A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children learning Dutch as a Second Language

Authors and affiliation:

Tessel Boerma[1],
Shula Chiat[2],
Paul Leseman[1],
Mona Timmermeister[1],
Frank Wijnen[1],
Elma Blom[1] (PI)

[1]Utrecht University
[2]City University London

Corresponding author:

Tessel Boerma, MA
Utrecht University, Department of Education and Learning
Heidelberglaan 1
3584 CS Utrecht
The Netherlands
0031-302534963

Abstract

Purpose This study evaluated a newly developed quasi-universal nonword repetition task (Q-U NWRT) as a diagnostic tool for bilingual children with language impairment (LI) who have Dutch as a second language. The Q-U NWRT was designed to be minimally influenced by knowledge of one specific language, in contrast to a language-specific (L-S) NWRT to which it was compared.

Methods 120 monolingual and bilingual children with and without LI participated (30 per group). A mixed-design ANOVA was used to investigate the effects of LI and bilingualism on the NWRTs. Receiver Operating Characteristic analyses were conducted to evaluate the instruments' diagnostic value.

Results Large negative effects of LI were found on both NWRTs, whereas negative effects of bilingualism only occurred on the L-S NWRT. Both instruments had high clinical accuracy in the monolingual group, but only the Q-U NWRT had high clinical accuracy in the bilingual group.

Conclusions This study indicates that the Q-U NWRT is a promising diagnostic tool to help identify LI in bilingual children learning Dutch as a second language. The instrument was clinically accurate in both a monolingual and bilingual group of children and seems better able to disentangle language impairment from language disadvantage than more language-specific measures.

Introduction

It is often a challenge for clinicians to determine whether or not a bilingual child has language impairment (LI). Results from studies suggest a tendency to misdiagnose bilingual children (Grimm & Schulz, 2014; Bedore & Peña, 2008; Salameh, Nettelbladt, Håkansson, & Gullberg, 2002; Smeets, Driessen, Elfering, & Hovius, 2010). Both under- and over-diagnosis of LI are reported, indicating that LI is either overlooked or that language delays are mistakenly ascribed to LI. Inappropriate education and treatment could be the undesirable result, emphasizing the need to improve the assessment of bilingual children. The present study examines a newly developed diagnostic tool for bilingual children learning Dutch as a second language (L2) that might support a more reliable diagnosis.

One of the reasons why identification of bilingual children with LI is challenging is that delays in language development can arise from impairment but also from external factors such as insufficient exposure to and, consequently, limited knowledge of the target language (Kohnert, 2010). Many cultural minority children grow up learning a first (minority) language at home and a second (majority) language outside of their homes in a different context (e.g. at day care or elementary school). The language skills of these children may vary immensely when they enter elementary school, depending on several factors such as the amount of bilingual exposure (Thordardottir, Rothenberg, Royard, & Naves, 2006) and the quality of input (Scheele, Leseman, & Mayo, 2010). Moreover, relative language ability in both languages changes as a function of age and learning opportunities and differs depending on which aspect of language is tested (e.g. Kohnert & Bates, 2002). The influence of these factors makes it difficult to determine the source of a child's language problems.

The diagnosis is further complicated by partially overlapping language profiles of typically developing (TD) bilingual children and monolingual children with LI. In the area of morphosyntax, LI-like patterns of acquisition of grammatical morphemes are found for TD

L2 learners of English (Paradis, 2005). Similarly, comparable developmental pathways in the acquisition of tense morphology and word order have been observed for children learning Swedish as L2 and monolingual Swedish children with LI (Håkansson, 2001; Håkansson & Nettelbladt, 1996). In Dutch, gender acquisition is reported to be vulnerable in both L2 learners and children with LI (Orgassa & Weerman, 2008) and the ability to inflect discriminated well in a monolingual, but not in a bilingual group of children in the Netherlands (Blom, de Jong, Orgassa, Baker, & Weerman, 2013). Finally, Grüter (2005) found no differences between L2 learners of French and monolingual French children with LI in their production and comprehension of object clitics. These behavioral similarities between the language profiles of bilingual children and children with LI can lead to cases of missed and mistaken identities (Gutiérrez-Clellen, 1996).

Bilingual TD children often also perform poorly on standardized language measures. Weaker performance can be explained by the distributed characteristic of bilingual learning, for instance concerning lexical knowledge (Oller & Pearson, 2002). The vocabulary size of bilingual children might be smaller compared to monolingual children when one language is measured, but similar when lexical knowledge in both languages is considered (Hoff et al., 2012). Another explanation for why bilingual children perform poorly on standardized measures is that these measures are "knowledge-dependent" (Campbell, Dollaghan, Needleman, & Janosky, 1997), disadvantaging bilingual children with less experience of the language of testing (e.g. Restrepo & Silverman, 2001). Thus, standardized language measures used for diagnosing LI in monolingual children may not be equally useful for bilingual children. Accordingly, language-based processing measures such as nonword repetition tasks (NWRT) have been proposed to complement traditional language tests. The advantage of such processing tasks is that they are less dependent on language knowledge, but tap into more basic cognitive underpinnings of language such as phonological processing and short-

term memory (Chiat, 2015; Gathercole, 2006). In this way, such measures remain sensitive to the presence of LI while minimizing the role of language-specific knowledge, hereby holding promise for differential diagnosis. The present study further explored this in a sample of monolingual Dutch children and bilingual children who were L2 learners of Dutch.

The nonword repetition task (NWRT)

NWRTs have been widely used as a measure of phonological short-term memory in various populations (for a review, see Coady & Evans, 2008). In this task, participants repeat nonsense words that conform to the phonotactics of their native language. It is a task that involves temporary storage and retrieval of novel strings and, in this manner, mimics word learning (Gathercole, 2006). This is reflected in the strong relationship between NWRT performance and vocabulary acquisition (e.g. Gathercole & Baddeley, 1989). The NWRT has also often been used to investigate differences between children with and without LI. Below, we review studies that have evaluated the use of a NWRT as a diagnostic instrument in both monolingual and bilingual children with and without LI (see also Chiat, 2015).

Nonword repetition in children with LI and its potential for differential diagnosis

The detrimental effect of LI on NWRT performance is robust and has been found in many studies and across languages (e.g. De Bree, Rispens, & Gerrits, 2007; Dispaldro, Leonard, & Deevy, 2013; Gathercole & Baddeley, 1990). As nonword repetition appears to be one of the most effective single predictors of language learning ability (Gathercole, 2006), several studies have investigated whether a NWRT can be used as a clinical marker to identify LI in children. Although results from some studies with monolingual children suggest that a NWRT cannot be used as a stand-alone tool due to sensitivity levels below 80% (e.g. Conti-Ramsden, Botting, & Faragher, 2001), others report sensitivity and specificity above 90%,

indicating high accuracy in identifying children with LI and TD respectively (e.g. Dispaldro et al., 2013; Gray, 2003; Kalnak, Peyrard-Janvid, Forssberg, & Sahlén, 2014).

As most work has been done with monolingual children from a cultural majority, the question remains whether the NWRT can also be used to support the diagnosis of children with different language experiences. Research on children from a cultural minority and children from low socio-economic status (SES) backgrounds suggests that it can. Ellis Weismer and colleagues (2000) examined a population-based sample of children and showed that a NWRT is a culturally nonbiased measure of language processing. Children from various cultural minorities performed similarly to children from the cultural majority on this NWRT even though their scores on standardized language measures were lower. Similar findings were reported by Campbell and colleagues (1997), who suggested that processing-dependent measures, such as a NWRT, could reduce bias in language assessment. Furthermore, research with children from low SES backgrounds confirms that differences in language experience have more influence on knowledge-based measures of vocabulary and grammar than processing-based NWRTs (Engel, Santos, & Gathercole, 2008).

NWRTs with bilingual children: potential pitfalls

Although some studies illustrate the diagnostic promise of NWRTs for the detection of LI in children with diverse language experiences, recent research with bilingual children also identifies potential pitfalls. Some studies did not find similar performance of monolingual and bilingual TD children on NWRTs. First, Kohnert, Windsor and Yim (2006) observed lower NWRT scores of bilingual children compared to monolingual children. Their study also included a group of monolingual children with LI and the diagnostic power of the measure was not sufficient to separate the bilingual TD children from the children with LI. Second, Engel de Abreu (2011) found no differences between monolingual and bilingual TD

children on working memory tasks, but did find group effects on the NWRT with higher scores in the monolingual group. The effect disappeared when vocabulary was controlled which suggests that performance on a NWRT relies on language-specific lexical knowledge.

