Narrative abilities of monolingual and bilingual children with and without language impairment: Implications for clinical practice

Tessel Boerma, Paul Leseman, Mona Timmermeister, Frank Wijnen, & Elma Blom

Utrecht University

Keywords: language impairment, bilingualism, narrative, assessment, internal states

Corresponding author:

Tessel Boerma, MA

Educational and Learning Sciences

Utrecht University

Heidelberglaan 1

3584 CS Utrecht

Phone: +31 30 253 4963

Email address: T.D.Boerma@uu.nl

Abstract

Background Understanding and expressing a narrative's macrostructure is relatively independent of experience in a specific language. A narrative task is therefore assumed to be a less biased method of language assessment for bilingual children than many other norm-referenced tests and may thus be particularly valuable to identify language impairment (LI) in a bilingual context.

Aims The present study aimed to investigate the effects of LI and bilingualism on macrostructural narrative skills. Moreover, it evaluated the diagnostic validity of a narrative task within a monolingual and bilingual sample.

Methods & Procedures 5 and 6 year old monolingual and bilingual children with and without LI (*N*=33 per group) were tested on production and comprehension measures of narrative macrostructure. A MANCOVA was used to investigate the effects of LI and bilingualism on their narrative abilities. Binary logistic regressions were conducted to evaluate the instrument's diagnostic value.

Outcomes & Results Negative effects of LI were found on all narrative measures, whereas no effects of bilingualism emerged. The narrative task adequately differentiated between both monolingual and bilingual children with and without LI, with story elements related to internal states being more effective than elements related to the basic episode structure.

Conclusions & Implications This study confirms the hypothesis that measures of narrative macrostructure are not biased against children who have less experience with a particular language, like bilinguals. In addition, it indicates that using narratives to assess children's language abilities can support the identification of LI in both a monolingual and bilingual context.

What this paper adds?

What is already known on this subject?

A narrative task is considered a valuable tool for the identification of language impairment in children. However, it is unknown whether this also holds for bilingual children who are often less experienced with the target language than monolingual children.

What this study adds?

This study demonstrates that a narrative task, analyzed on the macrolevel, can successfully differentiate between children with and without LI in monolingual and bilingual learning contexts. Story elements related to internal states had high discriminative value, suggesting that children with LI may have notable difficulty with understanding and expressing the feelings of the protagonists.

Introduction

A narrative task provides rich information about the linguistic development of children in an ecologically valid way and is considered a valuable clinical tool (Botting, 2002). It may be especially valuable for the challenging identification of bilingual children with language impairment (LI). Language delays of bilingual children can arise from impairment but also from insufficient exposure to and, consequently, limited knowledge of the target language (Kohnert, 2010). A narrative task taps into knowledge that goes beyond the specifics of a particular language (Gagarina et al., 2012) and is therefore thought to be a less biased method of language assessment for bilingual children than many other norm-referenced tests (Paradis, Genesee, & Crago, 2010). The present study contributes to a growing body of literature studying narrative abilities of children with LI (e.g. Bishop & Donlan, 2005; Dodwell & Bavin, 2008) and bilingual children (e.g. Pearson, 2002; Uccelli & Paez, 2007). It uses a four-group design with monolingual

and bilingual children with and without LI, thus allowing for systematically investigating the effects of LI and bilingualism on narrative skills. This research is the first to evaluate the use of a narrative task as diagnostic tool with the purpose of supporting the identification of LI in a bilingual context.

Measuring narrative skills

A narrative is an oral sequence of real or imaginary events and is often considered a primary means by which people construct and communicate their actions (Bruner, 1986). Due to the complex nature of storytelling, requiring integration of various cognitive, linguistic and social skills, narratives have been studied as an index of development in several domains (for a review see Liles, 1993) and, therefore, are of interest to both researchers and clinicians. In this section, we discuss frequently used methods to test narrative skills and analyze children's narratives.

Most studies investigate children's productive narrative ability, either through story generation or story retelling based on picture sequences. Story generation is assumed to be more difficult than story retelling as there is no benefit of a prior presented script, thus reflecting the narrator's internalized narrative organization. Retelling gives the advantage of a more structured elicitation, allowing for the experimenter's control over aspects such as length and complexity (Liles, 1993). In the present study, a less often used method was employed that combines story generation and story retelling. A model story which gives contextual support is presented to the child before a comparable but different story is introduced to elicit the child's own story. The advantage of this method is that it provides the child with an example of how a well-structured narrative is formed while at the same time allowing for the elicitation of the child's independent narrative formulation abilities (Gagarina et al., 2012). In addition, the current study uses

comprehension questions to test children's narrative skills, which can be a valuable complement to the assessment of narrative production (Liles, 1993).

Narrative ability can be measured on the level of the macrostructure and microstructure of a story. The *Story Grammar* model of Stein and Glenn (1979), used for macrolevel analyses, describes the universal structural organization of a story that consists of a setting and one or more episodes. Skilled narrators use a basic episode structure consisting of a causally connected goal-attempt-outcome sequence to interpret and represent the story events (Trabasso & Nickels, 1992). Next to these sequences, experienced narrators enrich story episodes by including references to the protagonists' thoughts and feelings, further on called internal states (Bamberg & Damrad-Frye, 1991). Narratives can also be analyzed on a more local (word or sentence) level through measures of microstructure, which refers to the use of lexical and grammatical elements such as reference and conjunction to create local cohesion. Due to cross-linguistic differences in these elements, microstructure may be more language-specific than macrostructure (Berman & Slobin, 1994). Below, we first review studies that have investigated narrative performance of bilingual children on the macro- and microlevel, after which we turn to research that has focused on differences between children with and without LI.

