IDENTIFYING DISORDER WITHIN DIVERSITY FOR CHINESE-ENGLISH BILINGUALS: EXPLORING THE CLINICAL UTILITY OF NARRATIVE MACROSTRUCTURE

by

Lina Bauer

B.Sc., The University of British Columbia, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Audiology and Speech Sciences)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Identifying disorder within diversity for Chinese-English bilinguals: exploring the clinical utility of narrative macrostructure

submitted by Lina Bauer in partial fulfilment of the requirements for

the degree of Master of Science

in Audiology and Speech Sciences

Examining Committee:

Paola Colozzo, Associate Professor, School of Audiology and Speech Sciences, UBC
Supervisor

Stefka Marinova-Todd, Associate Professor, School of Audiology and Speech Sciences, UBC
Supervisory Committee Member

Ruth Hanson, Clinical Instructor, School of Audiology and Speech Sciences, UBC
Supervisory Committee Member

Abstract

English Language Learners (ELLs) are sequential bilingual children who speak a minority language at home and learn English through school programs. ELLs are more likely than monolingual children to be under- or over- diagnosed with Developmental Language Disorder, a language disorder of unknown biomedical origin. This study aims to describe the range of abilities of ELLs on language-based measures in English thereby providing benchmarks to identify children who are not performing as expected. Narrative macrostructure and nonword repetition were of interest as performance on these measures seems to be less specific to a given language.

Chinese-English ELLs completed a story retell, analyzed for macrostructure and microstructure, and a nonword repetition task in Grade 2 (*N* = 75) and Grade 3 (*N* = 68). Parent report and English Language Learning (ELL) teacher evaluations provided information about the child's first language development history and current second language abilities. Speech-Language Pathologists (SLPs) reported which children had been referred to them for suspected language difficulties.

In line with developmental expectations for monolinguals, the ELLs used a variety of story grammar elements and showed emergence of mental states in their stories. Results from the language measures were generally consistent with ELL teacher evaluations and SLP report but also identified other children who may require additional supports. Macrostructure seemed to be a stronger indicator of low language proficiency in Grade 2 than Grade 3. Not all

microstructure measures were equally informative in this Chinese-English sample. Length of exposure to English was not related to macrostructure scores.

The findings further our understanding of typical language characteristics of Chinese-English sequential bilinguals. In line with prior research, results suggest that narrative macrostructure reflects language skills that are less language specific and, when supplemented with other data, holds promise for identifying bilingual children with a language disorder. Moreover, macrostructure could contribute to identifying children who would otherwise be missed. Evaluating Chinese-English bilingual children using English monolingual norms for verbal fluency and grammatical accuracy could misidentify these children with a language disorder. ELL teacher evaluations may be useful to screen for children who require further evaluation.

Lay Summary

For many school-age children, the language they first learn at home is not the language they use at school. With enough time, most children will have sufficient knowledge of the second language to succeed in school, but a few may have a language disorder. In this study, we examined the language abilities of Chinese-English bilinguals, through assessments and reports, in Grade 2 and Grade 3. In line with prior research, our results suggest that the ability to tell stories with a well-developed plot draws upon skills that are not specific to one language. There was variability in storytelling abilities with most, but not all, students producing detailed well-structured stories. In combination with other measures, a child's storytelling ability may be useful for distinguishing between children who need more time to learn the second language from those who have a language disorder.

Preface

This thesis presents original intellectual work completed by the author, Lina Bauer, under the supervision of Paola Colozzo with guidance provided throughout by the supervisory committee members Ruth Hanson and Stefka Marinova-Todd.

The current project used archive data from a longitudinal study originally designed by Ruth Hanson, Christy Whitely, and Paola Colozzo and approved by the UBC Behavioural Research Ethics Board (certificates H15-00498 and H16-00465). Speech-Language Pathologists at the Vancouver School board participated in data collection. Ruth Hanson transcribed the nonwords and oversaw the scoring for the nonword repetition tasks. The author contributed to all other data reduction and coding, and provided direction to research assistants to maintain high levels of reliability. Research assistants participated in the transcription of the narratives and macrostructure scoring. The author was involved from the beginning in developing both macrostructure rubrics, first working in collaboration with Paola Colozzo, Ruth Hanson, and Cristy Whitely, then taking a lead role in development, trialing, and revisions. A preliminary version of one rubric was piloted by a previous student, Ivana Prpic, who also participated in its early development.