Further research that looked into the relationship between language exposure and NWRT skills supports this claim. NWRT performance appears to be significantly influenced by language exposure (Gutiérrez-Clellen & Simon-Cereijido, 2010; Sharp & Gathercole, 2013; Summers, Bohman, Gillam, Peña, & Bedore, 2010). Due to individual differences in language exposure, Gutiérrez-Clellen and Simon-Cereijido (2010) found fair specificity (82%), but inadequate sensitivity (61% or lower) when a NWRT was used in just one language of the bilingual child. The study by Thordardottir and Brandeker (2013) partially supports these findings. They found significant associations between performance on an English NWRT and amount of English input in English-French bilingual children. Nonetheless, the strength of the association between NWRT performance and input was substantially weaker than associations between measures of vocabulary and input. In addition, no significant correlation was found between amount of French input and performance on a French NWRT. According to the authors, the difference between the English and the French NWRT in terms of their relation with amount of input can be explained by the characteristics of the nonword items. In contrast to the English NWRT, the items in the French NWRT were simple in terms of phonological complexity, syllable structure and stress pattern which made them relatively immune to effects of amount of exposure. Consequently, French NWRT performance of TD children was relatively high despite low levels of French exposure, resulting in an adequate sensitivity of 85% and slightly lower specificity of 79%.

Manipulating properties of NWRTs

The study by Thordardottir and Brandeker (2013) is not the only one suggesting that the diagnostic potential of a NWRT is dependent on particular characteristics of the nonwords. A meta-analysis by Graf-Estes, Evans and Else-Quest (2007) showed that the effect of LI, which should be maximized for optimal clinical value, is influenced by item properties such as syllable length and wordlikeness or phonotactic probability. Children with LI appear to perform weakly across all nonword lengths, but show greater difficulty with longer items (e.g. three-five syllables) compared to shorter ones (e.g. one-two syllables) relative to children with TD (e.g. Bishop, North, & Donlan, 1996). With regard to wordlikeness or phonotactic probability, results are less clear. Some studies have found a greater disadvantage for children with LI compared to TD children on low phonotactic probability items than on high phonotactic probability items (Munson, Kurtz, & Windsor, 2005), whereas others failed to find this difference (e.g. Coady, Evans, & Kluender, 2006). One factor that may also affect differences in the magnitude of LI is the scoring method used, although research on this topic is limited. Dispaldro and colleagues (2013) found that scoring the number of items correct produced a larger effect of LI than scoring the percentage of phonemes correct. Using a different NWRT, Graf-Estes and colleagues (2007) also scored children's responses with both methods and reported that the magnitude of group differences was greater when scoring the percentage of phonemes correct. Results from both studies show that it is important to take the scoring method into account, and suggest that the effect of scoring method may be different depending on the NWRT that is used.

While effects of LI need to be maximized in order to create a useful diagnostic tool, effects of bilingualism, such as amount of exposure, should be minimized. Item properties might also contribute to this. Correct repetition of items with low phonotactic probability or wordlikeness is influenced to a lesser extent by amount of exposure and sub-lexical knowledge than correct repetition of items with high phonotactic probability or wordlikeness

(Engel de Abreu, Baldassi, Puglisi, & Befi-Lopez, 2013; Messer, Leseman, Boom, & Mayo, 2010; Gathercole, 1995). This implies that one approach to diminishing the bilingual disadvantage on nonword repetition is by using items with low phonotactic probability or wordlikeness in the L2 of the child, at the same time allowing for a larger effect of LI (Munson, Kurtz, & Windsor, 2005). A downside of using this approach with bilingual children is its infeasibility, requiring a constant development of appropriate instruments due to the multitude of language combinations that are encountered in clinical practice.

A different approach to making NWRT performance relatively immune to effects of bilingualism is by creating an instrument that maximizes its applicability across languages (Chiat, 2015). Rather than incorporating specific features that only exist in a limited set of languages, such a test would be composed of sequences of phonemes that are "compatible with cross-linguistically diverse constraints on lexical phonology" (Chiat, 2015: 15). For instance, nonwords with simple CVCV structures are relatively universal in terms of syllable structure, whereas nonwords with consonants clusters (e.g. CCV) are more language-specific. Not all languages allow consonant clusters and children who have been exposed to these languages may have difficulty repeating such complex structures. Languages differ with respect to many other aspects of lexical phonology, such as word length, suprasegmental characteristics and segmental inventories. A NWRT that optimally uses the most common features across many languages may diminish reliance on amount of exposure in a particular language. In situations where clinical assessment is difficult due to the heterogeneity of children's language environments, a language-based processing measure that is not modelled on one specific language and is, in that sense, as universal as possible, might be informative. The present study investigated the performance of monolingual and bilingual children on such an instrument and assessed its clinical applicability.

The present study

This study used a quasi-universal (Q-U) NWRT (Chiat, 2015) that has recently been developed to support the assessment of bilingual children. The term Q-U NWRT is employed throughout this study to refer to a version of this task that is meant for children learning Dutch as their L2. The main purpose of the present research was to investigate the effects of LI and bilingualism on Dutch children's performance on this Q-U NWRT relative to a language-specific (L-S) NWRT. Moreover, we aimed to evaluate the clinical potential of both tasks. To validate the Q-U NWRT, we also examined the effects of syllable length and, in view of future clinical use, we explored which scoring method would prove to be most effective in discriminating between children with and without LI. Effects of phonotactic probability were not analyzed as this factor is not manipulated within the Q-U NWRT.

Considering that previous research has shown robust effects of LI across many different NWRTs (Graf-Estes et al., 2007), we predicted that scores on both the Q-U and the L-S NWRT would reveal negative effects of LI, with larger effects as item length increases. However, a difference between the two NWRTs was anticipated with respect to effects of bilingualism. Performance on the L-S NWRT relies on language-specific knowledge of Dutch and hence, previous experience with Dutch. Therefore, bilingual children were expected to be disadvantaged by the L-S NWRT relative to monolingual children, implying a negative effect of bilingualism. For the Q-U NWRT, performance of monolingual and bilingual children was predicted to be similar. Regarding the clinical potential of the tasks, we hypothesized that the Q-U NWRT would have better diagnostic accuracy, sensitivity and specificity compared to the L-S NWRT in a bilingual group of children as performance on the latter partially depends on external factors that are not associated with LI.

Methods

Participants

This study included 120 children of whom the majority were 5 and 6 years old. Monolingual children with TD (MOTD), monolingual children with LI (MOLI), bilingual children with TD (BITD) and bilingual children with LI (BILI) were compared (*N*=30 in each of four groups). Children were regarded as monolingual if both parents always spoke Dutch to them. Children were regarded as bilingual if one or both parents were native speakers of another language than Dutch and spoke their native tongue with the child for an extensive period of the child's life. The bilingual children with and without LI all learned Dutch in an environment where Dutch is the majority language. The groups were matched on exposure to Dutch before the age of 4 and current exposure to Dutch at home (Table 1) based on a parental questionnaire (Questionnaire for Parents of Bilingual Children (PaBiQ); COST Action IS0804, 2011)[1]. Exposure to Dutch before the age of 4 was measured as the amount of Dutch input relative to the total amount of language input that the child received before this age (both inside and outside home context). Current exposure to Dutch at home was measured as the amount of Dutch input relative to the total amount of language input that the child heard from its mother, father, siblings and other adults that had frequent contact with the child. There were no significant differences between the bilingual groups in exposure to Dutch before the age of 4 ($F(1,58) = .06$, $p = .81$, $\eta_p^2 = .00$) nor in current exposure to Dutch at home ($F(1,58) = 1.9$, $p = .18$, $\eta_p^2 = .03$). The first languages of the bilingual TD children included Turkish (*N*=13), Tarifit-Berber (*N*=11) and Moroccan Arabic (*N*=6). The first languages of the bilingual children with LI were Turkish (*N*=8), Moroccan Arabic (*N*=7), Egyptian Arabic (*N*=3), Tarifit-Berber (*N*=2), Dari (*N*=2), Pashto (*N*=1), Suryoyo (*N*=1), Kirundi (*N*=1), Russian (*N*=1), Chinese (*N*=1), Portuguese (*N*=1), Danish (*N*=1) and Frisian (*N*=1).

