicate the clear-cut results of Dollaghan and Campbell's (1998) study. However, this was the case for others, too (e.g., Ellis-Weismer *et al.*, 2000; Girbau and Schwartz, 2008; Kohnert *et al.* 2006). This difference may be due to the selection of the participants: the children included in the present study were not diagnosed with SLI by a 'gold standard'.

The results indicate that NRT *alone* is not enough in ruling in and ruling out language impairment. This is in line with findings by, for example, Conti-Ramsden (2003) and Ellis-Weismer *et al.* (2000). However, still, it seems to be a promising measure in screening for (S)LI. For children scoring 'positive' in the task, additional testing will have to determine whether the child indeed has (S)LI. Four-syllable words were most accurate in identifying children with SLI, which suggests that non-word lists of four syllables might be the most effective for diagnostic purposes.

By using likelihood ratios, we were able to compare the results of the present studies with those obtained by others, which was considered to be very useful. There are however some drawbacks to this method. The rather small sample size in the present study has some serious implications for the calculations of likelihood ratios. Choi (1998) stated that likelihood ratios cannot be derived easily from experimental data because of limited sample size. Therefore, the results should be interpreted carefully. Another, possibly more accurate method, of determining the diagnostic efficiency of a task is *ROC analysis* will also be performed. For readability's sake, a discussion of this method and its results will not be provided in this thesis' text, but is reported in the Appendix.

In this first screening study, the multilingual children were divided into a language impaired and a non-impaired group based on their TSI scores. It is expected that the TSI contains a cultural and linguistic bias since this screening instrument has been developed for Dutch monolingual children. It might be the case that due to the linguistic bias in the TSI, the children identified as having a language impairment were in fact not impaired. Therefore, a follow-up of the language status of the children in the impaired group was conducted, and the NRT and TSI scores of these children were analysed again. This analysis will be discussed in the following sections.

3.2 Screening study part B: Follow-up study

3.2.1 Introduction

A follow-up of the language status of the children in the Multilingual LI group was conducted, in order to determine the accuracy of the TSI as well as the NRT. The aim of the follow-up study was to find out how many of the children identified by the TSI as 'having language problems', were indeed probably language impaired. The NRT and TSI scores of these children were examined in order to find out which of these two measures is most suitable to detect language impairment.

The follow-up included a subdivision of the children in the Multilingual LI group. Two years after testing, the GGD speech therapist had judged the child's language again, often based on second hand information, from the teacher or a speech therapist from a private practice. The information on language resulted in more children moving from the impaired into the non-impaired group. The analyses were repeated with two new groups of the same children, now based on information two years after the screening.[29]

[29] It should be kept in mind that the children were *not* tested again. The results of two years before were analysed again in the newly formed groups.

3.2.2 Procedure

To determine whether the TSI had identified language problems correctly, the status of the children in group Multilinguals LI (TSI score <34) was checked two years after testing. The information on language development and intervention provided by the GGD client data base was judged by two speech therapists. With the information gathered two years after testing, the children in the MuLI group were subdivided in two groups: children with language impairment, and children without language impairment.

If a child had (had) speech-language therapy, he or she was considered as having language impairment (group Mu2LI).[30] No improvement of performance in Dutch after long-term language therapy, is often a indication for language weakness. Children without reported language problems were labelled as non-impaired (Mu2NI). Two years after screening, there were 11 children with language impairment, and 13 children without language problems.[31] This implies that the TSI falsely predicted language impairment in almost half of the group. Of one child in the group, there were no data on intervention status. The results of that child were not included in this analysis.
It has to be noted that there are two important drawbacks that bias this second judgement of the child's language abilities. First, the GGD client data base might not be entirely updated. Second, more importantly, none of the multilingual children's language was assessed according to the 'gold standard'. In most cases, a standardised Dutch language test was used, measuring the second language of the child, and not the first language. This means that it remained unclear whether children with a low score had a developmental language disorder or a second language learning problem, which is a serious problem of the design of this study.

3.2.3 Analyses

The TSI scores, and the NRT scores (the raw scores, and the PPC scores) *of two years before* were analysed again in the newly formed groups. The means will be provided in table 15.

[30] The '2' in the names of the groups indicates that the group composition was defined by the outcomes of the 'second judgement'.

[31] It should be kept in mind that the diagnosis 'SLI' was not provided for the children with LI, since this requires multidisciplinary assessment.

***Table 15**. Mean age (in months) and scores with SD's for the two subgroups in the original MuNI group. For Mu2LI, n = 11, and Mu2NI, n = 13, except when otherwise specified (for raw score).*

	Mu2LI		**Mu2NI**	
	Mean	**SD**	**Mean**	**SD**
Age	66.73	3.13	64.73	2.70
TSI	21.45	5.97	29.15	4.08
Raw score	10.29 (*n* = 7)	5.96	10.30 (*n* = 10)	3.50
PPC[32]	84.11	10.53	84.14	9.05

A One-way ANOVA showed that the 'new' Multilingual LI group performed significantly worse than the 'new' Multilinguals NI on the TSI (F(1,23)=13.97, p=0.001). The groups hardly show any difference on the raw score and the PPC score. A One-Way ANOVA confirmed this observation: the difference on the raw score between the two groups was not significant (F(1,16)=0.000, p=0.995), or on the PPC score (F(1,23)=0.000, p=0.995).

To complete the picture, the children considered as having no language impairment can be added to the original group of multilinguals without language impairment. Now the three definitive groups can be compared. See table 16.

***Table 16**. Mean age (in months), and scores with SD's for TSI, raw score and PPC per group. For MuLI, n = 11, for MuNI & Mu2NI, n = 38, and for MoNI, n = 15, except when otherwise specified (for raw score).*

	Mu2LI		**MuNI & Mu2NI**		**MoNI**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age	66.73	3.13	65.00	2.97	67.50	2.67
TSI	21.45	5.97	35.61	6.28	43.93	4.59
Raw score	10.29 (*n* = 7)	5.96	13.81 (*n* = 32)	4.22	18.50 (*n* = 8)	3.89
PPC	84.11	10.53	89.60	7.49	94.10	4.38

The raw scores and PPC scores on the NRT are plotted in figures 6 and 7.

[32] In the PPC analysis, missing items were excluded from the calculations. When awarding a 0% score to missed repetitions, the mean (SD) PPC scores are: Mu2LI: 75.80 (17.26), MuNI & Mu2NI: 85.50 (13.87), MoNI: 91.23 (5.90). The difference between the two multilingual groups grew larger when including the missed items in this analysis.

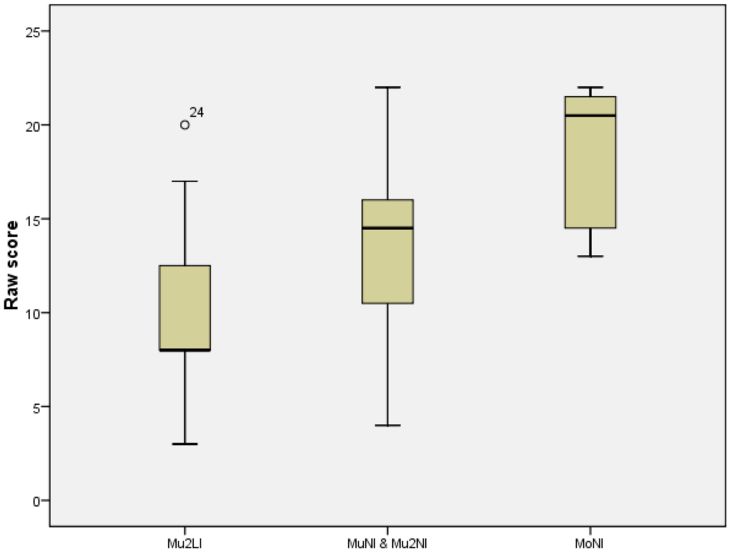

Figure 6. Raw scores per group, only for children with completed NRTs. The boxes contain 50% of all cases. The line in the box indicates median value. The whiskers indicate the maximum and minimum values. The open circle refers to a mild outlier.

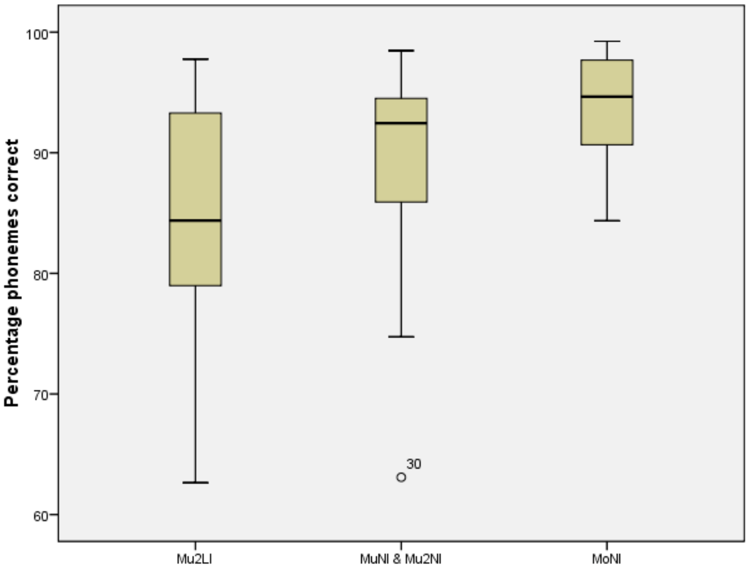

Figure 7. Percentage phonemes correct per group. The boxes contain 50% of all cases. The line in the box indicates median value. The whiskers indicate the maximum and minimum values. The open circle refers to a mild outlier.

A One-way ANOVA showed that there was a significant group effect on the TSI score (F(2,63)=46.60, p<0.001). A Tukey HSD post-hoc analysis confirmed that all three groups differ significantly from each other (p<0.001) (Monolinguals NI > Multilinguals NI > Multilinguals LI). There also was a significant group effect on the raw score of the NRT (F(2,46)=6.56, p=0.003). Games-Howell post-hocs showed that the Monolinguals performed significantly better than the Multilinguals LI (p=0.027), and the Multilinguals NI (p=0.029). There was no significant difference between the Multilinguals LI and Multilinguals NI (p=0.351). There was a significant group effect on the PPC score (F(2,63)=5.59, p=0.006). A Games-Howell post-hoc analysis showed that the Monolinguals performed significantly better than the Multilinguals LI (p=0.029), and the Multilinguals NI (p=0.026), while there was no significant difference between the Multilinguals LI and Multilinguals NI (p=0.274).