Narrative performance of bilingual children

Several studies have investigated narrative performance of bilingual children in their two languages. Results from this research suggest that cross-linguistic transfer may enable bilingual children to apply story grammar knowledge acquired in their first language (L1) to their second language (L2). That is, cross-language correlations were found for measures of narrative macrostructure, whereas this was typically not the case for microlevel variables (Pearson, 2002; Squires et al., 2014; Uccelli & Paez, 2007). Studies that have compared monolingual and

bilingual children's narrative performance also find differences between macro- and microstructure (Hipfner-Boucher et al., 2014; Pearson, 2002). Pearson (2002) used a story generation task and found that monolingual English children outperformed their bilingual Spanish-English peers on measures of microstructure. In contrast, no effect of bilingualism was found on measures of macrostructure. These results substantiate the idea that microstructural narrative ability is more sensitive to language-specific experience and input, whereas macrostructural narrative skills seem relatively independent of knowledge of a particular language. A recent study by Hipfner-Boucher and colleagues (2014) with a narrative retell task confirms this conclusion. Two groups of children who learned English as their L2 were compared with a group of monolingual English children on narrative macro- and microstructure. The children who primarily heard or spoke a language other than English at home performed significantly weaker on all measures of microstructure, whereas the group of children that was most often exposed to English was indistinguishable from the monolingual group. In terms of macrostructure, no differences between the three groups were observed. Thus, exposure to and use of English had a large impact on bilingual children's microstructural, but not on their macrostructural narrative abilities.

These findings are based on typically developing (TD) children. Cleave, Girolametto, Chen and Johnson (2010) elicited narratives of monolingual and bilingual children with LI. Whereas the bilingual children achieved lower scores than their monolingual peers on standardized tests tapping into morphosyntax, no differences between the groups were found on macro- or microstructural narrative measures. Although the effect of bilingualism on microstructure may thus be different for children with LI compared to children with TD, all studies together suggest that measures of narrative macrostructure do not disadvantage children that have received less input in a particular language, like bilinguals. Both bilingual children with

TD and LI score similarly on narrative macrostructure in comparison with monolingual peers with TD and LI respectively. Using a four-group design, the current research will investigate if these findings can be confirmed.

Narrative performance of children with LI

Narrative tasks are often found to be problematic for children with LI and appear to remain problematic into adolescence (Wetherell, Botting, & Conti-Ramsden, 2007). On the level of microstructure, studies have shown that children with LI perform weak on various aspects, including morphosyntax (Reilly, Losh, Bellugi, & Wulfeck, 2004), lexical richness and use of complex clauses (Gillam & Johnston, 1992). Children with LI are also outperformed by their TD peers when their narratives are analyzed on the macrolevel, producing fewer story structure elements, well-formed episodes and internal state terms (e.g. Bishop & Donlan, 2005; Reilly et al., 2004). This has also been found within a bilingual context (e.g. Paradis, Schneider, & Duncan, 2013; Squires et al., 2014). In addition, narrative comprehension of children with LI seems poor compared to children with TD (e.g. Dodwell & Bavin, 2008), although this has not been studied as extensively as narrative production.

The large majority of studies indicate that story telling is difficult for children with LI, but differences in measures of macrostructure between children with and without LI have not always been found. Studies by Iluz-Cohen and Walters (2012) and Norbury and Bishop (2003) indicated that the macrostructure of stories told by children with LI was not weaker than those told by children with TD. Even though it is possible that some of these divergent findings stem from methodological issues, such as a small sample size or the use of different inclusion criteria for LI (for a discussion see Duinmeijer, de Jong, & Scheper, 2012), they may indicate that macrostructural narrative skills are not necessarily weak for all children with LI. It may also be

that some story elements are more difficult for them than others. For example, Reilly and colleagues (2004) reported that children with LI were initially delayed in their production of the basic structure of the story's episodes (e.g. goals or outcomes), but quickly caught up with their TD peers. However, they consistently lagged behind in the use of elements related to internal states, i.e. those story elements that are not directly evident from the picture sequence and that represent the narrator's perspective on the events, such as frames of mind (Bamberg & Damrad-Frye, 1991). Together with other work that showed the limited use of cognitive state predicates by children with LI (Johnston, Miller, & Taller, 2001), this may suggest that story components related to the internal states of the protagonists are particularly problematic for children with LI, and, therefore, hold promise for differential diagnosis. The current study further explored this.

A narrative task as assessment tool

A narrative task is suggested to be an ecologically valid way to assess communicative competence in typical and atypical populations (Botting, 2002) and is part of clinical assessment test batteries. Widely used normed instruments are, for instance, the Renfrew Bus Story (Renfrew, 1969) and the Edmonton Narrative Norms Instrument (ENNI; Schneider, Dubé, & Hayward, 2005). However, only a few studies have evaluated the diagnostic validity of narrative tasks. Schneider, Hayward and Dubé (2006) investigated the ability of the ENNI to identify LI in children. Overall, nearly 81% of the children were classified correctly by using story grammar analysis. Although the authors conclude that this result holds promise for differential diagnosis, they emphasize the need to use a combination of tests due to the heterogeneity within the LI group. Pankratz, Plante, Vance and Insalaco (2007) reported that 84% of the children with LI were correctly classified by the Bus Story, which is within the acceptable range (Plante & Vance, 1994). However, the fact that nearly 22% of the TD children were misclassified led the authors to

conclude that this narrative task was not suitable for diagnostic purposes. Pankratz et al. (2007) used a combination of one measure of macrostructure and one measure of microstructure. Classification accuracy of the macrostructural measure alone was better than the classification accuracy of the measure of microstructure alone, especially due to relatively weak performance of TD children from a cultural minority on the latter. This suggests that adding other measures tapping into macrostructural narrative skills can possibly improve the instrument's clinical value, in particular where children with differing language experiences are concerned.

Little is known about the diagnostic validity of a narrative task within a sample of bilingual children. The research reviewed above suggests that measures of narrative macrostructure do not disadvantage bilingual children (Hipfner-Boucher et al., 2014; Pearson, 2002), in contrast to other standardized language tests that tap into language-specific knowledge, such as vocabulary or grammar tests (e.g. Restrepo & Silverman, 2001). That is, even bilingual children with TD perform poorly on these tests due to having received less input in the language of testing compared to monolingual children. Most of these standardized measures used for diagnosing LI in monolingual children, therefore, cannot separate a language delay that arises from external factors from a genuine LI, making identification of bilingual children with LI especially challenging (Kohnert, 2010). Assuming that macrostructural narrative skills depend less on knowledge of or experience with one particular language, a narrative task may hold promise for differential diagnosis in a bilingual population. The present study examined this in a sample of monolingual Dutch children and bilingual children who are L2 learners of Dutch.