[1] This questionnaire is the short version of a longer questionnaire piloted by research groups in several countries within COST Action IS0804, which was in part based on the ALEQ (Paradis, 2011) and the ALDeQ (Paradis et al., 2010)

TD children were recruited via regular elementary schools. Children with LI were recruited through two national organizations in the Netherlands (Royal Dutch Kentalis and Royal Auris Group) that provide diagnostic, care and educational services for children with language difficulties. All children with LI had been diagnosed by licensed professionals on the basis of a standardized protocol (Stichting Siméa, 2014). A score of at least 2 standard deviations (*SD*) below the mean on an overall score of a standardized language assessment test battery or a score of at least 1.5 *SD* below the mean on two out of four subscales of this standardized language assessment were the inclusion criteria for LI in this study. The standardized instruments that were used for diagnosis were the Dutch version of the Clinical Evaluation of Language Fundamentals (CELF-4-NL; Kort, Schittekatte, & Compaan, 2008) or the Schlichting Test for Language Production and Comprehension (Schlichting & Lutje Spelberg, 2010ab). The children with LI attended either special education (N=58) or regular education with ambulatory care (N=2; one bilingual child and one matched monolingual child). Exclusion criteria were the presence of a hearing impairment, intellectual disability and severe articulatory difficulties as determined by a certified professional.

The four groups of children were matched on age in months, nonverbal IQ and SES. Nonverbal IQ was measured with the short version of the Wechsler Nonverbal-NL (Wechsler & Naglieri, 2008) and SES was based on the education level of both parents. In cases where precise matching on child level was not possible, a child was matched on group level. Group characteristics are presented in Table 1. There were no significant age differences ($F(3,116) = .14$, $p = .94$, $\eta_p^2 = .00$) nor nonverbal IQ differences ($F(3,116) = 1.3$, $p = .28$, $\eta_p^2 = .03$) between any of the four groups. SES did differ significantly ($H(3) = 8.06$, $p = .045$), reflecting lower SES in the bilingual TD group compared to the monolingual TD group. Furthermore, there were significant differences between the groups with regard to gender due to the relatively small number of boys in the BITD group (χ^2 (3, N=120) = 8.9, $p = .03$).

[Insert Table 1 here]

Information on the Dutch language abilities of the children is provided by performance on three standardized measures testing receptive vocabulary (PPVT-III-NL; Schlichting, 2005), grammatical morphology (TAK Word Formation; Verhoeven & Vermeer, 2001) and knowledge of function words and word order (TAK Sentence Formation; Verhoeven & Vermeer, 2001). Norm-referenced quotient scores for the PPVT-III-NL and raw scores for both TAK measures are presented in Table 2. For the TAK measures, raw scores of the monolingual and bilingual groups were compared to norm groups that heard Dutch or a different language at home respectively (Figure 1).

[Insert Table 2 and Figure 1 here]

Instruments

Quasi-Universal

The Quasi-Universal (Q-U) NWRT (Chiat, 2015) was designed in collaboration with members of the COST Action IS0804 (Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment). The task contains 16 items that vary in length from two to five syllables. The items are constructed in such a way that they are in accordance with various constraints on lexical phonology in many languages (Chiat, 2015). The simple CVCV sequences of the items contain a limited range of consonants and vowels that occur in many languages. The designers of the Q-U NWRT offer a format that allows for adaptation to any particular language. For each of the 16 items, a set of four to six candidate options have been constructed from which a selection can be made. These candidate options are variations for each item that are matched for length, syllable structure, and segmental categories. This allows for some flexibility in case a proposed item is a real word in the relevant language or one of the segmental options does not occur in the target language. Once

particular options are selected, the phonemes within the items are produced with the phonetic qualities of the relevant language. Thus, the items still have certain language-specific characteristics, making them *quasi*-universal (for further discussion, see Chiat, 2015).

A Dutch version of this task was constructed for the purpose of this study. Candidate items that were real words in either the majority language Dutch or the three most common minority languages in this study (Turkish, Tarifit-Berber and Moroccan Arabic) were excluded, covering all languages of the TD children and 78% of the languages of the children with LI. Furthermore, all candidate items that included the plosive /p/ or the velar stop /g/ were excluded as /p/ does not originally occur in Tarifit-Berber and /g/ is very uncommon in Dutch. Sixteen items were chosen and recorded by a female native speaker of Dutch, producing the vowels and consonants with their Dutch phonetic qualities. Language-specific prosodic patterns were avoided by stressing all syllables equally, producing them with even length and pitch, apart from the final syllable lengthening which characteristically marks the end of an utterance (Chiat, 2015). In this way, a possible effect of language-specific prosodic knowledge, disadvantaging children with less experience in that language, was reduced. The final selection of items is presented in Appendix 1.

Language-Specific

The Q-U NWRT was compared to an adapted version of a task developed by Rispens and Baker (2012). This task is modelled on specific properties of Dutch and is thus an example of a Language-Specific (L-S) NWRT. The task was not designed for diagnostic purposes, but to investigate (sub)lexical processing in TD children, children with LI and children with reading problems. The original task contains 40 items equally divided between items of two to five syllables and of high and low phonotactic probability according to the Dutch phonotactic frequency database (Adriaans, 2006). Items in the L-S NWRT do not include consonant

clusters, which is analogous to the Q-U NWRT, but they do include final consonants and are CV…CVC sequences. The items are thus one phoneme longer than the items from the Q-U NWRT. Furthermore, items followed the regular Dutch stress pattern. For the present study, 24 out of the 40 items were selected to prevent fatigue due to the length of the task. The distribution of items with respect to syllable length and phonotactic probability was maintained for optimal differentiation. The final selection thus comprised 12 items of high and 12 of low phonotactic probability, each with three items per syllable length (2-5). The phonotactic probability of the items within the Q-U NWRT was also checked for Dutch and turned out to be higher than the low phonotactic probability items of the L-S NWRT, but lower than the items with high phonotactic probability (Q-U: -1.43; L-S High: -1.28; L-S Low: -2.02). This short version of the L-S NWRT was recorded by the same female native speaker of Dutch who also recorded the Q-U NWRT. The items are presented in Appendix 2.

Procedures and scoring

This research was screened by the Standing Ethical Assessment Committee of the Faculty of Social and Behavioral Sciences at Utrecht University. Criteria were met and further verification was not deemed necessary. Parents of participants signed an informed consent.

All participants were individually tested in a quiet room at their school. They completed a battery of tests in two separate sessions each lasting approximately one hour. The Q-U and L-S NWRT were the first tasks of the second session. Other tasks included working memory and attention tasks and will not be discussed in the current paper. The presentation format of the NWRTs was adapted from Engel de Abreu and colleagues (2013). Children were presented with a cartoon ‘alien’ that spoke a strange foreign language and wanted to teach this to the children. Two practice items familiarized the children with the procedure. This was followed by the first block with the first NWRT. After this, there was a

short break in which a friend of the alien was introduced that spoke a different nonsense language. Subsequently, the block with the second NWRT started. The order of the blocks was counterbalanced; half of the children began with the Q-U NWRT and half of the children with the L-S NWRT. Items within both NWRTs were prerecorded and presented in a pseudo-randomized order. Children were only allowed to hear each nonword once.

All responses of the children were recorded with a highly sensitive microphone (Samson Go Mic). They were transcribed offline and scored using two scoring methods: 1) percentage of items correct (PIC) and 2) percentage of phonemes correct (PPC). Whole-item accuracy was represented by an all-or-nothing score of correct or incorrect responses. Repetitions that included omissions or substitutions were considered incorrect, whereas repetitions with only additions were judged as correct as they do not reflect loss of information (Dollaghan & Campbell, 1998). Systematic substitutions of phonemes, reflecting articulation ability, were allowed. Second, the percentage of phonemes correct per item was calculated. The same procedure regarding omissions, (systematic) substitutions and additions was applied as with the first scoring method. In cases where the structure of an item was not maintained, syllable sequences in a child's response were aligned to the best corresponding target syllables before the number of phonemes correct was scored.

A second independent rater scored 75% of the data. For percentage of phonemes correct, the scores of the two raters overlapped in 94% of the cases for the Q-U NWRT and in 93% of the cases for the L-S NWRT. The intra-class correlation coefficient (ICC; absolute) was excellent (Q-U: .99; L-S: .98). For the percentage of items correct, scores of the two independent judges overlapped in 98% of the cases for both NWRTs. Again, the ICC was excellent (Q-U: .97; L-S: .96). Instances of disagreement were resolved by consensus.

Data analysis

All statistical analyses were performed using SPSS 22 (IBM Corp., 2013). Exploration of the data revealed that the variables PIC and PPC for both NWRTs were skewed. A square root transformation was applied to the data after which most variables were normally distributed, apart from the variables for two and five syllables. Therefore, non-parametric tests were done to check whether this affected the results, but no differences between parametric and non-parametric tests were found. The transformed variables will thus be used in all analyses with parametric tests. NWRT performance was not correlated with either SES or nonverbal IQ in any group, hence there was no need to control for prior differences between the groups.