The overall correlations naturally remained identical to those obtained in the first part of the screening study, as the same set of participants was used. Recall that, overall, there was a strong correlation between the TSI score and the PPC score (*r*(64)=0.86, p<0.001). For the children with completed NRTs, there was a moderate correlation between the raw score and the TSI score (*r*(46)=0.55, p<0.001), and a strong correlation between the raw score and the PPC score (*r*(46)=0.90, p<0.001).

For the Multilinguals LI group, there was no significant correlation between the TSI score and the PPC score (*r*(10)=0.33, p=0.327), or between the TSI score and the raw score on the NRT (*r*(6)=0.35, p=0.445). There was, however, a strong correlation between the raw score and the PPC score (*r*(6)=0.91, p=0.004). For the Multilinguals NI group, there was a weak correlation between the TSI score and the PPC score (*r*(37)=0.44, p=0.005), and a moderate correlation between the TSI score and the raw score (*r*(31)=0.52, p=0.003), and a strong correlation between the raw score and the PPC score (*r*(31)=0.87, p<0.001).

The correlations between the scores of the Monolinguals NI group were already discussed in section 3.1.4.1: there was no significant correlation between the TSI score and the PPC score (*r*(14)= -0.13, p=0,64), or between the raw score and TSI (*r*(7)=-0.43, p=0.292. There was a significant correlation between the raw score and the PPC score (*r*(7)=0.96, p<0.001).

Non-word length

An analysis of the performance per non-word length showed that the PPC score decreased with increasing non-word length (see figure 8). A two-way repeated measures ANOVA, with the PPC as dependent variable, Group (Multilinguals LI, Multilinguals NI, Monolinguals NI) as between-subjects factor, and Non-word Length (2, 3, 4, 5 syllables) as the repeated within-subjects factor showed a significant main effect for Group (F(1,57)=4.54, p=0.015) and Non-word Length (F(2.181,124,32)=42.99, p<0.001).[33] There was no significant interaction between Group and Non-word Length (F(4.36,124.32)=1.91, p=0.107). Games-Howell post-hocs further showed that the Monolinguals performed significantly better than Multilinguals LI (p=0.022). The Monolinguals also performed better than the Multilinguals NI, but the difference was not significant (p=0.059). There was no significant difference between the Multilinguals with and without impairment (p=0.415).

[33] Huynh-Feldt corrections for asphericity lead to corrected degrees of freedom.

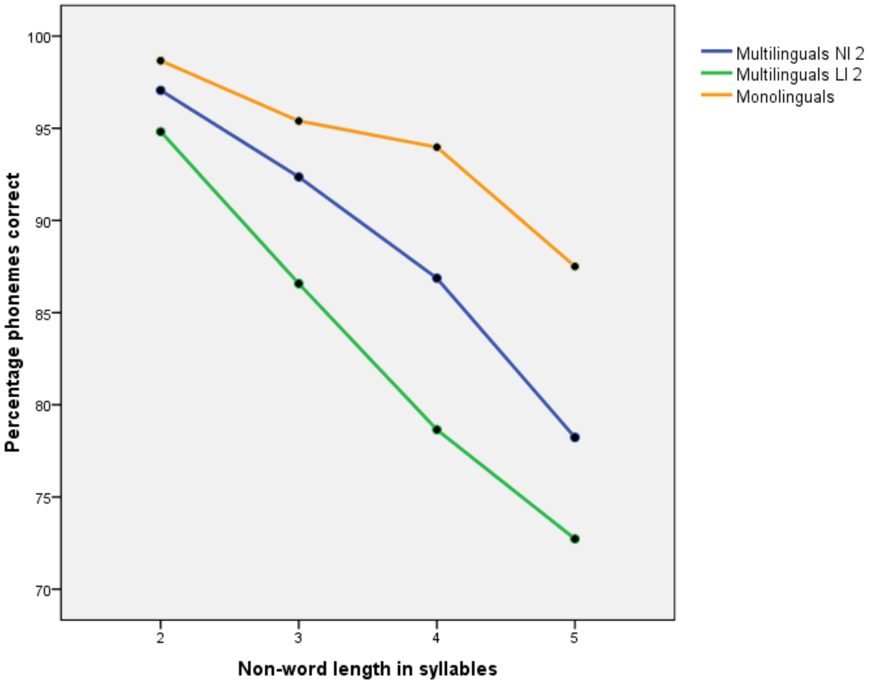

Figure 8. *Mean PPC per non-word length per group.*

Figure 8 shows that the difference between the groups of multilinguals with language impairment and the multilinguals without language impairment grew smaller, compared to the results from *Screening study part A*.

Phoneme errors

An analysis of the percentage phoneme substitution and omission per group lead to the following results:

Table 17. *Phoneme percentage substitution and phoneme percentage omission per group.*

	Mu2LI		**MuNI & Mu2NI**		**MoNI**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
PPS	22.04	7.62	19.25	5.22	16.93	4.73
PPO	19.29	9.40	13.92	7.47	18.70	20.01

The groups show an error pattern that is quite similar to that in *Screening study A*. A One-way ANOVA now showed that the group difference is not significant for either the percentage phoneme substitution (F(2,63)=2.66, p=0.078), or the percentage phoneme omission (F(2,59)=1.40, p=0.255).

3.2.4 Ruling in/ruling out language impairment with the NRT and the TSI

Likelihood ratios were calculated for the NRT (percentage phonemes repeated correctly) and the TSI, with the 'new' groups based on the follow-up information. A comparison between the accuracy of the NRT and the TSI will be made.

***Table 18**. Likelihood ratios for the percentage phonemes repeated correctly.*

PPC	Mu2LI (*n* = 11)		MuNI & Mu2NI (*n* = 38)		
	Number	Proportion	Number	Proportion	Likelihood ratio
≤81	5	0.45	5	0.13	3.45
82-93	3	0.27	16	0.42	0.65
≥94	3	0.27	17	0.45	0.61

The likelihood ratio for a positive test result (defined as ≤81) is 3.45, which means that a PPC score of ≤81 is about three times more likely to come from a child with language impairment than a child without language impairment. For a negative test result, this is 0.61. Recall the likelihood ratio levels obtained with the 'original' group division: respectively 7.00 and 0.30 for a positive and negative test result – these levels were better than the values obtained with the 'new' group division. However, the optimal cut-point remained almost identical, namely ≤81.

Given that the PPC4 score was previously identified as the most accurate measure, it was also analysed for this follow-up study. The optimal cut-off PPC score for the four-syllable non-words is now slightly lower, namely ≤74 (it was ≤78).

***Table 19**. Likelihood ratios for percentage phonemes correct for the four-syllable non-words (PPC4).*

PPC4	Mu2LI (*n* = 11)		MuNI & Mu2NI (*n* = 38)		
	Number	Proportion	Number	Proportion	Likelihood ratio
≤74	5	0.45	3	0.08	5.76
≥75	6	0.55	35	0.92	0.59

The new group division allows making a (limited) comparison between the NRT and the TSI. Because the non-impaired groups (the children scoring ≥34) were not checked for false negatives, there is no information available on true and false negatives identified by the TSI. However, an analysis of true and false positives can be made. In table 20, the likelihood ratio for a positive test result is provided.

***Table 20**. Likelihood ratio for a negative test result on the language screening tool TSI.*

	Mu2LI (*n* = 11)		Mu2NI (*n* = 38)		
	Number	Proportion	Number	Proportion	Likelihood ratio
< 34	11	1	13	0.34	2.90

The likelihood ratio for ruling in impairment with the TSI is 2.90, which is slightly lower than for the NRT (3.45), but the difference is very small. This might indicate that the NRT and the TSI are equally accurate measures to identify (S)LI. It should, however, be kept in mind that the size of the samples is small. Also, the score on the TSI was used as an inclusion criterion for one of either multilingual groups. Therefore, results may be influenced (i.e., a larger group difference on TSI). This indicates that results should be interpreted very carefully.

3.2.5 Discussion

In this first part of the screening study, children were divided into an 'impaired' and a 'non-impaired' group based on their TSI scores. In this follow-up study, it was attempted to make a better motivated distinction between children without language problems and children having language impairment (possibly, a specific language impairment). The language status of the children who failed on the TSI two years before was traced. This follow-up study showed which children had indeed language impairment (i.e., perhaps an indication for SLI). It turned out that (at least) 13 children in this group had no language problems, i.e., TSI had identified them incorrectly. However, the NRT did not show perfect performance either, given the presence of false negatives.

Within the larger sample (with the two other groups added to the newly divided group), the group differences were similar to those of the *Screening study part A*. However, the likelihood ratios for a positive test result grew smaller in comparison to those found in the 'original' sample, which implies that the predictive ability of the non-word repetition task turned out to be less accurate than initially hypothesised. The likelihood ratios for a positive test result did not show large differences for the total PPC and the TSI, which suggests that both are equally good at ruling in language impairment. However, the absence of information on (false) negatives identified by the TSI is a limitation, as it would have been interesting to find out whether children scoring high on the TSI, had in fact a language impairment (i.e., presence of false negatives). The results may have been biased since the TSI score was used as criterion for inclusion. Interestingly, the PPC4 score turned out to be a more accurate measure than the TSI. Children obtaining a PPC4 score of 74 or lower, are likely to have a language impairment. This implies that, when one would administer a task with only four-syllable non-words, the non-word repetition task is the best measure to identify (S)LI in multilinguals, compared to the (total) PPC and the TSI.

It was attempted to accurately determine whether a child had language impairment or not. However, none of the children in the 'final' impaired group were diagnosed with SLI by a 'gold standard' of diagnosis (the accepted reference used to identify SLI). Language tests by themselves are not sufficient for assessment of SLI. As was already discussed in section 2.2, identifying SLI involves multidisciplinary tests, which are – in the case of multilingual children – administered with an interpreter. To address this issue, the *Assessment study* compares children without language problems and children diagnosed with SLI conforming to the ‘gold standard’ by multidisciplinary tests with an interpreter. This allows us to formulate a more solid answer to the question whether the NRT is an accurate measure in discriminating multilingual language impaired children from typically developing multilinguals. The *Assessment study* addresses the question whether the NRT is a useful tool in assessment of SLI in multilinguals. This study will be discussed in the following chapter.