The present study

This study compared the narrative production and comprehension skills of monolingual and bilingual children with and without LI. The recently developed *Multilingual Assessment*

Instrument for Narratives (MAIN; Gagarina et al., 2012) was used to measure these skills on a macrostructural level. The first aim of the current study was to systematically investigate the effects of LI and bilingualism on macrostructural narrative abilities. Secondly, we aimed to examine the diagnostic validity of a narrative task within a monolingual and bilingual context. Our third aim was to explore if the diagnostic validity of the narrative task improved if children's understanding and production of story structure elements was differentiated in two classes with 1) elements constituting the basic episode structure and 2) elements relating to internal states of the main characters. This was motivated by previous research which showed that certain story elements may be more difficult for children with LI than other story elements (Reilly et al. 2004).

Although some studies suggest that story structure is not necessarily poor for all children with LI (Iluz-Cohen & Walters, 2012; Norbury & Bishop, 2003), most studies report weaker performance of children with LI compared to children with TD on measures of macrostructure. This has been found in monolingual (e.g. Bishop & Donlan, 2005; Reilly et al., 2004) and bilingual children (e.g. Paradis et al., 2013; Squires et al., 2014). Therefore, we predicted that children with TD would outperform children with LI on the macrostructural measures of the MAIN. Furthermore, based on the outcomes of previous research, we hypothesized that there would be no effects of bilingualism on any of the macrostructural measures of the MAIN for children with TD (Hipfner-Boucher et al., 2014; Pearson, 2002) or LI (Cleave et al., 2010). With regard to the clinical utility, we expected that the MAIN could contribute to a reliable differential diagnosis in both a monolingual and bilingual group of children, as a consequence of the hypothesized absent effect of bilingualism. Taking into account story elements related to internal states may improve the clinical value, as these may be particularly difficult for children with LI (Reilly et al., 2004).

Methods

Participants

A total of 132 children participated in this study of whom the majority (*N*=125) was 5 or 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*=33 in each of four groups). The children with LI were recruited through two national organizations in the Netherlands that provide diagnostic, care and educational services for children with language difficulties (Royal Dutch Kentalis and Royal Auris Group). These children were diagnosed with LI by licensed professionals on the basis of a standardized protocol. This meant that these children obtained 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, including speech production, auditory processing, grammatical knowledge and lexical-semantic knowledge. Delays on these subscales were determined with at least two appropriate tests per subscale. The most common test batteries used include the Dutch version of the Clinical Evaluation of Language Fundamentals (CELF-4-NL; Kort, Schittekatte, & Compaan, 2008) and the Schlichting Test for Language Production and Comprehension (Schlichting & Lutje Spelberg, 2010ab). Exclusion criteria for this study were the presence of a hearing impairment, intellectual disability and severe articulatory difficulties as determined by a certified professional. The children with LI attended special education (*N*=64) or regular education with ambulatory care (*N*=2; one bilingual child and one matched monolingual child). All children with TD were recruited via regular elementary schools and did not have reported language problems.

Children were considered monolingual if both parents always spoke Dutch to the child. A child was assigned to the bilingual group if at least one parent was a native speaker of another

language than Dutch and spoke their native tongue with the child for an extensive period of the child's life. All bilingual children were born in the Netherlands and learned Dutch in an environment where this is the majority language. The first languages of the bilingual TD children included Turkish (*N*=14), Tarifit-Berber (*N*=14) and Moroccan Arabic (*N*=5). The first languages of the bilingual children with LI were Turkish (*N*=10), Moroccan Arabic (*N*=7), Egyptian Arabic (*N*=3), Tarifit-Berber (*N*=2), Dari (*N*=2), Chinese (*N*=2), Pashto (*N*=1), Suryoyo (*N*=1), Kirundi (*N*=1), Russian (*N*=1), Portuguese (*N*=1), Danish (*N*=1) and Frisian (*N*=1).

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). SES was indexed by the average educational level of both parents of the child, based on the Questionnaire for Parents of Bilingual Children (PaBiQ; Tuller, 2015). In addition, we matched the two bilingual groups on exposure to Dutch before the age of 4 and current exposure to Dutch at home. 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. The observed wide range of exposure to Dutch is representative for the diverse groups of immigrants in the Netherlands (Centraal Bureau voor de Statistiek, 2015).

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 between any of the four groups ($F(3,128) = .22$, $p = .88$, $\eta_p^2 = .01$). Differences between nonverbal IQ did not reach significance between any of the four groups either ($F(3,128) = 2.4$, $p = .07$, $\eta_p^2 = .05$), although matching was not optimal due to a slightly higher IQ of the

monolingual children with TD compared to the other groups. SES did differ significantly ($H(3) = 9.87$, $p = .02$), reflecting lower SES in the bilingual TD group compared to the monolingual TD group. There were no significant differences between the groups with regard to gender (χ^2 (3, N=132) = 2.7, $p = .44$), nor were there significant differences between the bilingual groups in terms of exposure to Dutch before the age of 4 ($F(1,61) = .20$, $p = .66$, $\eta_p^2 < .001$) or current exposure to Dutch at home ($F(1,60) = 2.8$, $p = .10$, $\eta_p^2 = .04$).

[Insert Table 1 here]

For the purpose of the present study, children were tested on three language measures tapping into 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 and provide background information on the Dutch language abilities of the children. There was a significant group effect on all three measures (PPVT: $F(3,126) = 38.0$, $p < .001$, $\eta_p^2 = .46$; TAK Word Formation: $F(3,127) = 25.5$, $p < .001$, $\eta_p^2 = .38$; TAK Sentence Formation: $F(3,126) = 57.2$, $p < .001$, $\eta_p^2 = .58$). On the PPVT and the TAK Word Formation, the MOTD group outperformed the BITD group ($p < .001$) and the MOLI and BITD groups outperformed the BILI group ($p < .05$). There were no significant differences between the MOLI and BITD group. On the TAK Sentence Formation, all groups differed significantly ($p < .001$), except for the MOLI and BILI groups.