To investigate the effects of LI, bilingualism and syllable length on the NWRTs, a 2x4x4 mixed-design analysis of variance (ANOVA) was used. The analysis was run with Version of the NWRT as a within-subject factor with two levels (Q-U NWRT and L-S NWRT), Syllable Length as a within-subject factor with four levels (two, three, four and five syllables) and Group as a between-subject factor with four levels (MOTD, MOLI, BITD and BILI). Subsequently, post-hoc analyses (one-way ANOVAs and repeated measures ANOVAs) were conducted in case significant interactions between the three factors in the model were observed. Effect sizes are calculated using Cohen's *d* (1998).

A second analysis evaluated the clinical potential of the NWRTs by investigating to what extent the instruments predicted the absence or presence of LI in the monolingual and bilingual group of children. Receiver Operating Characteristic (ROC) curves were used to determine the optimal cut-off score for each NWRT associated with the highest sensitivity and specificity of the instrument (after Gutiérrez-Clellen & Simon-Cereijido, 2010). The ROC curve plots sensitivity and specificity for different NWRT scores that are observed in the data. Subsequently, the score that maximizes both sensitivity and specificity (as close to 1 as possible) is chosen as the optimal cut-off score of the instrument. For the purpose of this study, sensitivity can be defined as the proportion of children who are diagnosed with LI and

score below the optimal cut-off score whereas specificity is the proportion of TD children who score above this cut-off score. These measures thus indicate how well the instruments assign a child to the correct group. Sensitivity and specificity between 80% and 89% are considered fair, while rates above 90% are good (Plante & Vance, 1994). Likelihood ratios were also calculated to evaluate to what extent the instruments change the probability of the presence or absence of LI. In addition, diagnostic test accuracy of the NWRTs is estimated by the Area Under the Curve (AUC). The AUC is the probability that a randomly selected child with LI will score lower than a randomly selected child with TD and thus depends on the ability of the instruments to classify children with and without LI correctly (Tape, 2008). Tape's (2008) criteria for diagnostic test accuracy are applied (AUC of 1 = perfect; AUC of .90-1 = excellent; AUC of .80-.90 = good; AUC of .80-.70 = fair; AUC of .60-.70 = poor; AUC of 0.5 ≤ worthless).

Results of the above analyses using the two scoring methods (percentage of phonemes correct (PPC) and percentage of items correct (PIC)) were compared to identify the most effective method. Results from the outcome variable PPC are presented first. Subsequently, only clear differences for PIC compared to PPC are discussed to avoid redundancy. To control for possible misdiagnosis in our sample, all analyses described above were also conducted for a subsample of the participants, excluding children with LI and TD that scored unexpectedly high or low respectively on the TAK language measures. Analyses yielded similar results and are therefore not reported.

Results

Effects of LI, bilingualism and syllable length

Percentage of phonemes correct (PPC)

Table 3 presents the means and *SD*s of the PPC performance of the four groups of children on the two versions of the NWRT. Results revealed a significant main effect of Version ($F(1,116) = 148.5$, $p < .001$, $\eta_p^2 = .56$), a significant main effect of Syllable Length ($F(3,348) = 189.9$, $p < .001$, $\eta_p^2 = .62$) and a significant main effect of Group ($F(3,116) = 46.8$, $p < .001$, $\eta_p^2 = .55$). Significant interaction effects of Version × Group ($F(3,116) = 8.6$, $p < .001$, $\eta_p^2 = .18$), Syllable Length × Group ($F(9,348) = 2.0$, $p = .04$, $\eta_p^2 = .05$) and Version × Syllable Length ($F(3,348) = 23.0$, $p < .001$, $\eta_p^2 = .16$) were found and will be discussed below. The three-way interaction was not significant.

Pairwise comparisons showed that, independent of Group or Syllable Length, children's performance on the Q-U NWRT was better than on the L-S NWRT ($p < .001$). Furthermore, independent of Group or Version, performance deteriorated as item length in syllables increased ($p < .001$). Finally, the main effect of group showed that the two TD groups outperformed the two LI groups ($p < .001$). There were no statistically significant differences between monolingual and bilingual groups when the versions of the NWRTs and syllable lengths were collapsed.

[Insert Table 3 here]

The significant interaction between Version × Group indicated that effects of LI and bilingualism on performance of the NWRT differed depending on the version of the NWRT. Post-hoc analyses showed significant main effects of Group for both NWRTs separately (Q-U: $F(3,116) = 38.1$, $p < .001$, $\eta_p^2 = .50$; L-S: $F(3,116) = 40.0$, $p < .001$, $\eta_p^2 = .51$). Table 4 presents the results of the pairwise comparisons that show the effects of LI and bilingualism on the two versions of the NWRT. Children with LI performed significantly worse on both NWRTs in comparison with their TD peers. In the monolingual group, the effects of LI were largest for the L-S NWRT whereas in the bilingual group the Q-U NWRT led to the largest effect size. Furthermore, a significant negative effect of bilingualism was found for the L-S

NWRT in the TD group: the bilingual TD children scored lower than their monolingual peers. However, there were no differences between monolingual and bilingual TD children with respect to performance on the Q-U NWRT. Finally, the monolingual and bilingual groups with LI did not differ on either task.

[Insert Table 4 here]

Post-hoc analyses were performed to unpack the interaction between Syllable Length × Group (Figure 2) and showed larger effects of syllable length for monolingual and bilingual children with LI (Q-U: $\eta_p^2 = .57$; L-S: $\eta_p^2 = .59$) in comparison with their TD peers (Q-U: $\eta_p^2 = .45$; L-S: $\eta_p^2 = .31$). The TD groups significantly outperformed the LI groups on all syllable lengths (all $p < .01$). Within the LI group, monolingual and bilingual children did not differ on any of the syllable lengths. Within the TD group, the bilingual children performed significantly below the monolingual children on language-specific items with three, four and five syllables ($p < .01$). Other differences were not significant.

[Insert Figure 2 here]

Finally, effects of syllable length appeared to be different depending on the version of the NWRT. Two repeated measures ANOVAs for the NWRTs separately both revealed significant main effects of Syllable Length (Q-U: $F(3,357) = 122.6$, $p < .001$, $\eta_p^2 = .51$; L-S: $F(3,357) = 92.3$, $p < .001$, $\eta_p^2 = .44$). For the L-S NWRT, all syllable lengths differed from each other ($p < .001$) apart from syllable length four and five. For the Q-U NWRT, syllable lengths two and three did not differ whereas all other differences were significant ($p < .001$).

Percentage of whole-items correct (PIC)

Table 5 presents the means and *SD*s of the PIC performance of the four groups of children on the two versions of the NWRT. Results for this scoring method were similar to previous analyses with PPC and also revealed a significant main effect of Version ($F(1,116) = 274.9$, p

< .001, η_p^2 = .70), Syllable Length (F(3,348) = 345.0, p < .001, η_p^2 = .75) and Group (F(3,116) = 43.2, p < .001, η_p^2 = .53). Significant interaction effects of Version × Group (F(3,116) = 7.4, p < .001, η_p^2 = .16), Syllable Length × Group (F(9,348) = 5.1, p < .001, η_p^2 = .12) and Version × Syllable Length (F(3,348) = 9.2, p < .001, η_p^2 = .07) were found. The three-way interaction was not significant. Pairwise comparisons for PIC yielded the same outcomes as for PPC with the exception of a marginally significant difference (p = .047) between the monolingual and bilingual TD groups when the Versions of the NWRT and Syllable Length were collapsed.

[Insert Table 5 here]

Results from the post-hoc analyses showed a larger effect of LI in the monolingual group on the Q-U NWRT when PIC was employed compared to PPC (d = 2.44 vs. d = 2.12 respectively), but a smaller effect of LI in the bilingual group on the L-S NWRT (d = 1.20 vs. d = .92 respectively). Moreover, effects of syllable length became similar for the TD and LI groups when analyses were done with PIC. Overall patterns, however, were comparable.

Diagnostic accuracy, sensitivity and specificity

Percentage of phonemes correct (PPC)

Cut-off scores, sensitivity and specificity of the NWRTs are presented in Table 6. Although specificity was good for the L-S NWRT (93%) in the bilingual group of children, sensitivity was inadequate (63%). Over 35% of the bilingual children with LI were misclassified by the language-specific task. For the Q-U NWRT, specificity (93%) was the same in this group of children and sensitivity (83%) was clearly better. In the monolingual group, specificity and sensitivity were high for both NWRTs, with the highest levels for the L-S NWRT.