4. ASSESSMENT STUDY

4.1 Participants

In this study, children diagnosed with SLI conforming to the 'gold standard' are compared with children without language problems. Twenty-nine 4- to 7-year-old multilingual children participated in this study, divided into two groups:

1. Multilingual children diagnosed with SLI (MuSLI) (*n* = 14).[34] Mean age was 77.7 months (ranging from 66 to 93 months). The group included 11 male and 3 female subjects.
2. Multilingual children without SLI (MuNI) (*n* = 15). Mean age was 75.8 months (ranging from 56 to 92 months). The group included 7 male and 8 female subjects.

The design was to only include multilingual children speaking the most frequently spoken minority home languages in The Netherlands. Two matters were taken into consideration. First, it was already shown that – ranked by frequency in decreasing order – Turkish, Hind(ustan)i, Berber, Standard Arabic, and English are the largest minority home languages spoken by children in primary schools in The Hague (see section 2.3, table 1, for an overview of these languages). Secondly, an inventory of the home languages of 'allochtone' children enrolled for diagnostic research in all Dutch Speech and Hearing Centres was made to determine a ranking of the home languages of these children.[35] It was shown that the three most common home languages, ranked by frequency in decreasing order, were Turkish, Standard Arabic, and Berber. Therefore, it was decided to only include Turkish, Standard Arabic, and Berber speaking multilinguals into the experimental groups. The initial idea was to include 10 children from each home language, but it was difficult to find enough participants within the time available.

The children diagnosed with SLI were recruited via the Speech and Hearing Centre in The Hague, and the Cor Emousschool in The Hague (a special school for children with speech, language, and hearing problems). The selection criteria for the children in this group were:

a. Home language Turkish, Berber, or Standard Arabic.
b. No hearing difficulties.
c. Normal cognitive development (non-verbal IQ ≥85).
d. No known signs of neurological damage.

[34] A fifteenth child in this group was not able to complete the task, possibly due to shyness or fatigue. Data from this child were not included for this study.

[35] Inventory performed by Mirjam Blumenthal in the period between July 2002 and June 2003.

e. No developmental verbal dyspraxia or severe phonological disorder.[36]
f. Diagnosed as having SLI conforming to the 'gold standard' by multidisciplinairy tests (e.g., language tests, hearing tests, IQ measures) with interpreter.[37]

A questionnaire on language background and development was filled in by the teacher for each child in the Multilingual SLI group. The group of children diagnosed with SLI included nine Turkish speaking children, four Standard Arabic speaking children, and one Berber speaking child.

The group of multilingual children without SLI was selected to match the ages and languages of the children in the SLI group. This implies that the composition of the non-impaired group was only known after the selection of the children from the impaired group.
The selection criteria for the children in the non-impaired group were:
a. Home language Turkish, Berber or Standard Arabic.
b. No known hearing difficulties.
c. Normal cognitive development (non-verbal IQ ≥ 85).
d. No known signs of neurological damage.
e. No (history of) language problems.

The group of children without language problems included nine Turkish speaking children, five Standard Arabic speaking children, and one Berber speaking child.

4.2 Stimuli

The NRT was identical to the one used for the Screening study. See section 3.1.2 for more information.

4.3 Procedure and data analysis

Only the non-word repetition task was administered as part of the experiment. Scoring and data analyses were similar as for Study 1. See section 3.1.3 for more information. The check for reliability of the transcription and scoring was included in the check of the total sample of the screening study and assessment study.

[36] If a child was known to have the inability to produce speech sounds included in the non-word repetition task, the child was excluded from the experiment. This was determined through a questionnaire filled in by the speech therapist.

In the Multilingual SLI group, the NRT was administered by a speech therapist the Speech and Hearing Centre in The Hague in the period between August 2004 and May 2006. In the Multilingual NI group, the NRT was administered by a speech pathologist in Maastricht in the period between May 2007 and October 2007.

4.4 Results

Eight multilinguals with SLI (57%), and seven multilinguals without language problems (47%) repeated fewer than 24 targets, due to shyness or fatigue. The mean number of repeated target non-words for these children was 22.6 (SD=0.64), ranging from 21 to 23. Like in study 1, the four- and five-syllable words were avoided most. Of the missing cases, 28 percent were four syllable words, and 62 percent were five syllable words (only 1% two-syllable words, and no three-syllable words). Children with missing cases were not included in the analyses of the raw scores, but were included in the analysis of the PPC scores.

4.4.1 Analyses

In table 21 below, mean age, raw score on NRT, and mean percentage phonemes correct are presented per group.

***Table 21**. Mean age and scores with SD's for TSI, raw score and PPC per group. For MuSLI, n = 14, for MuNI, n = 15, except when otherwise specified (for raw score).*

	MuSLI		**MuNI**	
	Mean	**SD**	**Mean**	**SD**
Age	77.71	9.55	75.80	8.80
Raw score	10.17 (*n* = 6)	4.83	12.88 (*n* = 8)	5.41
PPC[38]	82.49	7.80	90.26	6.41

A One-Way ANOVA showed that there was no significant difference between the three groups with respect to age (F(1,28)=0.316, p=0.579).

[37] It turned out that in the examination of seven children an interpreter was involved. For four of these children, this happened recently before the experiment, for three of them approximately 2 years before the experiment, during the enrolment procedure for a special school for children with severe speech and language problems.

[38] In the PPC analysis, missing items were excluded from the calculations. When awarding a 0% score to non-words that were not repeated, the mean (SD) PPC scores are: MuSLI: 80.14 (9.05), and MuNI: 87.19 (6.50). The group difference did not grow larger.

The raw scores and the PPC scores are plotted in figures 9 and 10 below.

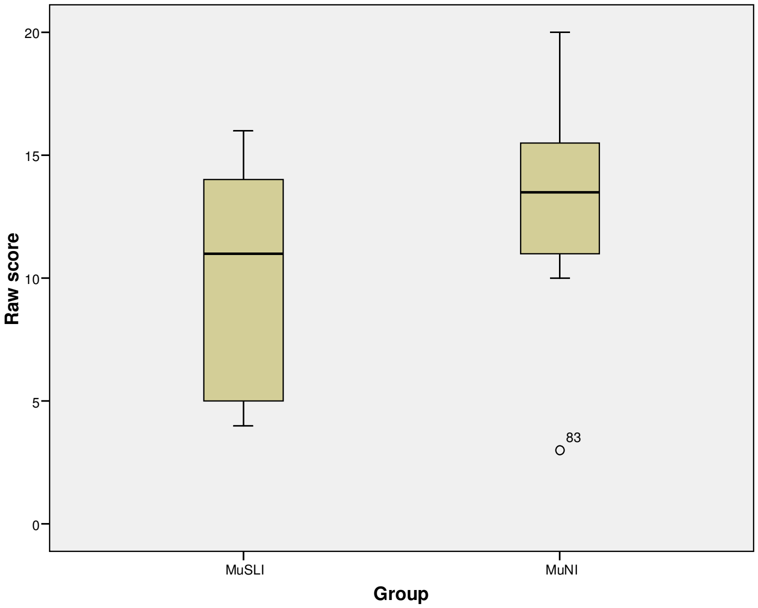

Figure 9. Raw scores per group, only for children with completed NRTs. The boxes contain 50% of all cases The line in the box indicates median value. The whiskers indicate the maximum and minimum values. The open circle refers to a mild outlier.

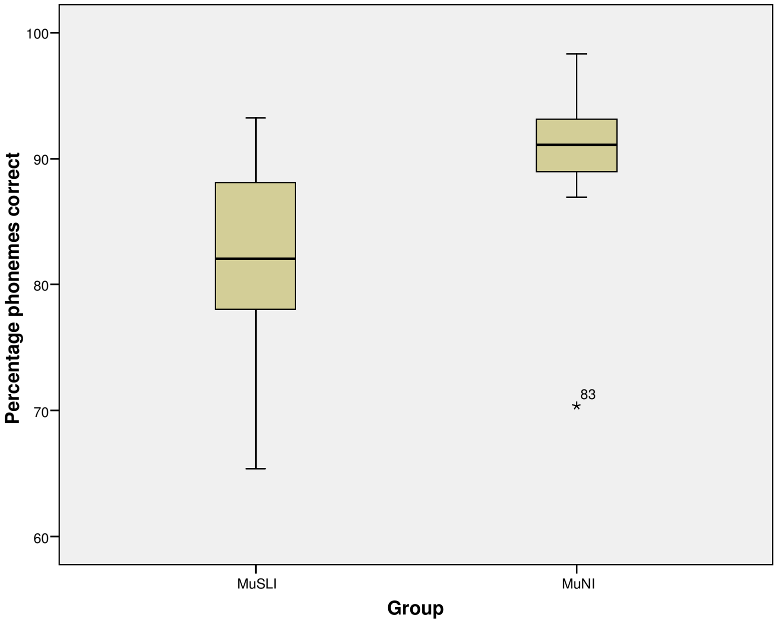

Figure 10. Percentage phonemes correct per group. The boxes contain 50% of all cases. The line in the box indicates median value. The whiskers indicate the maximum and minimum values. The asterisk refers to an extreme outlier.

The results of the two groups were compared by means of One-way ANOVA. There was no significant difference on the raw score (F(1,13)=1.03 , p=0.331).[39] The difference between Multilinguals with SLI and Multilinguals NI was significant on the PPC score (F(1,28)= 8.63 , p=0.007).

[39] This may be due to the very small sample sizes (*n* = 6 and *n* = 8).

For the children with completed NRTs, there was a strong correlation between the raw score and the PPC score (*r*(13)=0.92, p<0.001). This suggests that both scores are equally good in identifying SLI. However, this was not confirmed by the statistical analysis (possibly due to the small sample size).

Non-word length

A further inspection of the PPC showed that it decreases when non-word length increases. This effect is largest for the group of multilinguals with SLI, as is shown in figure 11 below.