[Insert Table 2 here]

Table 1: Characteristics of the participants

		Age in months		**Nonverbal IQ**		**Socio-Economic Status**		**Gender**	**Exposure to Dutch before 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	33	71.3 (6.7)	59-84	102.9 (13.6)	81-128	6.3 (2.1)	2-9	21 (64%)	n/a	n/a	n/a	n/a
MOLI	33	71.8 (7.3)	59-87	97.0 (14.5)	72-127	5.5 (1.9)	2-9	22 (67%)	n/a	n/a	n/a	n/a
BITD	33[a]	71.7 (7.1)	58-83	94.9 (14.1)	70-126	4.6 (2.2)	1-9	17 (52%)	41.8 (8.7)	25-60	52.0 (12.0)	29-79
BILI	33	72.8 (8.7)	58-86	94.5 (15.1)	71-124	5.7 (2.2)	2-9	23 (70%)	40.7 (11.0)	20-67	45.6 (17.1)	14-100[b]

Note: MOTD = monolingual typically developing; MOLI = monolingual language impaired;
BITD = bilingual typically developing; BILI = bilingual language impaired
[a]The parental questionnaire was missing for four bilingual TD children.
[b]Due to severe difficulties learning the 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	32[a]	109.9 (12.3)	78-137	33	16.4 (4.3)	7-24	33	29.6 (6.9)	12-40
MOLI	33	95.2 (12.6)	72-117	33	9.8 (4.3)	1-18	33	13.1 (7.6)	2-34
BITD	32[a]	94.4 (13.3)	59-122	33	11.9 (5.4)	0-20	32[c]	22.5 (7.3)	11-37
BILI	33	77.0 (11.7)	55-95	32[b]	6.5 (4.7)	0-15	32[c]	9.2 (5.9)	0-20

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

A NARRATIVE AS DIAGNOSTIC TOOL

Instruments

This study used the *Multilingual Assessment Instrument for Narratives* (MAIN; Gagarina et al., 2012), which was developed within the framework of the COST Action IS0804 *Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment*[1] as a tool to assess narrative comprehension and production of bilingual children from 3 to 10 years old. The MAIN offers several modes to elicit a narrative from children (story generation, retelling and telling after listening to a model story), which are preceded and/or followed by comprehension questions that target the story structure and internal states of the characters. To allow for assessment in both languages of a bilingual child, four comparable stories were created that were controlled for cognitive and linguistic complexity, parallelism in macrostructure and microstructure and cultural appropriateness and robustness (Gagarina et al., 2012:1). Each story is illustrated by six full-color picture sequences that represent the three episodes of the story. Each episode introduces one or more characters and allows for the description of the internal states of the character(s) (e.g. hungry or sad). In addition, each episode contains a goal (e.g, Cat wanted to catch the Butterfly), an attempt (e.g, Cat jumped forwards) and an outcome (e.g., Cat fell into the bush or Butterfly escaped) that can be expressed. The present study used the Dutch version of the MAIN.

Procedure

[1] A network that was set up thanks to funding of the European Cooperation in Science and Technology (COST) with the aim of coordinating research on linguistic and cognitive abilities of bilingual children with LI across different migrant communities (www.bi-sli.org).

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, including language, working memory and attention tasks (not reported in the present study), in two separate sessions each lasting approximately one hour. The MAIN was the final task of the second session. All children heard a model story (Dog or Cat) based on a picture sequence, which was followed by ten comprehension questions. Subsequently, the children were presented with another picture sequence (Baby Birds or Baby Goats) and were asked to tell their own story, again followed by ten comprehension questions. To ensure that version of the story could not influence the results due to possible differences in degree of difficulty, this variable (i.e. version) was taken into account when groups of participants were matched. There were no significant differences between the four groups with regard to version of the model story (χ^2 (3, N=132) = 1.4, p = .70) or version of the production story (χ^2 (3, N=132) = .12, p = .99). All narratives were recorded with a highly sensitive microphone (Samson Go Mic) and scored offline by a native speaker of Dutch.

To investigate the effects of LI and bilingualism on macrostructural narrative skills and evaluate the narrative task as clinical tool, we used the outcome measures that are offered by the MAIN. The MAIN comprises three measures of macrostructure for production: 1) Production (max. 17 points) which examined how many story structure elements children incorporated in their story. The story elements included the setting and, for each episode, the internal state as initiating event, goal, attempt, outcome, and internal state as reaction, 2) Number of Internal State Terms which are used by children when telling their own story, and 3) Structural Complexity, which measures how well-formed the child's story episodes are in terms of goal-attempt-outcome

sequences. Due to the high correlation of Structural Complexity with Production (r = .66, p < .001), only the first two outcome measures were used in our analyses. Next to measures assessing narrative production, the MAIN also comprises two measures of macrostructure for comprehension: 1) Comprehension of the model story (max. 10 points) and 2) Comprehension of the production story (max. 10 points). For each measure, children were asked to answer ten questions. In both sets of comprehension questions, three out of ten questions relate to the goals of each episode of the story, six questions refer to the internal states of the characters and one final question evaluates the ability to infer consequences of the events that have taken place (example questions are provided in the next section).

For the exploratory part of this study, three of the four outcome measures (Production, Comprehension of the model story and Comprehension of the production story) were restructured to examine if the diagnostic validity of the narrative task could be improved. The fourth outcome measure, Number of Internal State Terms, remained unaltered as this measure was not related to the story's structure. Restructuring the other three outcome measures was done as follows: First, the two sets of comprehension questions were combined for the sake of measurement quality. Subsequently, the comprehension questions and story structure elements were categorized into two classes. The first class contained the elements or questions related to Internal States (IS; e.g. 'How does the baby goat feel?' or 'The baby goat was scared'), including internal states as initiating event or as reaction. The second class consisted of the elements that constituted the Basic Episode Structure (BES; e.g. 'Why does the fox leap forward?' or 'The fox grabbed the baby goat'), including goal-related questions for Comprehension and goals, attempts and outcomes for Production. This led to four new variables: 1) Production_IS (max. 6 points), 2) Production_BES (max. 9 points), 3) Comprehension_IS (max. 12 points), and 4) Comprehension_BES (max. 6 points). A fifth new variable, 5) Comprehension_INF (max. 2

points), represented the comprehension questions that required inferencing (e.g. 'Who does the mother goat like best, the fox or the bird?'). Setting, a possible sixth variable, was removed from the analyses since this measure proved to be too unreliable, as will be explained below. Together with the Number of Internal State Terms, these five new variables were used to explore whether restructuring the MAIN in this way could improve differentiation between children with and without LI and could identify particular categories with high discriminative value.