Table of Contents

Abstract ..... **iii**

Lay Summary ..... **v**

Preface ..... **vi**

Table of Contents ..... **vii**

List of Tables ..... **xi**

List of Figures ..... **xii**

List of Abbreviations ..... **xiii**

Acknowledgements ..... **xiv**

Dedication ..... **xvi**

Chapter 1: Introduction **1**

1.1 English as Second Language Learners 2

1.2 Developmental Language Disorder 3

1.3 Language Development in English Language Learners: Difference and Disorder 4

1.4 Diagnosing Developmental Language Disorder 8

1.4.1 Considering L1 and L2 8

1.4.2 Reference Groups 11

1.4.3 Language Assessments That Show Promise for Assessing Bilingual Children in the LoI 12

1.5 The Present Study 27

1.6 Research Questions and Analyses 28

Chapter 2: Method **32**

2.1 Design 32

2.2 Participants 32
2.3 Materials and Procedures 37
2.3.1 ELL Teacher Evaluations and Information Provided by SLPs 39
2.3.2 Nonword Repetition 41
2.3.3 Narrative Retell 42
2.4 Analyses 51
Chapter 3: Results 52
3.1 Overview 52
3.2 Grade 2 53
3.2.1 Macrostructure 53
3.2.2 Microstructure 61
3.2.3 Nonword Repetition 62
3.2.4 Teacher Evaluations: ELL Matrices 62
3.2.5 Low Language Proficiency 63
3.3 Grade 3 66
3.3.1 Macrostructure 66
3.3.2 Microstructure 74
3.3.3 Nonword Repetition 76
3.3.4 Teacher Evaluations: ELL Matrices 76
3.3.5 Low Language Proficiency 76
3.4 Additional Analyses 80
3.4.1 Years of Exposure and Total Macrostructure Score 80
3.4.2 Monolingual Comparisons 80
Chapter 4: Discussion 85

4.1 Rubric Difficulty 85
4.2 Typical Narratives 87
4.3 Low-Scoring Stories 89
4.4 Identifying Developmental Language Disorder 90
4.4.1 Are There Some Children Who We Should Be Concerned About Based on the Language Measures Who Are Not Being Picked Up by the Reports? 91
4.4.2 Are the Language Measures Missing Some Children Who Have a Diagnosis? 92
4.5 Relationship Between ELL Matrices and Language Measures 93
4.6 Relationship between SLP Report and Language Measures 94
4.7 Considering Length of Exposure to English 96
4.8 Monolingual Comparisons 97
4.9 The Clinical Utility of Macrostructure 98
4.10 Clinical Implications 100
4.11 Limitations 101
4.12 Future Directions 102
Chapter 5: Conclusion 104
References 106
Appendices 120
Appendix A: Parent Questionnaire 120
Appendix B: Elicitation Protocol for *Pookins Gets Her Way* and *A Porcupine Named Fluffy* . 121
Appendix C: Macrostructure Analysis 122
C.1 Pookins Gets Her Way 122
C.2 A Porcupine Named Fluffy 125
Appendix D: Macrostructure Rubrics 128

D.1 Pookins Gets Her Way 128

D.2 A Porcupine Named Fluffy 135

Appendix E: Identifying Complete Episodes and Action-Consequence Pairs 142

E.1 Pookins Gets Her Way 142

E.2: A Porcupine Named Fluffy 144

List of Tables

Table 1 Demographic characteristics of the participants, by grade 34

Table 2 Language exposure, language history and maternal education at baseline (*N* = 75).... 35

Table 3 Data included in the current study..... 38

Table 4 Overview of children's performance on language measures and ELL Matrices, by grade 53

Table 5 Story grammar elements, by group 57

Table 6 Recombined story grammar elements, by group 58

Table 7 Expected number of story grammar elements in a typical story, by type..... 60

Table 8 Summary of microstructure variables for PGHW (*N* = 75)..... 61

Table 9 Story grammar elements, by group 70

Table 10 Recombined story grammar elements, by group 71

Table 11 Expected number of story grammar elements in a typical story, by type..... 74