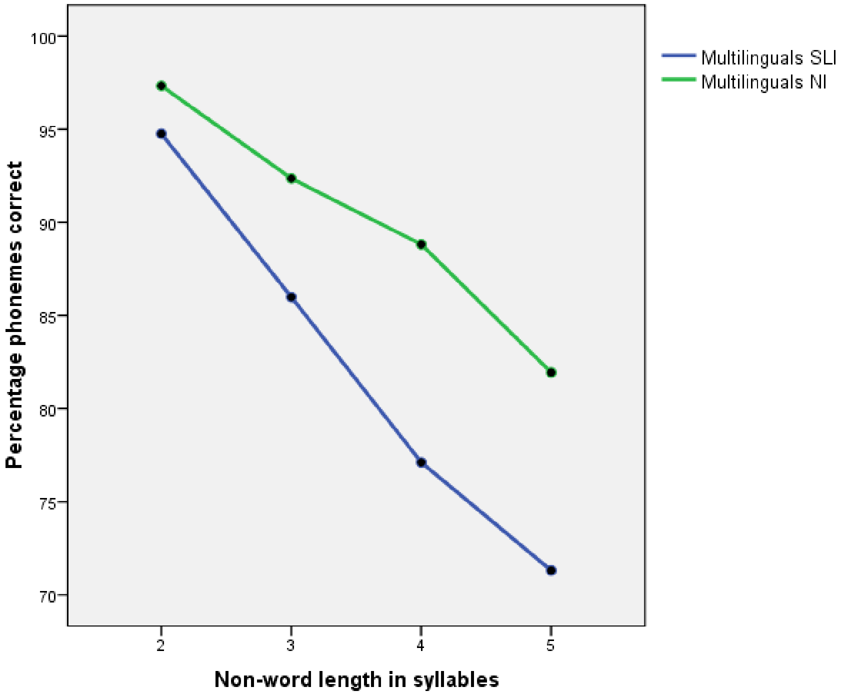

Figure 11. *Mean PPC per word length per group.*

A two-way repeated measures ANOVA with the PPC as dependent variable, Group (MuSLI, MuNI) as between-subjects factor, and Non-word Length (2, 3, 4, 5 syllables) as the repeated within-subjects factor showed a significant main effect for Group (F(1,27)=8.80 , p=0.006) and Non-word Length (F(3,81)=46.51 , p<0.001), and an interaction between Group and Non-word Length (F(3,81)=2.94 , p=0.038). The interaction between Group and Non-word Length seems to be caused by the relatively steep decrease of the PPC with increasing non-word length for the Multilinguals SLI, while the Multilingual NI group shows a more gradual decrease (see figure 11). A One-way ANOVA with post-hoc analysis was performed to determine the location of the interaction. This ANOVA, with Group as between-subjects factor and Non-word length (2, 3, 4, and 5 syllables) as dependent variables, showed the difference between the groups is only significant for non-words with four-syllables (F(1,28)=9.77, p=0.004), and five-syllables (F(1,28)=6.28, p=0.019). There was no significant difference between the groups on two-syllable non-words (F(1,28)=3.25, p=0.083), and three-syllable non-words (F(1,28)=3.50, p=0.072).

The children's ages ranged from four to seven years of age. It was previously shown that non-word repetition performance increases with age (see, for example, De Bree *et al.,* 2007). Possibly, the factor age has had some influence on the PPC scores in the present study. An analysis showed that there is a weak correlation between age and the PPC score (*r*(28)=0.44, p=0.016). The older the child, the higher the score on the NRT. Correlations were also found between age and PPC2 (moderate correlation (*r*(28)=0.59, p=0.001)), and PPC4 (weak correlation (*r*(28)=0.41, p=0.026)). The scores on the two other non-word lengths did not correlate (PPC3 (*r*(28)=0.343, p=0.068), PPC5 (*r*(28)=0.27, p=0.152).

Phoneme errors

Different error scores were calculated for the two possible error types: phoneme percentage substitutions, and phoneme percentage omissions. The results of the total phoneme error scores are presented in table 22.

***Table 22**. Phoneme percentage substitution and phoneme percentage omission per group.*

	MuSLI		MuNI	
	Mean	**SD**	**Mean**	**SD**
PPS	23.83	5.06	17.47	5.32
PPO	17.27	4.98	17.53	6.80

The error analysis showed that non-impaired multilinguals tend to make fewer substitution errors than Multilinguals with SLI. Substitution is a strategy applied more by the multilingual SLI group than in the non-impaired group. There is no substantial difference between the groups with respect to omission errors. A One-way ANOVA showed that the group difference with respect to the total phoneme percentage substitution was significant (F(1,28)=10.86, p<0.001). This was not true for the total phoneme percentage omission (F(1,28)=0.01, p=0.909).

4.4.2 Ruling in/ruling out language impairment with the NRT

Likelihood ratios for the PPC scores

The likelihood ratio for a positive test result (to rule in the presence of SLI), defined as a PPC of 84 or lower, is 8.57. The likelihood for a negative test result was 0.21 (see table 23).

***Table 23**. Likelihood ratios for the percentage phonemes repeated correctly (PPC), for language impaired and non-impaired multilingual, with optimal cut point for this method: PPC ≤84.*

PPC	**MuSLI (*n* = 14)**		**MuNI (*n* = 15)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤84	8	0.57	1	0.07	8.57
85-92	5	0.36	9	0.2	0.60
≥93	1	0.07	5	0.33	0.21

Likelihood ratios for the PPC per non-word length

PPC4 turned out to be the most informative. The likelihood ratio for a positive test result (to rule in the presence of SLI), defined as a PPC4 of 79 or lower, is 8.57. The likelihood for a negative test result was 0.40. PPC2 was less informative according to the outcomes of the likelihood calculations. See tables 24 - 27 below.

Table 24. *Likelihood ratios for percentage phonemes correct for the two-syllable non-words (PPC2).*

PPC2	**MuSLI (*n* = 14)**		**MuNI (*n* =15)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤90	5	0.36	1	0.07	5.36
91-97	6	0.43	7	0.47	0.92
≥98	3	0.21	7	0.47	0.46

Table 25. *Likelihood ratios for percentage phonemes correct for the three-syllable non-words (PPC3).*

PPC3	**MuSLI (*n* = 14)**		**MuNI (*n* =15)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤90	8	0.57	3	0.20	2.86
≥91	6	0.43	12	0.80	0.54

Table 26. *Likelihood ratios for percentage phonemes correct for the four-syllable non-words (PPC4).*

PPC4	**MuSLI (*n* = 14)**		**MuNI (*n* =15)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤79	8	0.57	1	0.07	8.57
80-90	3	0.21	6	0.40	0.54
≥91	3	0.21	8	0.53	0.40

Table 27. *Likelihood ratios for percentage phonemes correct for the five-syllable non-words (PPC5).*

PPC5	**MuSLI (*n* = 14)**		**MuNI (*n* =15)**		
	Number	**Proportion**	**Number**	**Proportion**	**Likelihood ratio**
≤71	7	0.50	2	0.13	3.75
72-86	5	0.36	8	0.53	0.67
≥87	2	0.14	5	0.33	0.43

4.5 Discussion

In the *Assessment study*, multilingual children with SLI were compared to multilingual children without language problems. The children with SLI performed significantly worse on repetition of non-words in comparison to the children without language problems. The longer the non-word, the more errors are made. The non-word length effect on the performance was largest for the SLI group. Furthermore, it was shown that a correlation exists between age and the PPC score: performance on the non-word repetition task increases with age (which was already shown by, for example, De Bree *et al.*, 2007). This implies that normative data should be collected per age (category).

There were some differences between the results of the assessment study and the screening study. For example, the difference between the mean PPC scores of the language impaired and non-impaired group was larger in the assessment study. This was confirmed by the larger likelihood ratios. This may be caused by the fact that the language impaired children in this group were diagnosed with specific language impairment conforming to a 'gold standard' of diagnosis. The language impaired children in the screening study may have had less severe language problems. Another difference between the studies was that the children in the assessment study obtained a lower mean PPC score than the children in the screening study – interestingly, this was also true for the non-impaired children. It is not clear what caused this difference, but it may be due to language (type) differences, or age.

Likelihood ratio analysis showed that a total PPC score of 84 or less indicates that a child is likely to have a language impairment. Conversely, if a child obtains a PPC score of 93 or greater, it is likely to have no language impairment. Similar to the screening study, the assessment study showed that PPC4 is the most accurate measure in discriminating between children with and without language impairment, at a PPC4 score of ≤79 for a positive test result, and ≥91 for a negative test result) From the levels of likelihood ratios can be concluded that the non-word repetition is able to identify children with SLI rather well, but it should never be the only language assessment tool in children with SLI. General issues concerning the current study are discussed in the following chapter.

5. General Discussion and Conclusion

The present study examined the performance of multilingual and monolingual children with and without (specific) language impairments on a non-word repetition task to determine the clinical usability of this task as a screening tool to detect language disorders in multilingual children from diverse linguistic backgrounds in The Netherlands. It aimed to answer the following research questions:

a. Does the non-word repetition task (NRT) discriminate between monolingual and multilingual children without SLI?
b. Is the NRT a reliable clinical tool for discriminating between multilingual children with and without (S)LI?
c. Is the NRT's reliability greater than the 'Taalscreeningsinstrument' (TSI)?

Question a, whether the non-word repetition task discriminates between monolinguals and multilinguals without SLI can be answered with *no,* which was hypothesised. The two screening studies showed that there are no significant differences between the two non-impaired groups. This result is in line with the results that were previously found by others (e.g., Campbell *et al.,* 1997; Conti-Ramsden, 2003; Dollaghan & Campbell, 1998; Ellis-Weismer *et al.*, 2000; Girbau & Schwartz, 2008). It indicates that the non-word repetition task is a culturally fair method of screening children for SLI since the task minimally depends on prior knowledge or experience.

Question b addressed the ability of the non-word repetition task to discriminate between multilinguals with and without (S)LI. The research question can be answered with *yes*. This result was expected, as it was also shown by Campbell *et al.* (1997), Conti-Ramsden (2003), Dollaghan and Campbell (1998), Ellis-Weismer *et al.* (2000), and Girbau and Schwartz (2008). The non-word repetition has the potential to be a valuable screening tool to detect language impairment in multilingual children. The percentage phonemes repeated correctly (PPC) is a suitable measure for discriminating between children with and without language impairment.[40] The *screening study* suggested that a total PPC score of 81 or less indicates that a child is likely to have a language impairment. If a child obtains a PPC score of 91 or greater, it is likely to have no language impairment. Of the four different non-word lengths, the PPC4 score (i.e., the percentage phonemes repeated correctly in four-syllable non-words) was shown to be the most accurate measure in discriminating between children with and without language impairment in both the screening studies and the assessment study. A PPC4 score of 78 or lower indicates that a child is probably

[40] Both scoring at phoneme level and the raw score provide accurate discrimination between children with and without SLI (e.g., Gray, 2003). Although phoneme level scoring can be viewed as a 'richer' way of analysing, from a clinical point of view it may not be preferred because it is relatively time consuming. Phoneme scoring requires recording of the responses and *offline* scoring at a later time, while raw scoring can take place immediately during administration of the task (i.e., *online).*

language impairment. The likelihood ratios for ruling in and ruling out language impairment exceeded those of the total PPC. This finding implies that the performance on four-syllable non-words is the most accurate measure for discriminating between children with and without language impairment, which has some clinical implications. It may be suggested to design a non-word repetition task of mainly four-syllable non-words for diagnostic utility.