[Insert Table 6 here]

Further examinations of the ROC curves identified large Areas Under the ROC Curve (AUC) for both NWRTs in the monolingual group and indicated excellent test accuracy (Q-U: area = .94, SE = .03, $p < .001$, CI 95 = .89 - .100; L-S: area = .95, SE = .03, $p < .001$, CI 95 = .91 - .100). In the bilingual group, test accuracy for the Q-U NWRT was excellent (area = .90, SE = .04, $p < .001$, CI 95 = .81 - .99), whereas it was fair for the L-S NWRT (area = .79, SE = .06, $p < .001$, CI 95 = .68 - .91).

Percentage of whole-items correct (PIC)

With the exception of some small differences, results were largely similar for the two scoring methods. Sensitivity (97%) increased for the Q-U NWRT in the monolingual group when PIC was employed, whereas it decreased slightly for the L-S NWRT (both 90%). In the bilingual group, sensitivity increased to 87% as specificity decreased to 83% for the Q-U NWRT. For the L-S NWRT, we observed similar patterns: sensitivity increased (77%) and specificity decreased (73%).

The AUC remained large for both NWRTs in the monolingual group (Q-U: area = .95, SE = .03, $p < .001$, CI 95 = .89 - .100; L-S: area = .95, SE = .03, $p < .001$, CI 95 = .90 - .100). In the bilingual group, test accuracy slightly decreased for both NWRTs, now ranging from good to fair (Q-U: area = .89, SE = .05, $p < .001$, CI 95 = .79 - .98; L-S: area = .76, SE = .07, $p < .001$, CI 95 = .63 - .89).

Discussion

The main aim of the present study was to evaluate the clinical applicability of the Dutch version of a newly developed quasi-universal nonword repetition task (Q-U NWRT) in a group of monolingual and bilingual children with and without LI. The new task was compared with a more traditional language-specific (L-S) NWRT. The Q-U NWRT was

designed to maximize phonological features most commonly represented across languages. Hence, performance on the Q-U NWRT should be minimally influenced by knowledge of one specific language, in contrast to performance on the L-S NWRT.

With respect to investigating effects of LI and syllable length, results were largely in line with our predictions. Large differences between children with and without LI, both monolingual and bilingual, were found on both NWRTs, strengthening the case for nonword repetition as a clinical marker of LI (Conti-Ramsden et al., 2001). TD children outperformed children with LI on all syllable lengths. When using percentage of phonemes correct as scoring method, the difference between the children with and without LI was largest for the longer items for both NWRTs. These findings are consistent with previous research (e.g. Dollaghan & Campbell, 1998) and show that the newly developed Q-U NWRT functions comparably to other NWRTs.

The observed effects of bilingualism in the TD groups corresponded with the predicted performance pattern. Due to item characteristics, children in all groups scored lower on the L-S NWRT than on the Q-U NWRT, but the L-S NWRT was particularly difficult for the bilingual TD children. The monolingual TD children outperformed their bilingual TD peers on the L-S NWRT, whereas their performance on the Q-U NWRT did not differ. The bilingual children were presumably disadvantaged on the L-S NWRT due to having less language-specific knowledge of Dutch to support memory representations needed to successfully repeat items from the L-S NWRT. This finding is consistent with previous work (Engel de Abreu et al., 2013; Engel de Abreu, 2011; Kohnert et al., 2006) and is also apparent in the scores on the language tests (see Table 2), which are substantially lower for the bilingual TD children than for their monolingual TD peers. Knowledge of Dutch did not appear to be as important for the Q-U NWRT as the two TD groups performed similarly.

In contrast to the TD group, no effect of bilingualism was found in the LI group on either NWRT, suggesting that the bilingual children with LI are not additionally disadvantaged by the L-S NWRT. A possible explanation for this is that the effect of language impairment outweighs the effect of language-specific knowledge. As a consequence of their impairment, both monolingual and bilingual children with LI have less language-specific knowledge of Dutch compared to children with TD. The impact of this effect on NWRT performance could be much more extensive than the effect of dual language learning, as is also indicated by the effect sizes of LI (d=2.50) and bilingualism (d=1.20). These findings are in line with other research that does not support a double delay in bilingual children with LI (Paradis, 2010). Another possible explanation as to why no additional effect of bilingualism was found in the groups with LI is potential misdiagnosis, reflected by the overrepresentation of bilingual children in special education (Smeets et al., 2010). Incorrectly diagnosed bilingual children with LI might be positively influencing NWRT performance, hereby masking effects of bilingualism. Even though we cannot rule out this possibility, analyses that excluded possibly misdiagnosed children did not support this explanation.

Although group comparisons are important, assessment in the clinical practice is always done at the level of the individual child. Overall, diagnostic accuracy proved to be excellent for both tasks in the monolingual sample. Moreover, sensitivity and specificity reached adequate levels. However, results for the two NWRTs diverged within the bilingual group. Over 35% of the bilingual children with LI were misclassified by the L-S NWRT. This replicates other work that also reported low sensitivity of a language-specific NWRT in a bilingual group of children (Gutiérrez-Clellen & Simon-Cereijido, 2010; Kohnert et al., 2006) and suggests that a language-specific NWRT ought to be used with caution. The diagnostic potential of the Q-U NWRT remained powerful in the group of bilingual children with adequate levels of sensitivity and specificity. The finding that the Q-U NWRT was

sensitive to LI in a heterogeneous group of children with diverse linguistic backgrounds suggests that this instrument is to be preferred over a language-specific task when used in clinical practice with bilingual children.

If the Q-U NWRT is used for clinical practice, it is important to know which method of scoring is most sensitive to LI. The results show that both scoring methods discriminated well between children with and without LI in both the group of monolingual and bilingual children. The number of items correct actually achieved the highest levels of sensitivity and specificity for the Q-U NWRT within the monolingual group of children, in line with other research (Dispaldro et al., 2013). Within the bilingual group, results for the two scoring methods were very similar. The practical implication of this finding is that scoring the number of items correct seems to work well for the Q-U NWRT, facilitating online scoring and making administration of the task less time-consuming.

The results of the present study indicate that the Dutch version of a Q-U NWRT can be a valuable tool for identifying children with LI that are L2 learners of Dutch. Further research in other language contexts, using different versions of the instrument, is needed to strengthen our findings. Furthermore, a limitation of the current research is that children were already diagnosed with LI, by stringent criteria. Many studies use a cut-off of -1.25 *SD* on two language domains as their inclusion criteria for LI (after Tomblin, Records & Zhang, 1996), whereas this study employed -1.5 *SD*. This might have enlarged the difference between the TD and LI groups, positively influencing the diagnostic accuracy of the instruments. The use of predefined groups instead of a population sample might have a similar effect. Previous research used a NWRT that distinguished children with and without LI excellently in predefined groups (Dollaghan & Campbell, 1998), but worked less well in a population-based sample (Ellis Weismer et al., 2000). To validate the findings of the current study, more data is needed from a large and representative sample of children. A second

consequence of using predefined groups is that we fully rely on previous diagnosis, as has been pointed out earlier. Given the overrepresentation of bilingual children in special education in the Netherlands (Smeets et al., 2010), certainty about adequate classification in our sample is not guaranteed. A final limitation of this study is that the bilingual children with LI were more heterogeneous in terms of their home languages, and thus their phonetic inventories, than the bilingual TD children. Whereas we excluded nonwords from the Q-U NWRT that were real words in all home languages of the TD group (i.e. Turkish, Tarifit-Berber and Moroccan Arabic), we could only check this post-hoc for the remaining home languages of the children in the LI group. Even though most items appeared to be true nonsense words in all languages of our sample, a few turned out to be meaningful words (e.g. /lita/ in Kirundi), which could have influenced the results. To check for the effects of home language, we compared NWRT performance between home language groups and found no differences. A study in larger and more homogeneous groups is needed to confirm this.

In addition, future research is needed to compare the Q-U NWRT to other instruments, particularly normed language-specific NWRTs that are currently being used in the clinical practice, and to other alternatives that have been proposed to aid assessment of bilingual children with LI. For example, Engel de Abreu and colleagues (2013) suggest that performance on working memory tasks involving numbers, such as digit span, are not affected by test language or cultural status and could therefore be used in assessment. In this study, 7 year old Portuguese-Luxembourgish language minority children performed equally well on digit span in either language and did not differ significantly from monolingual peers in Luxembourg or Brazil. The authors' explanation for this finding was that children are very familiar with numbers by the age of 7 due to extensive training. It would be relevant to test whether the clinical potential of a digit span task is comparable to the Q-U NWRT in children of that age, but also in younger children whose number knowledge is less well-entrenched.