Table 12 Summary of microstructure variables for APNF (*N* = 68) 75

Table 13 Performance of monolingual (age and grade matched) and bilingual children for microstructure variables obtained from PGHW story retells..... 82

Table 14 Performance of monolingual (age and grade matched) and bilingual children for microstructure variables obtained from APNF story retells..... 83

List of Figures

Figure 1 Proportion of children who included each item in their story retell 56

Figure 2 Participants who scored more than 1 SD from the mean for macrostructure, microstructure (two or more indicators), or NWR 63

Figure 3 Proportion of children who included each item in their story retell 67

Figure 4 Participants who scored more than 1 SD from the mean for macrostructure, microstructure (two or more indicators), or NWR 77

List of Abbreviations

%Errors: Percent utterances with errors

%MzWds: Percent total maze words

%PauseTime: Percent total pause time

APNF: A Porcupine Named Fluffy

DLD: Developmental Language Disorder

ELL/ELLs: English Language Learning/English Language Learners

L1: First language

L2: Second language

LoI: Language of instruction

MA-NDW: Moving-average number of different words

MLUw: Mean length of utterances in words

NWR: Nonword repetition

PGHW: Pookins Gets Her Way

SLP: Speech-Language Pathologist

TotNumUt: Total number of utterances

VPU: Verbs per utterance

WPM: Words per minute

Acknowledgements

This thesis would not have been possible without the work, expertise, and support of many people. In addition, I am grateful for the financial support I received through a SSHRC Canada Graduate Scholarship Master's Award.

To the Speech-Language Pathologists and research assistants who were involved in the initial stages of data collection and transcription, thank you! I did not get to meet many of you, but I appreciate the foundation you created.

I'd like to thank the families and teachers that took part in the study. In particular, thank you to the children whose story retells form the basis of this thesis. As someone who also learned English as a second language in school, I appreciate your involvement in this project despite how daunting it can be to be assessed in the second language. Listening to the story retells made me in awe of your capabilities, and I hope this thesis conveys your abilities sufficiently.

To the two Speech-Language pathologists, Dr. Ruth Hanson and Cristy Whitely whose ideas started this project, your passion to support bilingual children in the community is inspiring, and I thank you for the time you've spent on research tasks and collaboration. Paula, Hannah, and Danielle – thank you for your contribution by transcribing and coding stories. I also want to thank my committee members, Ruth and Dr. Stefka Marinova-Todd who provided me with feedback.

There are not enough sesquipedalian words to describe the appreciation I have for my supervisor, Dr. Paola Colozzo. Thank you for the encouragement, support, and mentorship.

Additionally, thank you for the time you dedicated to this project and the frequent words of motivation. Throughout this process, I reflected on how lucky I am to be one of your students. I truly could not have done this without you.

I also thank my friends and family for the support they've provided. All your encouragement is appreciated. A big thanks to my fellow thesis friends, Larissa Melville and Linda Wu. I didn't anticipate our friendship would blossom this way, but I'm so glad it did. Thank you for always being the biggest cheerleaders and making tea days extra memorable. A special thank you to my mom, who has helped me reach my goals in any way she can.

Dedication

To all those that provided me with kind words, laughs, and motivation. Especially to my mom, for her constant and unwavering support.

Chapter 1: Introduction

Identifying bilingual children who present with language-learning difficulties is of critical importance to provide them with timely specialized support in the primary grades—as would be the case for monolingual students. English Language Learners (ELLs) speak a home language that is not used by the larger community and are schooled in English. Given the lack of appropriate assessment tools and the practical challenges of implementing bilingual assessments, ELLs are at risk of over- or under-diagnosis of language disorder (Grimm & Schulz, 2014). Emerging evidence suggests, however, that select English-language measures, parent report on early milestones in the first language (L1), and teacher reports on current proficiency in English can, in combination, serve to identify ELLs with language disorders in the early school years. (Boerma & Blom, 2017; Paradis et al., 2013; Pratt et al., 2022; Schwob et al., 2021).

This study adds to the existing body of work aimed at providing guidance to clinicians when assessing ELLs to determine whether they may present with a language disorder. Participants were Chinese-English bilinguals who attend public schools in Vancouver, BC. The intent is to identify language measures that could help Speech-Language Pathologists (SLPs) ascertain which children present with low language proficiency in English by comparing them to each other—their true peers. In line with prior research, the current study focuses primarily on data obtained from narrative retells and explores the clinical utility of narrative macrostructure.