A screening tool is desired to overidentify rather than to underidentify (Washington & Craig, 2004). However, the task in the present study identified a substantial number of *false negatives*. Therefore, it may be advised to use a more 'conservative' cut-off score, e.g., a score that lies *above* the 'optimal' cut-off point in order to obtain more false positive scoring children. By doing so, the clinician can be sure to identify all children having language impairment. Those children will have to enrol in further language assessment to rule language impairment in or out.

It should be noted that performance on non-word repetition was previously found to be closely related to the level of language-specific phonotactic knowledge, as was shown by studies discussed in section 2.4.4. In the present study, the stimuli were not absolutely language neutral since complete language neutrality is impossible. However, the fact that the current task did not discriminate between multilinguals and monolinguals without language impairment suggests that this had no large effects on the performance of the children speaking more than one language.

Most researchers investigating the accuracy of the non-word repetition tasks used likelihood ratios to express the success of the task in discriminating between language impaired and non-impaired children. For that reason, the optimal cut-off points provided in this study were also derived by likelihood ratio calculations in order to compare the present study with other studies on non-word repetition. The small sample size in the present study has, however, some serious implications for the outcome of likelihood ratios. According to Choi (1998), likelihood ratios cannot be derived easily from experimental data because of limited sample size. Therefore, the results should be interpreted carefully. Additionally, another, potentially better method was applied: *ROC analysis* (see the Appendix). The differences between the outcomes of these of different methods of analysing the accuracy of the task indicate that the results also depend on the was of analysing. Using a larger sample may solve this problem.

Question c, 'Is the NRT's reliability greater than the 'Taalscreeningsinstrument' (TSI)?' cannot be answered with a definitive answer. It should be noted that the design of the screening study was not optimal to be able answer this question. The children in the 'impaired' group were not diagnosed with SLI by a 'gold standard', which has serious implications for the interpretation of the results. Additionally, the absence of information on false negatives on the TSI is a limitation of the design of the study. Previous studies (e.g., Dollaghan and Campbell, 1998) suggested that knowledge-based language measures (such as the TSI) would be inferior to a processing-based measure such as non-word repetition. The current study did not confirm this observation completely. The screening study part B (the *Follow-up study*) showed that in the multilingual sample, the total percentage phonemes repeated correctly did not exceed accuracy of the TSI. However, the PPC4 score turned out to be the better measure compared to either the total PPC score or the TSI. These results suggest that the TSI does have potential, however, it may be useful to include a (mainly four-syllable) non-word repetition task to the current screening tool in order to obtain a more solid outcome.

Taking the findings of the present experiments into consideration, the central question *'Is non-word repetition suitable as a screening tool to detect language disorders in multilingual children from diverse linguistic backgrounds in The Netherlands?'* may be answered with a careful *yes*. The non-word repetition task shows to be a promising clinical tool, specifically for the multilingual population. It may be one step forward to the problem of identifying SLI in multilingual children, but its results should be interpreted carefully. The findings of this study (together with those of previous studies) indicate that further language assessment is always necessary, because the performance on the non-word repetition task *alone* is never sufficient to rule in and rule out language impairment. It should be noted, that neither is the TSI able to do so, as is every screening tool. It is important to always complete the assessment with other measures, observations, and parental interviewing.[41] This makes it possible for the clinician to derive specific goals for speech and language therapy.

Summarising, the application of the NRT fits nicely in the need for accurate measures to detect language disorders in the growing multilingual population. The traditional language screening tool TSI is less accurate due to an assumed linguistic bias. The NRT is an accurate tool in discriminating between multilinguals with and without SLI. Additionally, the NRT is easy to administer and score, which are major advantages compared to other measures or procedures such as dynamic assessment, ethnographic interviewing, and language assessment with an interpreter. The task may be used to supplement the traditional language screening tool.

Suggestions for future research

The present study showed that the NRT may have a substantial clinical utility. Its diagnostic accuracy should be further investigated, in order to be able to possibly implement it as a screening tool to detect language disorders in a multilingual population.

The cut-off scores for ruling in or ruling out language impairment were not identical for the different samples in the current experiments. In future research, normative data of a large sample should be collected. In future research, the NRT performance of a larger group of multilingual children with and without SLI may be further investigated. It may be suggested to use mainly four-

[41] When assessing a multilingual child, the clinician should always be assisted by an interpreter.

syllable non-words, as those were shown to be the most accurate measure. A much larger experimental group is necessary to obtain to more reliable results, which is important in order to determine the accuracy of the task (such as likelihood ratio calculations or ROC analysis), and to obtain valid normative data. The Committee On Test Affairs Netherlands (COTAN) states that a sample size of at least 200 individuals is needed in order to make stable normative data.[42] Since non-word repetition performance appears to increase with age, normative data for different ages or age categories should be collected.

In the *screening study*, there was no information on whether the children in the 'impaired' group indeed had SLI. In future research, it is crucial to have the language impaired children diagnosed with SLI conforming to a 'gold standard' to be able to make a clear-cut distinction between the performance of language impaired and non-impaired children. The results of such a study may be compared to the performance of monolingual children with and without SLI, in order to identify the possible specific difficulties of multilinguals on the NRT.

Previous studies have mentioned the influence of the first language on the performance on the NRT. The home language (type) groups were too small to indicate whether this was also true for the present study. Therefore, a further investigation of NRT performance of children per home language group is necessary. A large influence of home language (type), will have major implications for the construction of the task. The language neutrality of the task used in the present study is debatable. However, as was already mentioned, it is impossible to create a completely language neutral task. The only possible solution to this issue, in my opinion, is to design a *language specific* task (i.e., a task following the phonotactic rules of the home language of the child), rather than a *language neutral* task. For future research, a language specific task may be designed specifically for each language group, e.g., the largest language groups Turkish, Berber, and Arabic.

[42] http://www.cotan.nl (last checked 22 October, 2008)

6. LITERATURE

Aarts, R., Extra, G. & Yağmur, K. (2004). *Multilingualism in The Hague*. In: Extra, G. & Yağmur, K. (eds.), *Urban Multilingualism in Europe. Immigrant Minority Languages at Home and School*. Clevedon: Multilingual Matters.

Appel, R. & Vermeer, A. (2000). *Tweede-taalverwerving en simultane taalverwerving*. In: Gillis, S. & Schaerlaekens, A. (eds.), *Kindertaalverwerving. Een handboek voor het Nederlands*. Groningen: Martinus Nijhoff Uitgevers.

Archibald, L.M.D. & Gathercole, S.E. (2006). Short-term and working memory in specific language impairment. *International Journal of Language & Communication Disorders,* 41(6), 675-693.

Archibald, L.M.D. (2008). The promise of non-word repetition as a clinical tool. *Canadian Journal of Speech Language Pathology and Audiology,* 32, 21-28.

Beers, M. (1995). *The phonology of normally developing and language impaired children*. Doctoral Dissertation. Amsterdam: University of Amsterdam.

Bishop, D.V.M. (2006). What causes Specific Language Impairment in Children? *Current Directions in Psychological Science,* 15, 217-221.

Bishop, D.V.M., North, T., Donlan, C. (1996). Nonword repetitions as a behavioural marker for inherited language impairment: evidence from a twin study. *Journal of Child Psychology and Psychiatry*, 36, 1-13.

Campbell, T., Dollaghan, C., Needleman, H. & Janosky, J. (1997). Reducing Bias in Language Assessment: Processing-Dependent Measures. *Journal of Speech, Language, and Hearing Research*, 40, 519-525.

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.

Choi, B.C.K. (1998). Slopes of a Receiver Operating Characteristic Curve and Likelihood Ratios for a Diagnostic Test. *American Journal of Epidemiology,* 148, 1127-1132.

Conti-Ramsden, G. (2003). Processing and Linguistic Markers in Young Children With Specific Language Impairment (SLI). *Journal of Speech, Language and Hearing Research*, 46, 1029-1037.

De Bree, E. (2007). *Dyslexia and phonology: A study of the phonological abilities of Dutch children at-risk of dyslexia*. Doctoral Dissertation. Utrecht: LOT. University of Utrecht.

De Bree, E., Rispens, J. & Gerrits, E. (2007). Non-word repetition in Dutch children with (a risk of) dyslexia and SLI. *Clinical Linguistics & Phonetics*, 21, 935-944.

De Jong, J. (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure*. Doctoral Dissertation. Groningen: Rijksuniversiteit Groningen.

De Jong, P.F. & Van der Leij, A. (1999). Specific Contributions of Phonological Abilities to Early Reading Acquisition: Results from a Dutch Latent Variable Longitudinal Study. Journal of Educational Psychology, 91, 450-476.

Dollaghan, C. & Campbell, T.F. (1998). Non-word Repetition and Child Language Impairment. *Journal of Speech, Language, and Hearing Research,* 41, 1136-1145.

Dollaghan, C.A. (2004). Evidence-based practice in communication disorders: what do we know, and when do we know it? *Journal of Communication Disorders,* 37, 391-400.

Ellis Weismer, S., Tomblin, J.B., Zhang, X.Z., Buckwalter, P., Gaura Chynoweth, J. & Jones, M. (2000). Non-word Repetition Performance in School-Age Children With and Without Language Impairment. *Journal of Speech, Language, and Hearing Research,* 43, 865-878.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27, 861-874.

Gathercole, S.E. & Baddeley, A.D. (1990). Phonological Memory Deficits in Language Disordered Children: Is There a Causal Connection? *Journal of Memory and Language,* 29, 336-360.

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

Gerrits, E. (2004). *De Non-Word Repetition test (NWR-test).* Unpublished project description. Audiologisch Centrum, Academisch Ziekenhuis Maastricht.