In summary, the key finding of the present study is that the Dutch version of a newly developed quasi-universal nonword repetition task is a promising diagnostic tool to help identify language impairment in bilingual children with Dutch as a second language. This task is designed to be minimally susceptible to experience in a specific language, in contrast to a more traditional language-specific task to which it was compared. Both instruments discriminated well between monolingual children with and without language impairment, but only the quasi-universal task was clinically accurate in a bilingual group of children as well. The quasi-universal task seems therefore suitable to disentangle language impairment from language disadvantage.

Acknowledgements

This work is part of the research program 'Cognitive development in the context of emerging bilingualism: Cultural minority children in the Netherlands' which is financed by a VIDI-grant awarded to dr. Elma Blom (PI) by the Netherlands Organization for Scientific Research (NWO). We thank the children, parents and schools that participated in the study. We give special thanks to dr. Judith Rispens for placing her instrument at our disposal.

References

Adriaans, F. (2006). PhonotacTools. [Computer program]. Utrecht Institute of Linguistics OTS, Utrecht University, the Netherlands.

Bedore, L., & Peña, E. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism,* 11(1): 1-29.

Bishop, D.V.M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*(4): 391–403.

Blom, E., de Jong, J., Orgassa, A., Baker, A., & Weerman, F. (2013). Verb inflection in monolingual Dutch and sequential bilingual Turkish–Dutch children with and without SLI. *International Journal of Language & Communication Disorders, 48*(4): 1-12.

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*(3): 519-525.

Chiat, S. (2015). Nonword Repetition. In: Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.) *Methods for assessing multilingual children: disentangling bilingualism from Language Impairment*. Bristol: Multilingual Matters.

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

Coady, J.A., Evans, J.L., & Kluender, K.R. (2010). The role of phonotactic frequency in nonword repetition by children with specific language impairments. *International Journal of Language & Communication Disorders, 45*(4): 494-509.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for Specific Language Impairment (SLI). *Journal of Child Psychology and Psychiatry, 42*(6): 741–48.

COST Action IS0804 (2011). Questionnaire for Parents of Bilingual Children (PaBiQ). http://www.bi-sli.org.

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(11-12): 935-944.

Dispaldro, M., Leonard, L.B., & Deevy, P. (2013). Real-word and nonword repetition in Italian-speaking children with Specific Language Impairment: A study of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research*, 56(1): 323-336.

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

Ellis Weismer, S., Tomblin, J., Zhang, X., Buchwalter, P., Chynoweth, J., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language and Hearing Research, 43*(4): 865–878.

Engel de Abreu, P.M.J. (2011). Working memory in multilingual children: Is there a bilingual effect? *Memory, 19*(5): 529-537.

Engel de Abreu, P.M.J., Baldassi, M., Puglisi, M.L., & Befi-Lopes, D.M. (2013). Cross-linguistic and cross-cultural effects on verbal working memory and vocabulary: Testing language minority children with an immigrant background. *Journal of Speech, Language and Hearing Research, 56*(2): 630– 642.

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

Gathercole, S. E. (1995). Is non-word repetition a test of phonological memory or long-term knowledge? It all depends on the non-words. *Memory and Cognition, 23*(1): 83–94.

Gathercole, S.E. (2006). Nonword repetition and word learning: the nature of the relationship. *Applied Psycholinguistics, 27*(4): 513-543.

Gathercole, S. E., & Baddeley, A.D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*(2): 200–213.

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*(3): 336–360.

Graf-Estes, K., Evans, J.L., & Else-Quest, N.M. (2007). Differences in the nonword repetition performance of children with and without Specific Language Impairment: A meta-analysis. *Journal of Speech, Language and Hearing Research, 50*(1): 177-195.

Gray, S. (2003). Diagnostic accuracy and test-retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders*, 36(2):129-151.

Grimm, A., & Schulz, P. (2014). Specific Language Impairment and Early Second Language Acquisition: The Risk of Over- and Underdiagnosis. *Child Indicators Research*, 7(4): 821-841.

Grüter, T. (2005). Comprehension and production of French object clitics by child second language learners and children specific language impairment. *Applied Psycholinguistics, 26*(3): 363-392.

Gutiérrez-Clellen, V.F. (1996). Language diversity: Implications for assessment. In K.N. Cole, P.S. Dale, & D.J. Thal (Eds.), *Assessment of communication and language,* 6: 29–56. Baltimore: Paul H. Brookes.

Gutiérrez-Clellen, V.F., & Simon-Cereijido, G. (2010). Using nonword repetition tasks for the identification of language impairment in Spanish-English-speaking children: Does the language of assessment matter? *Learning Disabilities Research & Practice, 25*(1): 48-58.

Håkansson, G. (2001). Tense morphology and verb-second in Swedish L1 children, L2 children and children with SLI. *Bilingualism: Language and Cognition, 4*(1): 85-99.

Håkansson, G., & U. Nettelbladt (1996). Similarities between SLI and L2 children. Evidence from the acquisition of Swedish word order. In J. Gilbert & C. Johnson (eds.), *Children's Language*, 9: 135-51. Mahwah, NJ: Lawrence Erlbaum & Associates.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, 39(1): 1-27.

Kalnak, N., Peyrard-Janvid, M., Forssberg, H., & Sahlén, B. (2014). Nonword Repetition – A Clinical Marker for Specific Language Impairment in Swedish Associated with Parents' Language-Related Problems, *PLoS ONE*, 9(2): e89544.

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

Kohnert, K., & Bates E. (2002). Balancing bilinguals II: Lexical comprehension and cognitive processing in children learning Spanish and English. *Journal of Speech, Language, and Hearing Research, 45*(2): 347-359.

Kohnert, K., Windsor, J., & Yim, D. (2006). Do language-based processing tasks separate children with primary language impairment from typical bilinguals? *Journal of Learning Disabilities Research & Practice, 21*(1): 19–29.

Kort, W., Schittekatte, M., & Compaan, E.L. (2008). *CELF-4-NL: Clinical Evaluation of Language Fundamentals*. Amsterdam: Pearson Assessment and Information B.V.

Messer, M.H., Leseman, P.P.M., Boom, J., & Mayo, A.Y. (2010). Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology, 105*(4): 306-323.

Munson, B., Kurtz, B.A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without language impairments. *Journal of Speech, Language, and Hearing Research, 48*(5): 1033–1047.

Oller, D.K., & Pearson, B. Z. (2002). Assessing the effects of bilingualism: A background. In: D. K. Oller & R. E. Eilers (Eds.), *Language and Literacy in Bilingual Children* (pp. 3-21). Clevedon, UK: Multilingual Matters.

Orgassa, A., & Weerman, F. (2008). Dutch gender in specific language impairment and second language acquisition. *Second Language Research, 24*(3): 333-364.

Paradis, J. (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with Specific Language Impairment. *Language, Speech and Hearing Services in the Schools, 36*(3): 172-187.

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

Paradis, J. (2011). Individual Differences in Child English Second Language Acquisition: Comparing Child-Internal and Child-External Factors. *Linguistic Approaches to Bilingualism, 1:3*: 213-237.

Paradis, J., Emmerzael, K., & Sorenson Duncan, T. (2010). Assessment of English Language Learners: Using Parent Report on First Language Development. *Journal of Communication Disorders, 43*(6): 474-497.

Plante. E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, 25:15–24.

Restrepo, M.A., & Silverman, S. (2001). Validity of the Spanish Preschool Language Scale-3 for use with bilingual children. *American Journal of Speech Language Pathology, 10(*4): 382–393.

Rispens, J., & Baker, A. (2012). Nonword repetition: the relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language and Hearing Research, 55*(3): 683-94.

Salameh, E.K., Nettelbladt, U., Håkansson, G., & Gullberg, B. (2002). Language impairment in Swedish bilingual children. A comparison between bilingual and monolingual children. *Acta Paediatrica, 9*(2): 229–234.

Scheele, A.F., Leseman, P.P.M., & Mayo, A.Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics, 31*(1): 117-140.

Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL*. Dutch version. Harcourt Assessment B.V., Amsterdam.

Schlichting, J.E.P.T., & Lutje Spelberg, H.C. (2010a), *Schlichting Test voor Taalbegrip: Manual,* Houten: Bohn, Stafleu, Van Loghum.

Schlichting, J.E.P.T., & Lutje Spelberg, H.C. (2010b), *Schlichting Test voor Taalproductie-II: Manual,* Houten: Bohn, Stafleu, Van Loghum.