The introduction will review the following topics: language development in ELLs; Development Language Disorder in monolingual and bilingual children; evidence-based considerations when assessing ELLs; promising language tasks and variables, with a focus on

narrative retell and macrostructure analysis. The chapter will conclude with research aims, questions, and hypotheses.

1.1 English as Second Language Learners

A child's first language is not always the same as the one they use to learn in school. English Language Learners (ELLs) are sequential bilingual children who learn English as an additional language through preschool or school programs and usually speak a minority language at home (Paradis, 2016). These bilingual students must learn the language of instruction (LoI) while simultaneously learning the academic content in that language (Cummins, 2000).

ELLs make up a large proportion of children in many school districts in BC, representing 11.5% of all students (Government of British Columbia, 2023). Statistics Canada report that 12.7% of the Canadian population speaks a non-official language at home with that number increasing in major metropolitan centres (Statistics Canada, 2022). Furthermore, more than 40% of students in the Vancouver school district speak a language other than or in addition to English at home (Vancouver School Board, 2023). These students are diverse along many dimensions including, languages spoken, when and for what reasons their family came to Canada, levels of parental education and other socio-demographic variables. Children from a few home language groups nonetheless make up a large proportion the ELL population; internal data from 2016-2017 report that 44% of multilingual children spoke a Chinese language at home (Vancouver Board of Education, 2016).

There is a gap between the time it takes for a child to acquire conversational proficiency versus academic proficiency in their second language (L2) (Cummins, 2000). It takes about two years of exposure to the L2 to reach conversational proficiency, while five to seven years are needed to reach academic proficiency in English-speaking classrooms (Cummins, 2000). Educators unaware of this difference may assume that conversational proficiency in English is an indicator of overall language ability, and hence incorrectly conclude that a bilingual child has a language-learning disability, such as when a bilingual child with native-like phonology is unable to achieve satisfactory grades (Cummins, 2000). Relatedly, ELLs who are in the process of learning English during the early elementary school years are at risk of being under- or over-diagnosed with DLD. This is because their L2 skills are, for a time, likely below average compared to their monolingual peers and thus deviate from the monolingual norms that are used to assess a child's language development (Paradis, 2016).

1.2 Developmental Language Disorder

A child's ability to understand and use spoken language is imperative for school success which in turn suggests early detection to be critical (Ziegenfusz et al., 2022). Parents or teachers who suspect communication difficulties may voice their concerns to a SLP who would then evaluate the child's abilities.

Following these assessments, some children will be diagnosed with Developmental Language Disorder (DLD, previously described as Specific Language Impairment), a childhood

language disorder that is not associated with a known biomedical cause (Bishop et al., 2017).[1] Children with DLD experience persistent language difficulties which make them susceptible to experience academic and social challenges (Kohnert, 2010). This population forms a heterogenous group. Children with DLD may have difficulties in one or more domains of language: morphology, syntax, vocabulary and semantics, and pragmatics. DLD can also be accompanied by difficulties in phonology (Bishop et al., 2017). While there is individual variability, there is evidence that children with DLD perform lower on broad academic, literacy (e.g., reading, spelling, writing, and oral narratives), and numeracy achievement compared to their typically developing (TD) peers (Ziegenfusz et al., 2022). Bilingual children can present with DLD just as some monolingual children do (Bishop, 2020; Bishop et al., 2017). A child with DLD will have impairments in all their languages (Genesee et al., 2021; Kohnert, 2010; Peña et al., 2023). Crucially, bilingualism is not the cause of DLD.

1.3 Language Development in English Language Learners: Difference and Disorder

The majority of ELLs have typical language-learning ability and, with enough time and sufficient opportunity, can develop a level of proficiency that will allow them to communicate successfully in the second language (L2) (Kohnert, 2010) and meet their needs across life contexts—including expectations at school.

[1] The term *Developmental Language Disorder* will be used in this paper regardless of whether research papers cited have referred to language disorders as *Specific Language Impairment* or other labels.