Gerrits, E. (2005). Taaldiagnostiek bij meertalige kinderen: problemen en oplossingen. *Toegepaste Taalwetenschap in Artikelen*, 74, 169-177.

Gerritsen, F.M.E. (1994). V*TO Taalscreenings-instrument (TSI) voor 3-, 4-, en 5-jarigen. Handleiding en verantwoording.* Lisse: Swets & Zeitlinger.

Girbau, D. & Schwartz, R.G. (2008). Phonological working memory in Spanish-English bilingual children with and without specific language impairment. *Journal of Communication Disorders*, 41, 124-145.

Graf Estes, K., Evans, J.L. & Else-Quest, N.M. (2007). Diffences in the Non-word Repetition Performance of Children with and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research,* 50, 177-195.

Gray, S. (2003). Diagnostic accuracy and test-retest reliability of non-word repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders,* 36, 129-151.

Kohnert, K., Windsor, J. & Yim, D. (2006). Do Language-Based Processing Tasks Separate Children with Language Impairment form Typical Bilinguals?, *Learning Disabilities Research & Practice,* 21, 19-29.

Laing, S. & Kamhi, A. (2003). Alternative Assessment of Language and Literacy in Culturally and Linguistically Diverse Populations. *Language, Speech, and Hearing Services in Schools,* 34, 44-55.

Leonard, L.B. (2002). *Children with Specific Language Impairment.* Cambridge, Massachusetts: MIT Press.

Munson, B., Kurtz, B.A. & Windsor, J. (2005). The influence of Vocabulary Size, Phonotactic Probability, and Wordlikeness on Non-word Repetition of Children With and Without Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, 48, 1033-1047.

Oller, J.W., Yan, R., Badon, L.C. & Oller, S.D. (2006). Multicultural and Diversity Issues: Disproportionate Representations of Language Minorities in Disordered & Gifted Programs. *Preliminary handout for convention ASHA*. http://convention.asha.org/2006/handouts/855_SC13Oller_John_W._091011_110206113400.doc

Sackett, D.L., Haynes, R.B. & Tugwell, P. (1985). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston: Little Brown.

Salameh, E.K., Nettelbladt, U., Håkansson, G. & Gullberg, B. (2002). Language impairment in Swedish bilingual children: a comparison between bilingual and monolingual children in Malmö. *Acta Paediatrica*, 91, 229-234.

Smith, B. (2006). Precautions regarding non-word repetition tasks. Commentaries. *Applied Psycholinguistics*, 27, 584-587.

Thorn, A.S.C. & Gathercole, S.E. (1999). Language-specific Knowledge and Short-term Memory in Bilingual and Non-bilingual Children. *The Quarterly Journal of Experimental Psychology,* 52, 303-324.

Tomblin, J.B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E. & O'Brien, M. (1997). Prevalence of Specific Language Impairment in Kindergarten Children. *Journal of Speech, Language, and Hearing Research,* 40, 1245-1260.

Verhoeven, L., & Vermeer, A. (2006). *Verantwoording Taaltoets Alle Kinderen (TAK)*. Arnhem: Centraal Instituut voor Toetsontwikkeling.

Washington, J.A. & Craig, H.K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology,* 13, 329-340.

Other

Centraal Bureau voor de Statistiek (CBS) : http://www.cbs.nl

The Committee On Test Affairs Netherlands (COTAN): http://www.cotan.nl

APPENDIX - ROC ANALYSES

1. INTRODUCTION

By using likelihood ratios, we were able to compare the results of the present studies with those obtained by others, which was considered to be very useful. It was already mentioned that likelihood ratios cannot be derived easily from experimental data because of limited sample size (Choi, 1998) (see section 2.4.5.2 in this thesis). Thus, the rather small sample size in the present study has some serious implications for the calculations of likelihood ratios. Therefore, the results should be interpreted carefully. Another, possibly more accurate method (as suggested by Choi, 1998) of determining the diagnostic efficiency of a task is *ROC analysis* will also be performed.[43] Although likelihood ratios were already used in the examination of the accuracy of the task, it may be useful to also include ROC analysis. The results will be discussed in this Appendix.

2. RECEIVER OPERATING CHARACTERISTIC ANALYSIS

A perfect test, with a *sensitivity* of 100%, is able to identify all impaired children. This perfect test should also label all non-impaired children as having no impairment, being 100% *specific*. The levels of sensitivity and specificity taken together indicate the level of *accuracy* of the test. A test can also give false outcomes. A test with a low accuracy identifies too many 'false positives' (i.e., overidentifications) and/or 'false negatives' (i.e., underidentifications).

With Receiver Operating Characteristic (ROC) analysis can be determined how well a certain measure discriminates between impaired and non-impaired individuals, by calculation true and false positives and negatives. A two-by-two *confusion matrix* (also called 'contingency table') can be made (Fawcett, 2006). See table 1 for a basic confusion matrix.

[43] For a comparison between likelihood ratio and ROC analysis, see Choi (1998).

Table 1*. Confusion matrix. 'Y' and 'N' represent hypothesised classes for a positive and negative test result. The lower case 'p' and 'n' represent the true classes of the individuals that participated. Matrix based on Fawcett (2006), page 862.*

		True class	
		positive (impaired)	**negative (non-impaired)**
Hypothesised class	*yes* ≤ x	True Positives (TP)	False Positives (FP)
	no > x	False Negatives (FN)	True Negatives (TN)
Column totals		P	N

From this confusion matrix, some performance metrics can be derived (Fawcett, 2006). Some common performance metrics are, for example:

False Positive Rate = FP/N

True Positive Rate, or Sensitivity = TP/P

Specificity = TN/ (FP+TN) = 1-(FP/N)

Accuracy = (TP+TN)/(P+N)

The trade-off between true positive rates (i.e., 'benefits', Sensitivity) and the false positive rates (i.e., 'costs', 1-Specificity) for different threshold values can be plotted in an ROC graph. Three basic ROC graphs with different levels of accuracy are provided in figure 2. The point on the curve that is closest to the upper left-hand corner of the graph is the optimum classifier, with optimal (high) true and (low) false positive rates.

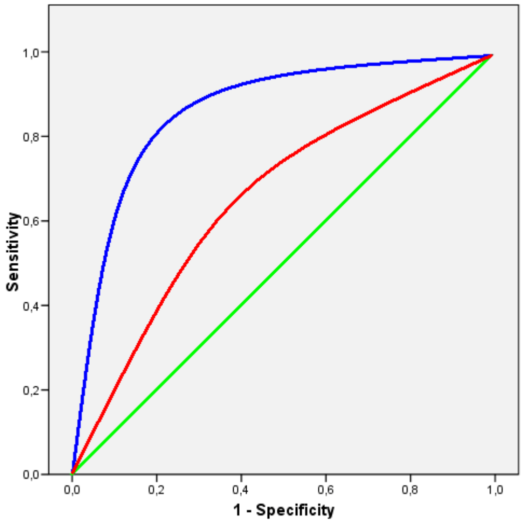

Figure 1*. Examples of three ROC curves. The curve on the left represents an excellent test, the curve in the middle a good test, and the curve on the right an unsuccessful test.*[44] *Graph based on picture on* http://gim.unmc.edu/dxtests/roc3.htm

[44] In this example, the classifiers are represented along a smooth line, in actual ROC calculations, the 'curve' is a step function (see the Results chapter). The larger the sample size, the smoother the curve.

The accuracy of a task is measured by the *area under the ROC curve*, which can be calculated by ROC analysis software.[45] A test can be classified in a scale ranging from 0.5 to 1, i.e., from 'unsuccessful' to 'excellent'.[46] The larger the area under the curve, the more accurate the diagnostic test. The following classifications can be provided given a certain level of the area under the curve:

0.9 – 1 = excellent
0.8 – 0.9 = good
0.7 – 0.8 = fair
0.6 – 0.7 = poor
0.5 – 0.6 = fail

For more information on ROC analysis, see Fawcett (2006).

3. Results

3.1 Screening Study part A

ROC analysis for the PPC scores

The true positive and false positive rates can be plotted in a ROC graph (see figure 6). In an ROC graph, the optimal classifier is the point closest to the upper left-hand corner. In this graph, the most accurate classifier of ruling in language impairment is a PPC score of ≤90, i.e., this threshold PPC score is the most accurate cut-off point of all possible cut-off points, with optimal true and false positive rates.

[45] See http://gim.unmc.edu/dxtests/roc3.htm (last checked 11 November, 2008).

[46] The calculations required for obtaining this value are rather complex, and are not further discussed in this thesis.

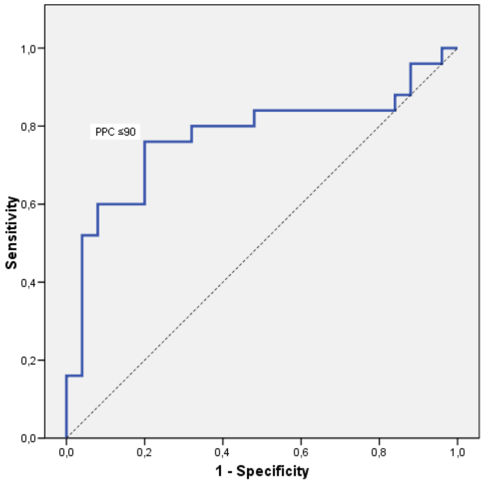

***Figure 2**. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC. Optimal classifier (PPC ≤90) is marked.*

Analysis further showed that the accuracy is 76% for a PPC score ≤90. For this threshold, the sensitivity (i.e., the probability that a language impaired child will be correctly identified by the task) is 76%. However, it does identify some false positives (24%). The level of specificity (i.e., the probability that a non-impaired child will be correctly identified by the task) for this threshold score is 76%. The area under the ROC curve in the graph above is 0.773. This is considered as 'fair' (i.e., not optimal).

A confusion matrix can be constructed in order to further examine the accuracy of the classifier 'PPC ≤90'. In table 2, the confusion matrix is provided with derivations for the threshold values. The matrix was computed following the method described by Fawcett (2006).