A second independent rater scored over 40% of the data, including narratives from all four groups of children. There was complete agreement between the two raters in 96% of the cases for Comprehension of the model story. The intra-class correlation coefficient (ICC; absolute) was .93. For Comprehension of the production story, the overlap between the two ratings was 92% and the ICC was .89. For Production, the two raters agreed in 84% of the cases. The ICC was .82. The ICC for the Number of Internal State Terms was .93. Inspection of all individual questions or elements within the subcomponents of the MAIN revealed that only the percentage of overlap between the two ratings for 'setting' was below 75% (i.e. 66%). This item was therefore not analyzed as individual story element and was removed from the exploratory part of the study. The overlap between the scores of the two raters for the remaining restructured variables was all above 80%.

Data-analysis

All statistical analyses were performed using SPSS 22 (IBM Corp., 2013). For all variables, except for the Number of Internal State Terms, the percentage correct was calculated. Skewness and kurtosis indicated that the variables were normally distributed, except for the Comprehension of the model story in the monolingual TD group which was moderately skewed. Parametric tests are reported, because no differences between parametric and non-parametric tests were found.

Results from the participant match showed that the four groups of children differed in terms of SES and, although not significant, in terms of nonverbal IQ. Therefore, these variables were included as covariates in a MANCOVA to test for the effects of LI and bilingualism, the first aim of the study. Impairment Status (LI and TD) and Language Group (monolingual and bilingual) were included as fixed factors and Comprehension of the model story, Production, Comprehension of the production story and Number of Internal State Terms as dependent variables. Subsequently, post-hoc analyses were conducted in case significant main and/or interaction effects emerged from the MANCOVA.

A second analysis corresponded to the second aim of the study and evaluated the clinical potential of the MAIN by investigating to what extent the instrument predicted the absence or presence of LI. A binary logistic regression was carried out with Comprehension of the model story, Production, Comprehension of the production story and Number of Internal State Terms as predictors of Impairment Status. The backward stepwise method was used to report results from the full model with all predictors and the most optimal model that maximized classification accuracy but minimized the number of predictors. In this study, sensitivity and specificity are based on the proportion of children with LI and TD, respectively, that are correctly identified as such by the MAIN measures. 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 the diagnostic utility of the instrument. For the purpose of the exploratory part of the study, another binary logistic regression was carried out with the restructured variables (Production_IS, Production_BES, Comprehension_IS, Comprehension_BES, Comprehension_INF, and Number of Internal State Terms) as predictors of Impairment Status, to explore whether restructuring the MAIN would improve classification accuracy and identify particular narrative elements that are highly sensitive to LI.

Results

Effects of LI and bilingualism

Table 3 presents the performance per group on the four subcomponents of the MAIN. The MANCOVA revealed a significant main effect of Impairment Status ($F(4, 119) = 19.5$, $p < .001$, $\eta_p^2 = .40$), whereas there was no main effect of Language Group nor an interaction effect of Impairment Status × Language Group. Covariates IQ and SES were also not significant. Subsequent *t*-tests showed effects of Impairment Status on all four subcomponents of the MAIN: Comprehension of the model story ($t(1, 90) = 5.2$, $p < .001$, $d = .90$), Production ($t(1, 130) = 7.5$, $p < .001$, $d = 1.30$), Comprehension of the production story ($t(1, 130) = 5.7$, $p < .001$, $d = .99$), and Number of Internal State Terms ($t(1, 113) = 5.8$, $p < .001$, $d = 1.01$).

[Insert Table 3 here]

Table 3: Narrative performance on the four subcomponents of the MAIN.

		Comprehension Model story %	**Production %**	**Comprehension Production %**	**Internal State Terms *N***
	N	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
MOTD	33	91.14 (9.3)	45.27 (11.9)	82.42 (16.8)	4.03 (2.8)
MOLI	33	81.2 (15.6)	29.77 (7.9)	62.66 (19.1)	2.36 (1.9)
BITD	33	90.24 (9.3)	44.03 (11.3)	77.19 (18.5)	5.30 (2.9)
BILI	33[a]	71.82 (23.2)	29.77 (14.0)	57.71 (24.4)	2.00 (2.0)

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

[a]The comprehension of the production story was missing for one bilingual child with LI.

Pairwise comparisons of the monolingual and bilingual groups with TD and LI show that the effects of Impairment Status were due to significantly weaker performance of the children with LI compared to the children with TD (Table 4). Effects were large on each subcomponent in the bilingual group, with the largest difference between children with TD and LI on the Number

of Internal State Terms that children produced. In the monolingual group of children with TD and LI, large effect sizes were observed on narrative Production and Comprehension of the production story, whereas medium effect sizes on Comprehension of the model story and Number of Internal State Terms were detected.

[Insert Table 4 here]

Table 4: Pairwise comparisons: effects of Impairment Status on MAIN performance.

			Comprehension model story		**Production**		**Comprehension production**		**Internal State Terms**	
Effect	**Comparisons**	***N***	*p*	*d*	*p*	*d*	*p*	*d*	*p*	*d*
Impairment Status	MOTD-MOLI	64	=.003	.77	<.001	1.53	<.001	1.10	=.007	.70
	BITD-BILI	64	<.001	1.05	<.001	1.09	<.001	.90	<.001	1.30

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

Diagnostic validity of the MAIN

The significant negative effects of Impairment Status and the absent effects of Language Group on macrostructural narrative abilities show that the MAIN might be a promising instrument to identify LI, irrespective of whether children are monolingual or bilingual. Therefore, a binary logistic regression was carried out for each language group separately to investigate how well the children with LI are differentiated from their peers with TD by using the four subcomponents of the MAIN as predictors. Results are presented in Table 5.

In the monolingual group, the full model that included all subcomponents of the MAIN as predictors was statistically significant compared to the intercept only model ($\chi^2(4, N=66) = 45.55$, $p < .001$, Nagelkerke $R^2 =. 665$). The classification results showed that 82% of all monolingual children were correctly classified as having TD or LI. A total of 85% of the children with LI (sensitivity) and 79% of the children with TD (specificity) were identified as such by the MAIN. Similar classification results were obtained from a reduced model with three predictors,

indicating that the Number of Internal State Terms did not significantly contribute to accuracy in assigning the status of LI. This optimal model remained significant ($\chi^2(3, N=66) = 44.91$, $p < .001$, Nagelkerke $R^2 = .658$) and did not show poorer fit compared to the full model ($\Delta\chi^2(1, N=66) = -.643$, $p = .423$).