Bilingual children vary in regard to the age at onset of L2 acquisition and length of exposure to and frequency of input for each of their languages—factors that combine to influence the course of a given child's L2 development (Grimm & Schulz, 2014). The child's L1 also impacts L2 acquisition. For example, in a study of bilingual children with various L1s, children whose L1 is marked for tense and agreement (e.g., Hindi, Punjabi, Urdu, Spanish, and Arabic) were more accurate on an English verb morphology assessment compared to children whose L1 is not (e.g., Mandarin and Cantonese) (Paradis, 2011). Mandarin and English are typologically distinct languages with morphosyntax being one area of stark difference, Mandarin having sparse morphosyntactic features. (Sheng, Yang, et al., 2023). Therefore, Chinese-English bilinguals may have different trajectories of English morphological development compared to other bilingual children whose L1 marks tense (Sheng, Yang, et al., 2023).

Patterns of slow morphological acquisition in English for TD children whose L1 is Chinese have been reported in the literature (Paradis et al., 2013). In a study by Paradis et al. (2016), 18 Chinese-English sequential bilinguals (mean age = 8;5, mean length of exposure = 4;3) were tested for English verb morphology using the Test of Early Grammatical Impairment (TEGI) over three time points. Not only was there variation between the levels of accuracy, but also evidence for a plateau in the mastery of certain markers. By the time the children had 6;4 years of L2 exposure, only 39% had acquired native-like levels for all tested morphemes (Paradis, 2005). By the end of the study, irregular past tense, DO auxiliary, and Dropped Marker (i.e., grammatical judgement of whether a tense marker can be omitted) were less likely to be

mastered indicating that subject-verb agreement could be of exceptional difficulty for Chinese-English bilinguals (Paradis et al., 2016).

Given this high level of heterogeneity, as a group, children who are in the process of learning English as a L2 during the early elementary school years are at risk of being over- and under-diagnosed with DLD. Over-diagnosis can occur because levels of proficiency in the L2 are, for a time, likely below that of their monolingual peers and thus deviate from the monolingual norms if these are used as a basis of comparison when assessing a child's language development (Paradis, 2016). To further complicate diagnosis of DLD in bilingual children, the linguistic characteristics and error patterns of a child learning English as a L2 may overlap with those of monolingual children with DLD (Paradis, 2016). In the early stages of English L2 development, bilingual children frequently omit grammatical morphemes which is a common characteristic of DLD in monolingual English-speaking children (Paradis, 2005). Therefore, differences or delays in L2 development could be incorrectly interpreted as evidence of a language disorder if the child's bilingual status is not considered. However, if we consider the child's overall linguistic background, differences or delays in L2 development may be a result of not-yet-sufficient exposure (Paradis, 2016).

Of course, clinicians want to avoid incorrectly diagnosing children with DLD (i.e., over-diagnoses) as such errors could have negative psychosocial consequences for the child and their family. Moreover, intervention time and resources are limited and most often insufficient to meet the needs of all children with speech-language needs, so that there is a concern of providing specialized services to children who do not require them. With a desire to be cautious, a clinician may take a wait-and-see approach due to assumptions that the child has

not had enough time to learn the L2 thereby postponing in-depth assessments or intervention until the child has had more time to become immersed in the LoI (Genesee et al., 2021). Based on anecdotal evidence, school policies or clinician practices may also contribute to this. Delays of 2 or more years have been reported by Genesee et al. (2021). While a wait-and-see approach is often appropriate for many children who are immersed in the L2, it can be a barrier for a child with a language impairment to receive timely intervention, leading to wait-and-fail (i.e., under-diagnoses) (see Genesee et al., 2021 for discussion). Other contributors to under-identification include the lack of bilingual clinicians and limited availability of appropriate assessment tools (Peña et al., 2023). Therefore, the search for a feasible assessment approach for ELLs with acceptable levels of sensitivity and specificity continues.

Timely referrals and correct identification are high stakes issues for the education system, as prevalence rates of DLD are estimated at approximately 7-8% for children entering school (Whiteside & Norbury, 2017). This translates to approximately two children in every 30-seat classroom. Like other neurodevelopmental disorders (e.g., Autism Spectrum Disorder, Attention-Deficit/Hyperactivity Disorder), DLD can manifest in any child whether they are growing up in monolingual or multilingual contexts (American Psychiatric Association, 2013). ELLs must be appropriately assessed in order to determine with a strong probability whether a language disorder is present, or instead, whether the child would likely benefit from language enrichment opportunities (i.e., supports specifically aimed at ELLs). Manifestations will change as a child develops and grows older, but language difficulties will persist into later school years, adolescence, and even into adulthood (Botting, 2020).