Sharp, K.M., & Gathercole, V.C.M. (2013). Can a novel word repetition task be a language neutral assessment tool? Evidence from Welsh–English bilingual children. *Child Language Teaching and Therapy, 29*(1): 77-89.

Stichting Siméa (2014). Indicatiecriteria: auditief en/of communicatief beperkte leerlingen. Retrieved from: http://www.simea.nl/dossiers/passend-onderwijs/brochures-po/14-simea-brochure-indicatiecriteria-digitaal.pdf

Smeets, E., Driessen, G., Elfering, S., & Hovius, M. (2010). Allochtone leerlingen en speciale onderwijsvoorzieningen. Nijmegen: ITS.

Summers, C., Bohman, T.M., Gillam, R.B., Peña, E.D., & Bedore, L.M. (2010). Bilingual performance on nonword repetition in Spanish and English. *International Journal of Language & Communication Disorders, 45*(4): 480-493.

Tape, T. (2008). Interpreting diagnostic tests. From http://gim.unmc.edu/dxtests/Default.htm

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

Thordardottir, E., Rothenberg, A., Rivard, M.E., & Naves, R. (2006). Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages? *Journal of Multilingual Communication Disorders, 4*(1): 1–21.

Tomblin, J.B., Records, N.L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*(6): 1284–1294.

Verhoeven, L., & Vermeer, A. (2001). Taaltoets Alle Kinderen (TAK). Arnhem: Cito.

Wechsler, D., & Naglieri, J.A. (2008). Wechsler Nonverbal Scale of Ability. Purchased from: http://www.pearsonclinical.nl/wnv-nl-wechsler-non-verbal

Appendix 1 – Dutch version of the Quasi-Universal NWRT (derived from Chiat, 2015)

Syllable length	Orthography	IPA International Phonetic Alphabet	Prosody ˈ even stress and pitch ˌ falling pitch
2	Sieboe	sibʊ	ˈSieˌboe
	Lietaa	lita	ˈLieˌtaa
	Naakie	naki	ˈNaaˌkie
	Noelie	nʊli	ˈNoeˌlie
3	Baamoedie	bamʊdi	ˈBaaˈmoeˌdie
	Zieboelaa	zibʊla	ˈZieˈboeˌlaa
	Loemiekaa	lʊmika	ˈLoeˈmieˌkaa
	Naaliedoe	nalidʊ	ˈNaaˈlieˌdoe
4	Noekietaalaa	nʊkitala	ˈNoeˈkieˈtaaˌlaa
	Ziebaalietaa	zibalita	ˈZieˈbaaˈlieˌtaa
	Lietiesaakoe	litisaku	ˈLieˈtieˈsaaˌkoe
	Kaazoeloemie	kazʊlʊmi	ˈKaaˈzoeˈloeˌmie
5	Toeliekaasoemoe	tʊlikasʊmʊ	ˈToeˈlieˈkaaˈsoeˌmoe
	Maaloeziekoebaa	maluzikuba	ˈMaaˈloeˈzieˈkoeˌbaa
	Sieboenaakielaa	sibʊnakila	ˈSieˈboeˈnaaˈkieˌlaa
	Liedaabiemoedie	lidabimʊdi	ˈLieˈdaaˈbieˈmoeˌdie

Appendix 2 – the Language-Specific NWRT (Rispens & Baker, 2012)

Syllable length	Phonotactic probability	Orthography	IPA International Phonetic Alphabet
2	high	Raanom	rɑnɔm
		Daanes	dɑnɛs
		Woosel	wosɛl
	low	Luubuf	lybʏf
		Kuimup	kœymʏp
		Joefeum	jʊfø:m
3	high	Kaaroodin	kɑrodɪn
		Voopeeket	vopekɛt
		Deevoenos	devʊnɔs
	low	Veujoetup	vø:jʊtʏp
		Nuigeusup	nœyxø:sʏp
		Muihuuguf	mœyhyxʏf
4	high	Liekoovoepar	likovʊpɑr
		Kooviewaalan	koviwɑlɑn
		Liejootaanig	lijotɑnɪx
	low	Guiweusoegeer	xœywø:sʊxɪr
		Meufuusuinef	mø:fysœynɛf
		Juuvuigoowuf	jyvœyxowʏf
5	high	Wookaaloemoodon	wokɑlʊmodɔn
		Baamerienooves	bɑmɛrinovɛs
		Tieloniedaanag	tilɔnidɑnɑx
	low	Fuugiwuinoefep	fyxɪwœynʊfɛp
		Geumuwoekuubir	xø:mʏwʊkybɪr
		Nuijigeufuusut	nœyjɪxø:fysʏt

Table 1: Characteristics of the participants

		Age in months		**Nonverbal IQ**		**Socio-Economic Status**		**Gender**	**Exposure to Dutch before the age of 4**		**Current exposure to Dutch at home**	
	N	Mean (*SD*)	Range	Mean (*SD*)	Range	Mean (*SD*)	Range	Nr. of boys	Mean (*SD*)	Range	Mean (*SD*)	Range
MOTD[a]	30	71.7 (6.7)	59-84	102.5 (14.4)	81-128	6.6 (2.1)	2-9	20 (67%)	n/a	n/a	n/a	n/a
MOLI[a]	30	71.9 (7.3)	59-87	97.8 (12.8)	72-118	5.7 (2.0)	2-9	22 (73%)	n/a	n/a	n/a	n/a
BITD[a]	30[b]	71.4 (7.5)	54-83	96.7 (14.1)	70-126	4.8 (2.4)	1-9	12 (40%)	42.3 (8.1)	25-57	50.7 (13.9)	23-83
BILI[a]	30	72.6 (8.8)	58-86	96.0 (14.8)	71-124	5.7 (2.3)	2-9	21 (70%)	41.7 (10.8)	20-67	45.2 (17.1)	14-100[c]

[a]MOTD = monolingual typically developing; MOLI = monolingual language impaired;
BITD = bilingual typically developing; BILI = bilingual language impaired
[b]Parents of one bilingual TD child were not willing to give information about their education level.
[c]Due to severe difficulties learning their native tongue, parents of one child with LI decided to consistently speak Dutch to the child when he entered elementary school (explaining the 100% current exposure to Dutch at home). Before this, he was exposed to Dutch 50% of the time.

Table 2: Dutch language skills of the four groups of children.

	PPVT			TAK Word Formation			TAK Sentence Formation		
	N	Mean (*SD*)	Range	***N***	Mean (*SD*)	Range	***N***	Mean (*SD*)	Range
MOTD[a]	29[b]	111.4 (13.1)	78-137	30	16.5 (4.5)	7-24	30	30.4 (6.0)	17-40
MOLI[a]	30	94.8 (13.0)	70-117	30	10.5 (3.3)	5-18	30	10.3 (7.3)	2-34
BITD[a]	29[b]	94.1 (12.2)	59-119	30	11.6 (5.2)	0-20	30	21.5 (7.3)	4-35
BILI[a]	30	78 (10.3)	55-95	29[c]	6.9 (4.7)	0-15	29[d]	9.8 (5.7)	2-20

[a]MOTD = monolingual typically developing; MOLI = monolingual language impaired;
BITD = bilingual typically developing; BILI = bilingual language impaired
[b]For one MOTD and one BITD child, the normed score for the PPVT was not available due to incorrect assessment procedures.
[c]For one BILI child, the TAK Word Formation was terminated due to the child's refusal to cooperate.
[d]For the same reason, one TAK Sentence Formation from a (different) BILI child was terminated.