The full model was also statistically significant compared to the intercept only model in the bilingual group of children ($\chi^2(4, N=66) = 36.35$, $p < .001$, Nagelkerke $R^2 = .565$). The classification results revealed that the proportion of correctly classified children (84%) was slightly higher than in the monolingual group. Sensitivity was 79% and specificity was 88%. Without resulting in a decrease in classification accuracy, the model could be further reduced by removing Production as predictor. This optimal model remained significant ($\chi^2(3, N=66) = 35.85$, $p < .001$, Nagelkerke $R^2 = .559$) and did not show poorer fit compared to the full model ($\Delta\chi^2(1, N=66) = -.507$, $p = .477$).

[Insert Table 5 here]

Table 5: Binary logistic regression with the four subcomponents of the MAIN as predictors of Impairment Status.

Group	Model[a]	Variables included	B	S.E.	Wald	*p*	Sn	Sp	LR+	LR-
Monolingual	Full	Production	-.216	.070	9.611	.002	85%	79%	4.0	.19
		Comprehension production	-.067	.031	4.522	.033				
		Comprehension model	-.052	.035	2.253	.133				
		Number of IS Terms	.131	.165	.630	.427				
	Optimal	Production	-.192	.060	10.182	.001	85%	79%	4.0	.19
		Comprehension production	-.066	.030	4.650	.031				
		Comprehension model	-.047	.033	1.974	.160				
Bilingual	Full	Number of IS Terms	-.459	.179	6.574	.010	79%	88%	6.5	.24
		Comprehension production	-.028	.018	2.491	.114				
		Comprehension model	-.039	.026	2.273	.132				
		Production	-.024	.034	.495	.482				
	Optimal	Number of IS Terms	-.516	.164	9.909	.002	79%	88%	6.5	.24
		Comprehension model	-.042	.026	2.756	.097				
		Comprehension production	-.029	.018	2.735	.098				

Note: Sn = Sensitivity; Sp = Specificity; LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio; IS = Internal State
[a]The full model is the model with all predictors and the optimal model is the minimal model that generated the highest sensitivity and specificity;

Diagnostic validity: restructuring the MAIN

As described under 'Procedure', the MAIN was restructured for the purpose of the exploratory part of this study to examine whether the diagnostic validity of the narrative task could be improved. Together with the Number of Internal State Terms, five new variables (see Appendix 1 for their descriptive statistics) were used as predictors for a second binary logistic regression for each language group, of which results are presented in Table 6.

The full models with all six predictors were statistically significant from the intercept only model in both the monolingual ($\chi^2(6, N=66) = 41.63$, $p < .001$, Nagelkerke $R^2 =. 624$) and bilingual group ($\chi^2(6, N=66) = 37.32$, $p < .001$, Nagelkerke $R^2 =.576$), comparable to the previous models presented in Table 5. In the monolingual group, model reduction eventually resulted in the exclusion of the predictors Number of Internal State Terms and Comprehension of the Basic Episode Structure, optimally classifying 85% of the children with LI and 82% of the children with TD correctly. Model fit did not decrease ($\Delta\chi^2(2, N=66) = -1.32$, $p = .517$) and this optimal model remained significant ($\chi^2(4, N=66) = 40.31.$, $p < .001$, Nagelkerke $R^2 =.609$). In the bilingual group, the exclusion of three predictors (Comprehension of the Basic Episode Structure, Comprehension Inference and Production of the Basic Episode Structure) resulted in the optimal model, with both a sensitivity and specificity of 85%, which was equal to the full model. The optimal model was again significant ($\chi^2(3, N=66) = 36.79.$, $p < .001$, Nagelkerke $R^2 =.570$) and did not decrease in fit ($\Delta\chi^2(3, N=66) = -0.532$, $p = .912$).

[Insert Table 6 here]

Table 6: Binary logistic regression with six subcategorized predictors of the MAIN of Impairment Status.

Group	Model[a]	Variables included	B	S.E.	Wald	*p*	Sn	Sp	LR+	LR-
Monolingual	Full	Production_BES	-.107	.041	6.758	.009	76%	82%	4.2	.30
		Comprehension_IS	-.078	.030	6.659	.010				
		Production_ IS	-.046	.027	2.857	.091				
		Comprehension_INF	.021	.013	2.745	.098				
		Comprehension_ BES	-.034	.031	1.178	.278				
		Number of IS Terms	-.025	.161	.024	.876				
	Optimal	Production_ BES	-.113	.040	7.904	.005	85%	82%	4.7	.19
		Comprehension_IS	-.082	.030	7.613	.006				
		Production_ IS	-.050	.023	4.987	.026				
		Comprehension_INF	-.021	.013	2.790	.095				
Bilingual	Full	Number of IS Terms	-.580	.198	8.553	.003	85%	85%	5.6	.18
		Comprehension_IS	-.054	.023	5.595	.018				
		Production_ IS	-.019	.021	.757	.384				
		Production_BES	.012	.025	.237	.626				
		Comprehension_INF	.004	.013	.078	.780				
		Comprehension_ BES	.002	.027	.005	.943				
	Optimal	Number of IS Terms	-.530	.176	9.121	.003	85%	85%	5.6	.18
		Comprehension_IS	-.051	.019	7.381	.007				
		Production_ IS	-.019	.021	.822	.364				

Note: Sn = Sensitivity; Sp = Specificity; LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio; BES = Basic Episode Structure; IS = Internal States; INF = Inference

[a]The full model is the model with all predictors and the optimal model is the minimal model that generated the highest sensitivity and specificity

Discussion

The present study compared the narrative skills of a group of monolingual and bilingual children with and without language impairment (LI). The recently developed *Multilingual Assessment Instrument for Narratives* (MAIN; Gagarina et al., 2012) was used to assess the ability to comprehend and produce a narrative's macrostructure.

The first aim of this research was to investigate the effects of LI and bilingualism on children's macrostructural narrative ability. Results confirmed our predictions and the outcomes of recent studies with two-group comparisons (e.g., Cleave et al., 2010; Hipfner-Boucher et al., 2014; Paradis et al., 2013; Squires et al., 2014). In both the monolingual and bilingual group, negative effects of LI were found on all components of the MAIN. Children with LI had weaker narrative comprehension, produced fewer story structure elements and expressed a smaller number of internal state terms than children with TD. Furthermore, no effects of bilingualism were observed. The bilingual children had similar narrative production and comprehension skills compared to their monolingual peers, indicating that they were not disadvantaged by the measures of narrative macrostructure due having received less input in Dutch. Together with the observation that the bilingual children scored substantially lower than the monolingual children on three measures tapping into vocabulary, morphology and syntax (see Table 2), this finding supports the hypothesis that a narrative is a less biased method of language assessment for bilingual children than many other tests tapping into language-specific knowledge (Paradis et al., 2010).