1.4 Diagnosing Developmental Language Disorder

A diagnosis of DLD usually depends on direct assessment of the child using standardized tests, language samples, as well as gathering background information from parents (Grimm & Schulz, 2014). Test scores alone do not warrant a diagnosis of DLD as the results of standardized assessments do not necessarily reflect the functional impacts of communication difficulties. A diagnosis of DLD is only warranted if language difficulties persist over time and there is a significant impact on everyday social interactions and/or academic achievement (Bishop et al., 2017). Additional considerations to prevent misdiagnosis when assessing a bilingual child include considering their competence in both their languages, comparing to an appropriate reference group, and judicious task selection when assessing proficiency in L2.

1.4.1 Considering L1 and L2

Any assessment for ELLs needs to consider the child's linguistic competence in both L1 and L2, since the child may be dominant in the minority home language rather than the language of testing (Paradis, 2016). While it may be ideal to assess a bilingual child in their first language, this is not always feasible, particularly in populations with diverse languages spoken (Boerma & Blom, 2017). Barriers to assessing both languages include the lack of bilingual SLPs across languages and the absence of validated and culturally appropriate assessments for each language (Boerma & Blom, 2017; Paradis et al., 2013). Even if samples can be collected in a child's first language, it can be challenging to analyze them due to a lack of research on typical and atypical language development in these minority languages (Paradis et al., 2013).

Thankfully, there are other ways of obtaining information regarding development in the L1. Moreover, there is growing evidence that select English-language measures, parent report, and teacher reports can, in combination, serve to identify ELLs with language disorders in the early school years (Boerma & Blom, 2017; Paradis et al., 2013).

1.4.1.1 Parent and Teacher Report

In many cases, it seems that assessing an ELL in English is the most feasible option. However, information about a child's first language skills is crucial and should not be overlooked (Bonifacci et al., 2020). It is not always possible to assess L1 proficiency directly and so an alternative is to obtain information about the child via a background questionnaire (Paradis et al., 2013). Parents, or other primary caregivers, offer an important perspective into the child's early development (Abutbul-Oz & Armon-Lotem, 2022). The delayed acquisition of language milestones in L1 can be the first indication of language impairment that persists into childhood (Dale et al., 2003). For example, language difficulties at age 2 is a risk factor for language difficulties at age 3 and 4 years (Dale et al., 2003). Additionally, language difficulties that are evident after age 5 are likely to persist into adolescence (Bishop et al., 2017; Charest et al., 2019; Stothard et al., 1998). Parents typically have observed their child's language development from birth and can compare it to that of other children in their family or social circle that share similar environmental factors, which is unlikely to be possible for most professionals (e.g., SLP, teacher). Furthermore, parents have knowledge about their child in their natural environment and context (Abutbul-Oz & Armon-Lotem, 2022).

The Alberta Language and Development Questionnaire (ALDeQ) (Paradis et al., 2010) consists of four sections: *early milestones* (e.g., How old was your child when he/she first spoke a word?), *current first language abilities* (e.g., Compared with other children of the same age, how do you think that your child expresses him/herself?), *behaviour patterns and activity preferences* (e.g., How quickly/how easily does your child learn new things?), and *family history* (e.g., Is there anyone among the child's immediate family or other relatives who had difficulties learning to read and write, in speaking and pronunciation, slow to learn to talk?). Paradis et al. (2010) found that early milestones and current first language abilities were respectively strongest and second strongest sections for discriminating between the TD and language impaired children. Similarly, a study by Grimm and Schulz (2014) investigating predictors, contributing factors, and misdiagnosis rates in German monolingual and bilingual TD children and children with a language disorder, found that late onset of the single-word and multi-word stage, and family history of DLD were predictive. However, Grimm and Schulz (2014) reason that child- and family-related factors should complement standardized assessments since at most half of the children would have been identified based on reported data for contributing factors alone. Furthermore, in settings where there are no standardized measurements for monolingual and bilingual children, a child's language development should be assessed using informal tools such as parent and teacher reports and spontaneous speech samples in the LoI (Grimm & Schulz, 2014).