***Table 2**. Confusion matrix for the prediction: positive test result is PPC ≤ 90.*

		True class	
		positive	**negative**
Hypothesized class	*yes* ≤ 90	19	6
	no > 90	6	19
Column totals		25	25

According to most norm-referenced (language) tests, individuals obtaining a score of -1.3 SD or more below the *mean PPC score of the non-impaired multilingual group* (mean score was 92.45, with an SD of 4.58 see section 3.1.4.1, table 7, in the thesis' text) can be regarded as 'slightly impaired', which, in this case, is a PPC score of *86.50* or lower (92.45-(1.3*4.58)). A score of -1.7 SD or more below the mean score of the group can be regarded as 'moderately impaired', which is in this case a PPC score of *84.66* or lower (92.45-(1.7*4.58)). In tables 3 and 4, confusion matrices are provided with performance metrics for these two threshold values.

Table 3. *Confusion matrix for the prediction: positive test result is PPC ≤ 84.66 (mean - (1.7*SD)).*

		True class	
		positive	**negative**
Hypothesized class	*yes* ≤ 84.66	12	1
	no > 84.66	13	24
Column totals		25	25

The accuracy is lower for this specific cut-off point, namely 72%. The sensitivity is too low, with 48%, while the specificity is considerably high (96%). This implies that with a PPC cut-off score of 84.66, the level of true positives is low, but it is accompanied by a high level of true negatives.

Table 4. *Confusion matrix for the prediction: positive test result is PPC ≤ 86.5 (mean - 1.3*SD).*

		True class	
		positive	**negative**
Hypothesized class	*yes* ≤86.5	15	3
	no > 86.5	10	22
Column totals		25	25

At the PPC cut-off point ≤86.5, the accuracy is higher: 74%, with a sensitivity of 60% and a specificity of 88%.

Summarising, the –1.3SD and –1.7SD points were not identified as the optimum cut-off points, compared to the PPC ≤90 threshold score.

ROC analysis for the PPC per non-word length

The ROC analysis was also conducted for the individual non-word lengths.

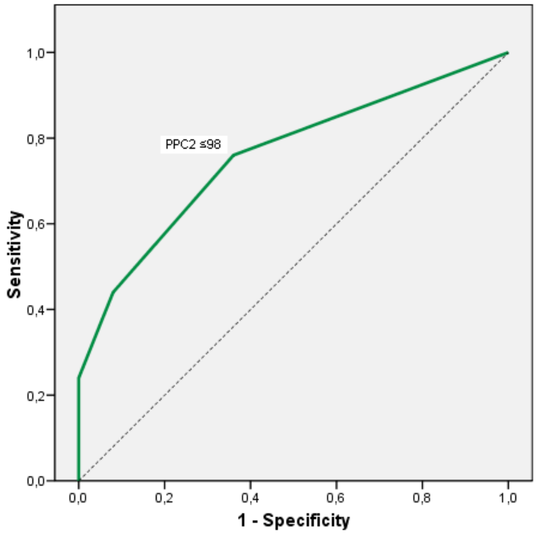

Figure 3. *Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC 2. Optimal classifier (PPC2 ≤98) is marked.*

The accuracy of this measure at the cut-off point PPC2 ≤98 is 70%, with a sensitivity of 76%. The specificity of the measure was 64%, which means that it did not identify many true negatives. The area under the curve is 0.758, which indicates that it is a ‘fair’ measure.

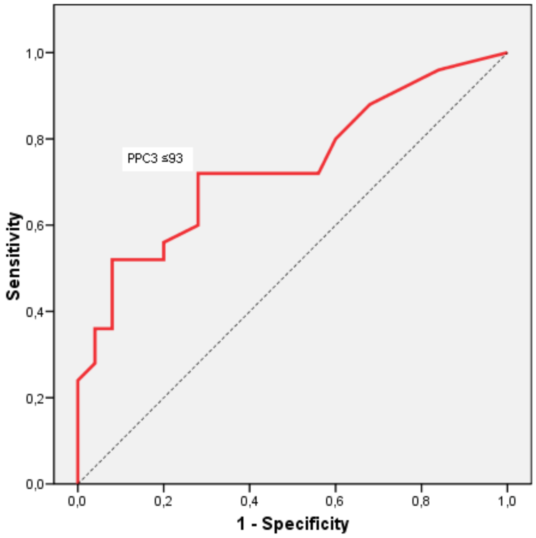

***Figure 4**. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC3. Optimal classifier (PPC3 ≤93) is marked.*

The accuracy of this measure at the cut-off point PPC3 ≤93 is 68%, with a sensitivity of 72%, and a specificity of 64%. The area under the curve is 0.737, which indicates that it is a ‘fair’ measure. Compared to the PPC2 score, this measure performed worse in identifying children with LI correctly.

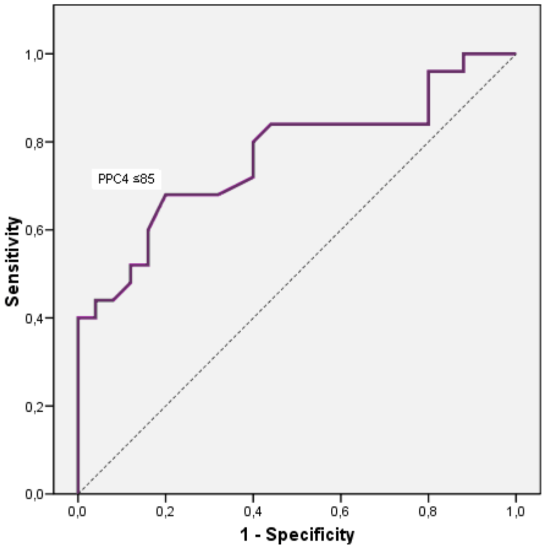

***Figure 5**. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC4. Optimal classifier (PPC4 ≤85) is marked.*

Compared to the PPC2 and PPC3 score, this measure performed better. At the cut-off point PPC4 ≤85, the accuracy is 74%, with a sensitivity of 68%, and a specificity of 80%. The high level of specificity indicates that the measure is good at identifying true negatives, i.e., the probability that a non-impaired child is correctly identified by the task is high. However, this is not desired for a screening task which desirably overidentifying impairment (such that further assessment can rule it out). The area under the curve is 0.768, which indicates that it is a 'fair' measure.

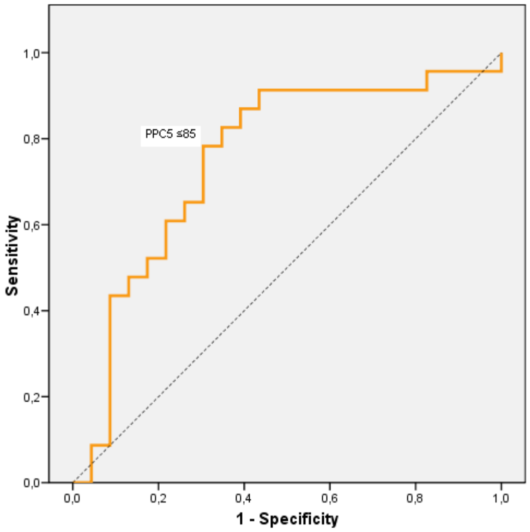

***Figure 6**. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC5. Optimal classifier (PPC5 ≤85) is marked.*

The PPC5 did not perform as well as the PPC4. At the cut-off point PPC5 ≤85, the accuracy is 68%, with a sensitivity of 76%, and a specificity of 65%. In this measure, the level of false positives is higher (higher 'overidentification'), which is good considering the use of the task as a screening tool. The area under the curve is 0.752, which indicates that it is a 'fair' measure.

Discussion

ROC analysis showed that the classifier PPC ≤90 is the most accurate measure. It is fairly accurate, with an accuracy level of 76%. ROC analyses for the individual non-word lengths showed that the PPC4 score is the most accurate measure with an accuracy of 74% (at a threshold PPC4 score of ≤85). PPC4 turns out to be a more accurate classifier than the other measures, with the largest area under the curve. At a threshold PPC4 score ≤85 it is the most accurate in discriminating between multilingual children with and without language problems, with an accuracy of 74%. PPC3 and PPC5 are the less accurate classifiers with an accuracy of 68% at threshold PPC scores of respectively ≤93 and ≤85. Therefore, it may be suggested that the PPC4 can be used for identifying (S)LI in multilingual children.

Both likelihood ratios (see section 3.1.4.2) and ROC analysis indicated that the PPC4 score is the most accurate measure. There are, however, some differences between both methods, for example with respect to the optimal cut-off points that can be used for ruling in language impairment. For the total PPC score, the likelihood ratio analysis showed PPC ≤ 80 to be the most accurate threshold score, while ROC analysis pointed out that a PPC score ≤ 90 performed best in identifying children with (S)LI. The 'optimal' cut-off points of the performance per non-word length also showed some differences between likelihood ratios and ROC analysis:

PPC2:	likelihood ratios: ≤ 93	ROC: ≤ 98
PPC3:	likelihood ratios: ≤ 86	ROC: ≤ 93
PPC4:	likelihood ratios: ≤ 78	ROC: ≤ 85
PPC5:	likelihood ratios: ≤ 70	ROC: ≤ 85

These differences are substantial, and have some serious implications for diagnosis and clinical utitlity. It means that the level of accuracy is dependent of which method of analysis is used. Therefore, results should always be treated carefully.

3.2 Screening study part B: Follow-up study

In this sample, the accuracy of the PPC score at the cut-off point PPC ≤85 is 71%. The sensitivity of the measure is 55%, which is not very high. The specificity is 76%. The area under the curve is 0.663, which indicates that it is a *poor* measure in distinguishing between children with and without (S)LI. It identifies few true positives, which may be a problem when using this task as a screening tool.