The second aim of the study was to evaluate the use of the MAIN as a diagnostic tool. The results of this study were comparable to previous work that studied the diagnostic validity of a narrative task (Pankratz et al., 2007; Schneider et al., 2006) in that over 80% of the children were correctly classified as TD or LI by the MAIN. This was true within both a monolingual and bilingual context which suggests that a narrative task, analyzed on the

macrolevel, can support a reliable differential diagnosis irrespective of children's linguistic background. However, even though overall classification accuracy was above 80%, specificity in the monolingual group and sensitivity in the bilingual group only reached 79%, misclassifying 21% of the monolingual children with LI and bilingual children with TD. This is not sufficient for clinical purposes (Plante & Vance, 1994).

The third part of the current study therefore explored whether the sensitivity and specificity of the MAIN could be improved by taking into account specific properties of the narrative elements. This idea was motivated by the possibility that elements related to internal states may be more difficult for children with LI than other elements, such as the expression or comprehension of the basic episode structure, including goals, attempts and outcomes (Reilly et al., 2004). After restructuring the MAIN, clustering elements that were related to internal states and clustering elements that were part of the basic structure of the episode, classification accuracy improved. Sensitivity and specificity reached acceptable levels (> 80%), both within the monolingual and the bilingual group of children. In line with our hypothesis, comprehension and production of internal states showed high discriminative value and contributed to the optimal model that generated the highest sensitivity and specificity, in contrast to comprehension of the goals of the episodes.

These findings suggest that children with LI, either monolingual or bilingual, might have relatively more difficulty with understanding and expressing the feelings and intentions of the protagonists than with the basic episode structure. The limited use of cognitive state predicates by children with LI has already been shown by Johnston and colleagues (2001) and is in line with other work on the relation between language development and theory of mind (ToM) or perspective-taking skills. Children with LI are reported to be delayed in their ability to understand and report their own and others' mental states (e.g., Farrant, Fletcher, & Maybery, 2006). Together these findings indicate that the inclusion of comprehension and

production of internal states within a narrative task could support the identification of LI. Besides these similarities, differences between the language groups also emerged, indicating that a less biased method of language assessment, such as a narrative task, may ideally still consist of different elements for monolingual and bilingual children. Whereas the production of the goals, attempts and outcomes was an important predictor of LI in the monolingual group, the number of internal state terms used by a child contributed best to classification accuracy in the bilingual group. It is conceivable that bilingualism affects these story elements differently. For example, there are indications that bilingualism enhances the abovementioned ToM skills in TD children (e.g. Greenberg, Bellana, & Bialystok, 2013), whereas this may not apply to mere expressions of an episode's basic structure. As such a bilingual advantage was not observed for all our measures relating to internal states, the data from this study only partially support this hypothesis. Future research is needed to examine the effects of bilingualism on these distinct story elements.

The results of the present study indicate that measures of narrative macrostructure can support differentiation between monolingual and bilingual children with and without LI, especially when story elements related to internal states are taken into account. However, due to the heterogeneity in the population of children with LI, we echo the conclusion of Schneider and colleagues (2006) that a narrative task should not be used as a stand-alone tool, but can be informative about LI in combination with other measures. Furthermore, a limitation of this study is the use of predefined groups. We included groups of children that were already diagnosed with LI by stringent criteria (-1.5 *SD*). This has presumably enlarged the difference between the children with TD and LI, positively influencing the diagnostic validity of the narrative task. In addition, even though these stringent criteria were used for the purpose of guaranteeing adequate classification in our sample, the possibility of misdiagnosis cannot be ruled out. The diagnostic validity of the MAIN was determined based

on groups of children that were diagnosed with LI on the basis of standardized measures that may have been biased against children with less experience with Dutch. Future research with a large and representative population sample without predefined groups of children is therefore needed to validate the findings of the current study. Moreover, future studies should examine if the results from the present study can be generalized to specific bilingual groups with varying dominance patterns.

To conclude, this study was the first to evaluate the use of a narrative task as diagnostic tool with the purpose of supporting the identification of LI in a bilingual context. Results from group comparisons showed that measures of narrative macrostructure are difficult for children with LI, whereas bilingualism did not affect performance. After taking into account specific properties of the story elements, the narrative task adequately differentiated between both monolingual and bilingual children with TD and LI, with story elements related to internal states being particularly effective. This study thus showed the promises of a narrative task to alleviate the clinical challenge of assessing bilingual children.

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 the narrative and discourse working group of the COST Action IS0804 for developing the MAIN. ***Declaration of interest***: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

BAMBERG, M. and DAMRAD-FRYE, R., 1991, On the ability to provide evaluative comments: further explorations of children's narrative competencies. *Journal of Child Language*, 18(3), 689-710.

BERMAN, R.A. and SLOBIN, D.I., 1994, *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum.

BISHOP, D.V.M. and DONLAN, C., 2005, The role of syntax in encoding and recall of pictorial narratives: Evidence from specific language impairment. *British Journal of Developmental Psychology*, 23(1), 25-46.

BOTTING, N., 2002, Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy*, 18(1), 1-21.

BRUNER, J., 1986, *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

CENTRAAL BUREAU VOOR DE STATISTIEK (CBS; Statistics Netherlands), 2015, *Bevolking; Generatie, geslacht, leeftijd en herkomstgroepering, 1 januari.* Den Haag. Retrieved from: www.cbs.nl.

CLEAVE, P.L., GIROLAMETTO, L.E., CHEN, X. and JOHNSON, C.J., 2010, Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders*, 43(6), 511-522.

DODWELL, K. and BAVIN, E.L., 2008, Children with specific language impairment: an investigation of their narratives and memory. *International Journal of Language and Communication Disorders*, 43(2), 201-18.

DUINMEIJER, I., DE JONG, J. and SCHEPER, A.R., 2012, Narrative abilities and memory and attention in children with a Specific Language Impairment, *International Journal of Language & Communication Disorders,* 47(5), 542-555.

FARRANT, B., FLETCHER, J. and MAYBERY, M.T., 2006, Specific language impairment, theory of mind, and visual perspective taking: evidence for simulation theory and the developmental role of language. *Child Development*, 77(6), 1842–1853.