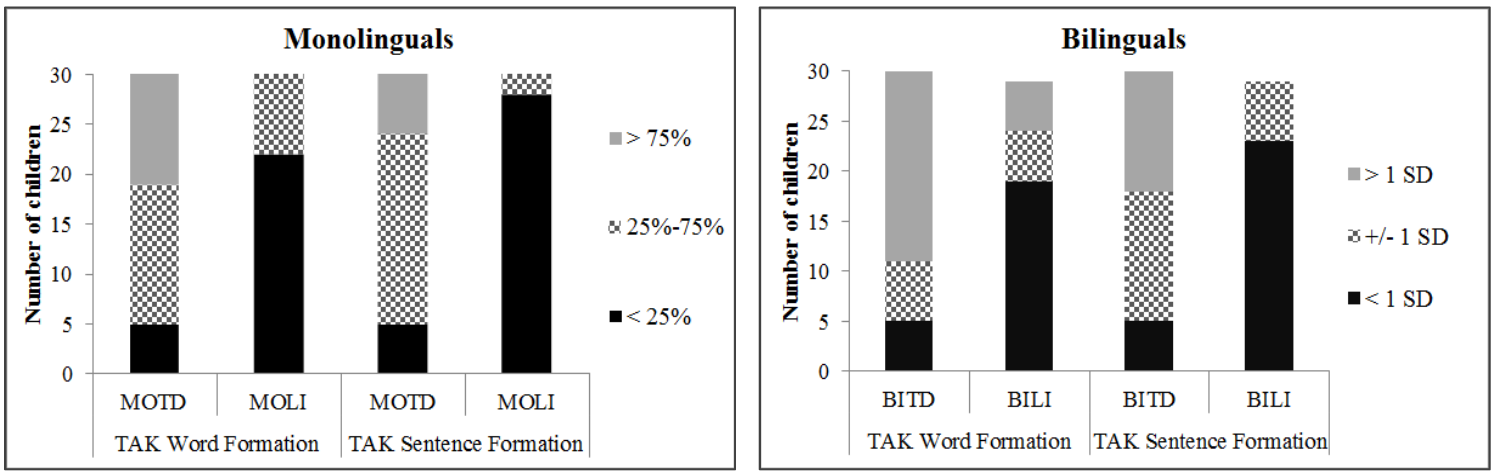

Figure 1: Categorization of children per group according to norms of the TAK Word and Sentence Formation.

Note: TAK norm categories differ for the monolinguals and bilinguals

MOTD = monolingual typically developing; MOLI = monolingual language impaired;
BITD = bilingual typically developing; BILI = bilingual language impaired

Table 3: Percentage of Phonemes Correct on the two versions of the NWRT for the four groups of children.

			Monolingual				**Bilingual**			
			TD[a]		LI[a]		TD[a]		LI[a]	
NWRT	**Syllables**	***N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Quasi-Universal	**All**	120	**88.1**	**6.5**	**67.4**	**12.2**	**86.3**	**6.2**	**69.0**	**12.2**
	2		95.6	6.0	87.7	9.1	97.5	4.3	87.3	9.5
	3		92.3	5.5	73.8	13.1	91.3	7.9	77.5	14.3
	4		92.3	5.3	70.6	17.4	89.1	7.1	73.4	16.6
	5		78.6	14.8	52.4	16.4	76.1	13.0	51.8	17.6
Language-Specific	**All**	120	**82.0**	**6.9**	**58.4**	**11.4**	**73.4**	**7.4**	**60.6**	**13.1**
	2		89.9	4.1	76.7	11.4	85.8	7.3	76.2	11.0
	3		89.2	6.9	68.9	16.0	80.7	10.8	69.2	17.1
	4		82.1	11.0	53.8	14.4	71.6	8.9	56.1	15.4
	5		73.7	10.2	46.6	12.2	64.7	11.9	50.2	14.5

[a]TD = Typically Developing; LI = Language Impaired

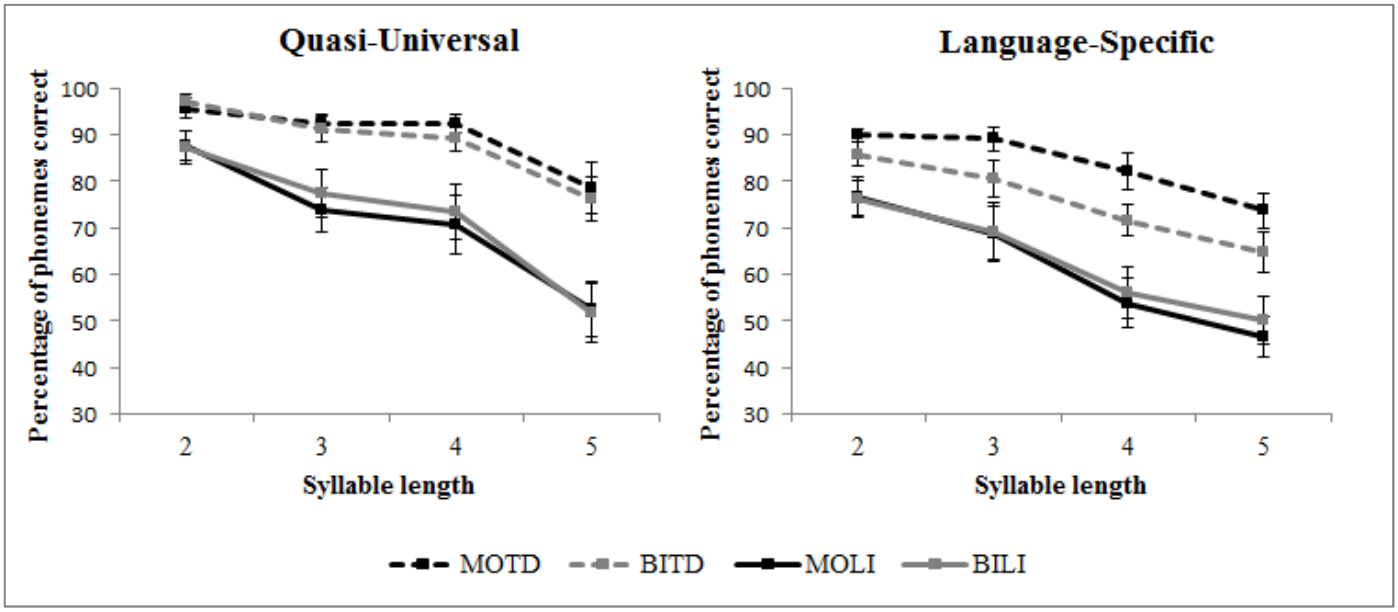

Figure 2: Percentage of Phonemes Correct on the NWRTs per syllable length; error bars represent -/+ 2 standard errors.
MOTD = monolingual typically developing; MOLI = monolingual language impaired;
BITD = bilingual typically developing; BILI = bilingual language impaired

Table 4: Pairwise comparisons: effects of LI and bilingualism on NWRT performance – based on the percentage of phonemes correct

			Quasi-Universal		Language-Specific	
Effect	**Comparisons**	***N***	*p*	*d*	*p*	*d*
Language Impairment	MOTD-MOLI[a]	60	<.001	2.12	<.001	2.50
	BITD-BILI[a]	60	<.001	1.79	<.001	1.20
Bilingualism	MOTD-BITD[a]	60	=1.00	.28	<.001	1.20
	MOLI-BILI[a]	60	=1.00	-.13	=1.00	-.18

[a]MOTD = monolingual typically developing; MOLI = monolingual language impaired; BITD = bilingual typically developing; BILI = bilingual language impaired

Table 5: Percentage of Items Correct on the two versions of the NWRT for the four groups of children.

			Monolingual				**Bilingual**			
			TD[a]		LI[a]		TD[a]		LI[a]	
NWRT	**Syllables**	***N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Quasi-Universal	**All**	120	**59.6**	**15.0**	**25.4**	**13.0**	**55.1**	**13.7**	**28.6**	**17.1**
	2		86.7	17.0	60.8	26.0	88.3	17.0	56.7	29.3
	3		64.2	21.5	21.7	23.4	64.2	26.0	30.3	28.9
	4		59.7	24.1	13.6	20.7	48.1	24.2	25.3	27.1
	5		25.6	26.2	4.2	11.5	18.3	24.2	2.5	7.6
Language-Specific	**All**	120	**36.6**	**11.4**	**11.7**	**9.0**	**23.9**	**9.3**	**14.3**	**11.4**
	2		60.0	12.1	32.8	21.2	53.2	16.4	33.0	22.7
	3		45.6	19.5	10.6	16.7	26.6	18.9	17.3	19.3
	4		29.4	24.2	1.8	5.4	11.4	14.9	5.4	13.2
	5		10.6	15.5	1.1	4.2	3.3	6.8	1.2	4.7

[a]TD = Typically Developing; LI = Language Impaired

Table 6: Optimal cut-off scores, sensitivity (Sn), specificity (Sp), positive likelihood ratios (LR+) and negative likelihood ratios (LR-) – based on the percentage of phonemes correct

	All Children						**Monolinguals**						**Bilinguals**					
	N	Cut-off	Sn	Sp	LR+	LR-	*N*	Cut-off	Sn	Sp	LR+	LR-	*N*	Cut-off	Sn	Sp	LR+	LR-
Quasi-Universal	120	78.1	83%	92%	10.4	.18	60	77.7	83%	90%	8.3	.19	60	78.1	83%	93%	11.9	.14
Language-Specific	120	72.7	87%	77%	3.8	.17	60	72.7	93%	93%	13.3	.08	60	63.8	63%	93%	9	.40