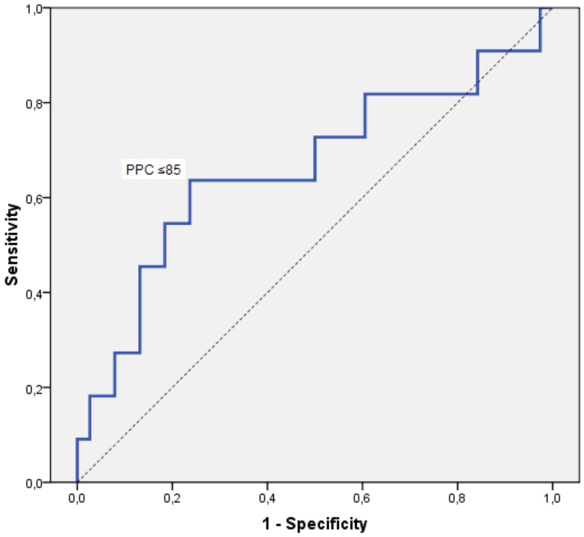

Figure 7*. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC. Optimal classifier (PPC ≤85) is marked.*

The PPC score of the four-syllable non-words was also examined:

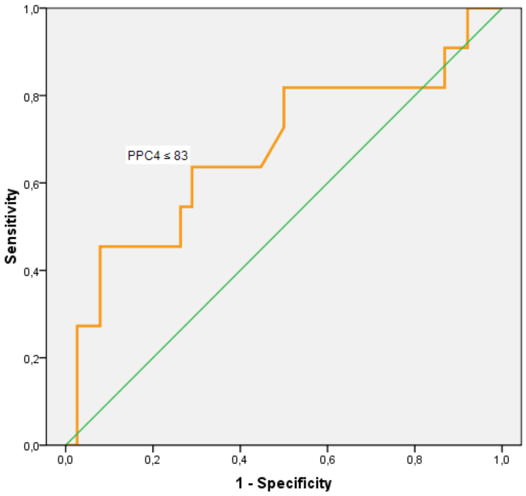

Figure 8. *Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC4. Optimal classifier (PPC4 ≤83) is marked.*

The accuracy of this measure at the cut-off point PPC4 ≤83 is 69%. The sensitivity of the measure is 64%, and the specificity is 71%. The area under the curve is 0.677, which indicates that it is a poor measure in distinguishing between children with and without (S)LI.

Discussion

In this sample, both the total PPC and the PPC of the four-syllable non-words were classified as *poor* measures by the area under the curve. This implies that neither are accurate enough measures. It might be the case that the unequal sample size of the groups was of influence in these results. Likelihood ratios indicated other optimal cut-off scores for both PPC (≤81, with ROC: ≤85), and PPC4 (≤74, with ROC ≤81).

3.3 Assessment Study

ROC analysis for the PPC scores

The ROC analysis showed that the diagnostic accuracy of the measure is *good*. Analysis showed that the accuracy is 79% for a PPC score ≤88. The sensitivity (the probability that a language impaired child will be correctly identified by the task) is 79%. The level of specificity (the probability that a non-impaired child will be correctly identified by the task) for this measure is 88%. These levels are rather high. The area under the ROC curve is 0.800. This is considered as ‘good’, which implies that this threshold score is accurate in discriminating children with SLI from children without SLI.

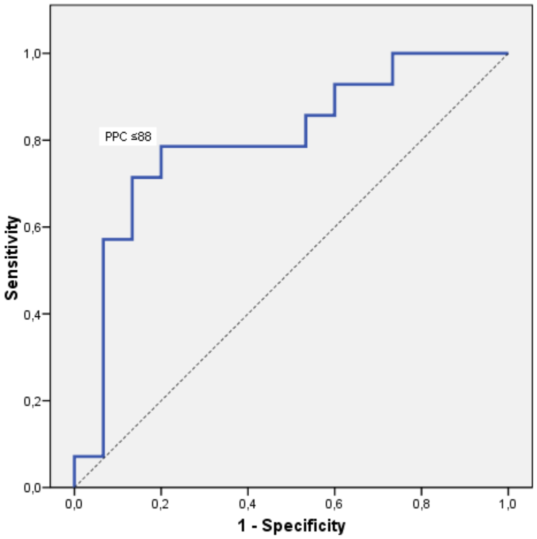

Figure 9. *Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC. Optimal classifier (PPC ≤88) is marked.*

ROC analysis for the PPC per non-word length

From the ROC graph (see figure 10), the optimal cut-off point of PPC2 cannot be derived easily. Therefore, calculations of sensitivity and specificity are not provided. The area under the curve is 0.686, which means that it is a *poor* measure.

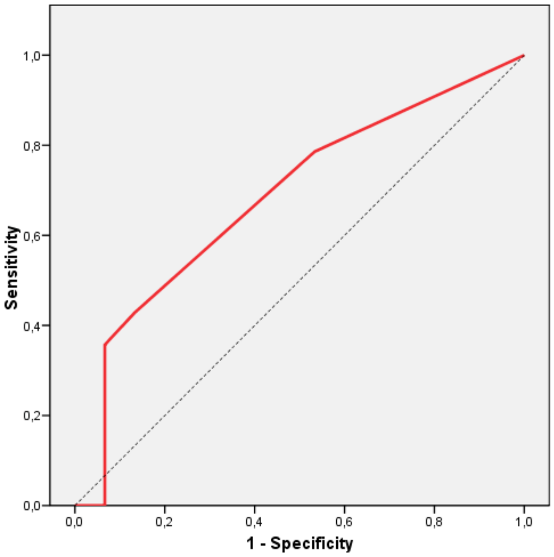

Figure 10. *Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC2. No optimal classifier is marked.*

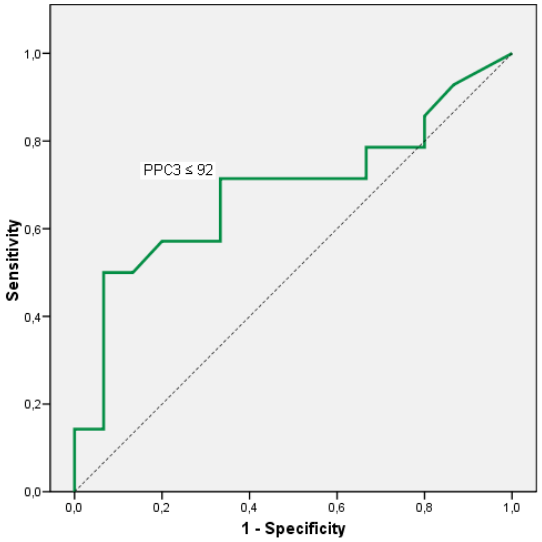

Figure 11*. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC3. Optimal classifier (PPC3 ≤92) is marked.*

The accuracy of this measure at the cut-off point PPC3 ≤83 is 69%. The sensitivity of the measure is 71%, and the specificity is 67%. The area under the curve is 0.686, which indicates that it is a poor measure.

Both PPC2 and PPC3 turned out to be *poor* measures, as indicated by the low level of 'area under the curve'. PPC 4 and PPC5 however, were better measures. Given the larger area under the curve (0.812), PPC4 is considered to be a *good* measure, with an accuracy of 76% at the PPC4 ≤85 threshold score. The sensitivity of the task is 79%, the specificity is 73%.

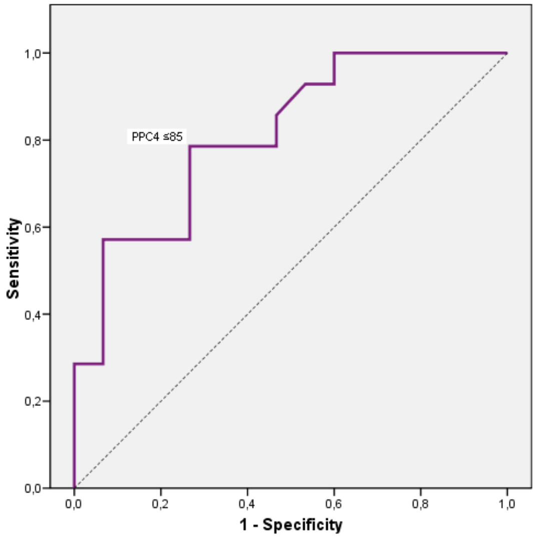

Figure 12*. Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC4. Optimal classifier (PPC4 ≤85) is marked.*

PPC5 is a *fair* measure (area under the curve: 0.781), with an accuracy of 79% at the PPC cut-off point ≤76. The sensitivity of the measure was 79%, and the specificity was 80%.

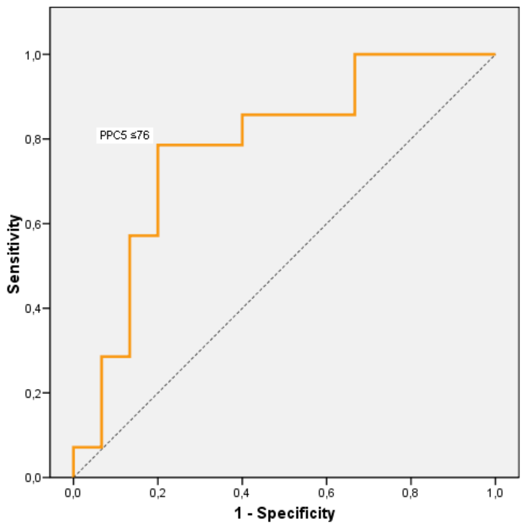

Figure 13. *Receiver Operating Characteristic curve: a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different possible cut-off points of the PPC5. Optimal classifier (PPC5 ≤76) is marked.*

Discussion

In this sample, both the total PPC and the PPC4 turned out to be good measures in distinguishing between multilinguals with and without SLI. Again, the PPC4 score was shown to be the most accurate measure. In comparison to the screening studies, in this study more clear-cut outcomes were found. This may be due to the fact that the children with SLI were included on the basis of a 'gold standard' diagnosis, which may have boosted the group differences. However, also in this study, the 'optimal' cut-off scores for ruling in language impairment were not similar in likelihood ratio analysis and ROC analysis. For the total PPC score, likelihood ratio calculation identified the cut-off point of PPC ≤84, while ROC analysis identified ≤ 88. The same was true for the PPC2 (likelihood ratios: ≤90, ROC: inconclusive), PPC3 (likelihood ratios: ≤90, ROC: ≤ 92), PPC4 (likelihood ratios: ≤79, ROC: ≤ 85), and PPC5 (likelihood ratios: ≤71, ROC: ≤76).

4. General discussion and conclusion

In the present study, PPC4 was shown to be the most accurate measure. This was confirmed by both the likelihood ratio calculations as well as the ROC analysis. However, the results obtained with the ROC analysis were not identical to those found with likelihood ratios calculations. This implies that the level of accuracy is dependent of which method of analysis is used. Therefore, results should always be interpreted very carefully. There cannot be given a conclusive 'optimal' cut-off score. The small sample size has probably influenced the results. Therefore, in future research, the same experiment should be repeated in a larger sample, in order to obtain reliable cut-off scores to rule in impairment. As is shown by the more solid, clear-cut results of the assessment study, inclusion in the 'impaired' group should always be based on assessment with a 'gold standard'.

5. LITERATURE

Choi, B.C.K. (1998). Slopes of a Receiver Operating Characteristic Curve and Likelihood Ratios for a Diagnostic Test. *American Journal of Epidemiology,* 148(11), 1127-1132.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.