GAGARINA, N., KLOP, D., KUNNARI, S., TANTELE, K., VÄLIMAA, T., BALČIŪNIENĖ, I., BOHNACKER, U. and WALTERS, J., 2012, MAIN – Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics,* 56, 1-155.

GILLAM, R.B. and JOHNSTON, J.R., 1992, Spoken and written language relationships in language/learning-impaired and normally achieving school-age children. *Journal of Speech and Hearing Research, 35*(6), 1303-1315.

GREENBERG, A., BELLANA, B. and BIALYSTOK, E., 2013, Perspective-taking ability in bilingual children: extending advantages in executive control to spatial reasoning. *Cognitive Development*, 28(1), 41-50.

HIPFNER-BOUCHER, K., MILBURN, T., WEITZMAN, E., GREENBERG, J., PELLETIER, J. and GIROLAMETTO, L., 2014, Narrative abilities in subgroups of English language learners and monolingual peers. *International Journal of Bilingualism*, 1-16.

ILUZ-COHEN, P. and WALTERS, J., 2012, Telling stories in two languages: Narratives of preschool children with typical and impaired language. *Bilingualism: Language and Cognition*, 15(1), 58-74.

JOHNSTON, J.R., MILLER, J.F. and TALLAL, P. J., 2001, Use of cognitive state predicates by language-impaired children. *International Journal of Language & Communication Disorders*, 36(3), 349–370.

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

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

LILES, B., 1993, Narrative discourse in children with language disorders and children with normal language: A critical review of the literature. *Journal of Speech and Hearing Research*, 36(5), 868–882.

NORBURY, C.F. and BISHOP, D.V.M., 2003, Narrative skills of children with communication impairments. *International Journal of Language & Communication Disorders*, 38(3), 287-313.

PANKRATZ, M.E., PLANTE, E., VANCE, R. and INSALACO, D.M., 2007, The diagnostic and predictive validity of The Renfrew Bus Story. *Language, Speech, and Hearing Services in Schools*, 38(4), 390–399.

PARADIS, J., GENESEE, F. and CRAGO, M., 2010, *Dual language development and disorders: A handbook on bilingualism and second language learning* (2nd Edition). Baltimore, MD: Brookes.

PARADIS, J., SCHNEIDER, P. and SORENSON DUNCAN, T., 2013, Discriminating children with language impairment among English language learners from diverse first language backgrounds. *Journal of Speech, Language and Hearing Research*. 56(3), 971-981.

PEARSON, B.Z., 2002, Narrative competence among monolingual and bilingual school children in Miami. In: Oller, D.K., Eilers, R.E. (Eds.). *Language and literacy in bilingual children* (Clevedon, UK: Multilingual Matters), pp. 135-174.

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

REILLY, J., LOSH, M., BELLUGI, U. and WULFECK, B., 2004, "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language*, 88(2), 229-247.

RENFREW, C., 1969, *The Bus Story: A test of continuous speech*. Old Headington: Oxford.

RESTREPO, M.A. and 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.

SCHNEIDER, P., DUBÉ, R.V. and HAYWARD, D., 2005, The Edmonton Narrative Norms Instrument. Retrieved from University of Alberta Faculty of Rehabilitation Medicine website: http://www.rehabmed.ualberta.ca/spa/enni.

SCHNEIDER, P., HAYWARD, D. and DUBÉ, R.V., 2006, Storytelling from pictures using the Edmonton Narrative Norms Instrument. *Journal of Speech-Language Pathology and Audiology*, 30(4), 224–238.

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.

STEIN, N.L. and GLENN, C., 1979, An analysis of story comprehension in elementary school children. In R.O. Freedle (Ed.), *New directions in discourse processing* (Norwood, NJ: Ablex), Vol. 2, pp. 53–120.

SQUIRES, K.E., LUGO-NERIS, M.J., PEÑA, E.D., BEDORE, L.M., BOHMAN, T.M. and GILLAM, R.B., 2014, Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language and Communication Disorders*, 49(1), 60-74.

TRABASSO, T. and NICKELS, M., 1992, The development of goal plans of action in the narration of a picture story. *Discourse Processes,* 15(3), 249-275.

TULLER, L., 2015, Clinical use of parental questionnaires in multilingual contexts. In: Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.), *Methods for assessing multilingual children: disentangling bilingualism from Language Impairment*. Bristol: Multilingual Matters.

UCCELLI, P. and PÁEZ, M., 2007, Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among English and Spanish skills. *Language, Speech, and Hearing Services in Schools,* 38(3), 225-236.

VERHOEVEN, L. and VERMEER, A., 2001, Taaltoets Alle Kinderen (TAK). Arnhem: Cito.

WECHSLER, D. and NAGLIERI, J.A., 2008, Wechsler Nonverbal Scale of Ability. Purchased from: http://www.pearsonclinical.nl/wnv-nl-wechsler-non-verbal

WETHERELL, D., BOTTING, N. and CONTI-RAMSDEN, G., 2007, Narrative in adolescent specific language impairment (SLI): a comparison with peers across two different narrative genres. *International Journal of Language and Communication Disorders*, 42(5), 583–605.

Appendix 1

Descriptive statistics for the restructured predictors of the MAIN

		Production_ IS %	**Production_ BES %**	**Comprehension_ IS %**	**Comprehension_ BES %**	**Comprehension_ INF %**
	N	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
MOTD	33	30.81 (20.5)	58.92 (13.5)	86.26 (14.1)	95.35 (8.8)	63.64 (28.7)
MOLI	33	13.13 (15.5)	44.44 (12.1)	68.94 (17.6)	85.15 (16.3)	57.58 (37.8)
BITD	33	29.80 (18.0)	56.23 (15.9)	80.05 (16.8)	91.16 (14.1)	74.24 (28.3)
BILI	33[a]	13.64 (16.4)	44.78 (20.9)	58.08 (24.2)	78.69 (16.8)	57.58 (37.8)

Note: MOTD = monolingual typically developing; MOLI = monolingual language impaired; BITD = bilingual typically developing; BILI = bilingual language impaired; IS = Internal States; BES = Basic Episode Structure; INF = Inference

[a]The comprehension of the production story was missing for one bilingual child with LI. Therefore, Comprehension_BES and Comprehension_IS are based on the comprehension of the model story for this child.