Similarly, teacher report has clinical utility in identifying children with a language disorder. Pratt et al. (2022) explored the use of the Inventory to Assess Language Knowledge (ITALK), a parent and teacher questionnaire, to screen for DLD in Spanish-English bilinguals. The

ITALK consists of questions related to speech intelligibility, vocabulary knowledge, sentence length, grammaticality, and language comprehension. The authors found that teacher report of English and parent report of Spanish showed the best classification accuracy for children in kindergarten. However, in second grade, the strongest predictors were parent and teacher reports in English (Pratt et al., 2022). The authors reasoned that this may be due to the increased exposure to English for the children and parents. Children would be expected to increase their English language proficiency with time given sufficient quality and quantity of opportunities to hear and speak the LoI. These increased learning opportunities may lead children to experience a shift in language dominance from L1 to English (Kohnert, 2010), although variability in language dominance patterns is to be expected (Peña et al., 2023).

Shifts in language dominance may also lead to L1 attrition. When the child who speaks a minority language at home begins English preschool or school, there is often a reduction in exposure to the L1 if it is not supported in the educational environment (Paradis et al., 2010). As L1 attrition is common among ELL children, clinicians must be aware that parent report of low L1 abilities once a child has entered school may in some cases be a sign of L1 attrition rather than a language disorder (Paradis et al., 2010).

1.4.2 Reference Groups

Another issue with assessing bilinguals is determining a reference group. For the reasons mentioned previously, TD bilingual children often score below monolingual norms on standardized assessments, which may lead in some cases to incorrect identification of DLD in

ELLs. Paradis et al. (2013) were interested in determining if TD children and those with a language disorder could be differentiated based on norm-referenced English-language measures and a parent questionnaire. They found that 24% to 78% of their sample of TD children (aged 4;10 to 7;2) scored below monolingual expectations depending on the test – a result that could likely contribute to overdiagnoses of DLD. Since comparing bilinguals to monolingual norms would likely lead to overidentification of language disorders, a preferable alternative is to compare bilinguals to other bilinguals (Paradis, 2016). In practice, standardized norm-referenced instruments with normative samples consisting of bilingual children are very rare (e.g., BESA for Spanish-English bilingual 4- to 6- year olds, Peña et al., 2018), and to our knowledge, no such test has been developed for ELLs who speak the main home languages reported in Vancouver.

1.4.3 Language Assessments That Show Promise for Assessing Bilingual Children in the LoI

ELL children form a heterogenous group as they come from various language backgrounds and have variable experiences learning English. Furthermore, L1 typology impacts L2 acquisition so that children with different L1s may have dissimilar trajectories learning English (Hao et al., 2018). As previously reported, TD Chinese-English bilinguals might have slower morphology acquisition in English compared to children whose L1 is structurally more similar (Paradis, 2011, 2016). Therefore, an ideal language assessment should be less sensitive to language-specific knowledge. Assessments of this nature include nonword repetition and narratives.

1.4.3.1 Nonword Repetition

Nonword repetition (NWR) is described as a processing measure that targets language-learning capacity rather than prior knowledge of the L2 (Paradis, 2016). A systematic review of the discriminative power for detection of DLD in monolingual and bilingual children found that NWR tasks had a large mean effect size for children aged 2;0 to 8;11 (Schwob et al., 2021). Subgroup analyses showed slight variations in effect sizes depending on linguistic status of the children (monolingual or bilingual) and language specificity of the task (language-specific or quasi-universal). The effect size was larger for the monolingual subgroups (1.61) compared to the bilingual subgroups (1.36) and for the quasi-universal task (1.73) versus the language-specific task (1.52). However, language status, age, and language specificity of the task did not significantly influence the effect sizes. The authors note that NWR is a promising diagnostic tool but underscore that it is imperative to consider the child's linguistic background and complement with other assessments to obtain a comprehensive clinical picture and improve diagnostic accuracy (Schwob et al., 2021).

Other studies offer additional support for merging data sources. For example, Boerma and Blom (2017) report that the combination of results from a parent questionnaire (early language development in L1) and NWR when assessing bilingual children (mixed L1 and Dutch; mean age 6;0; range 4;10 to 7;2), increased sensitivity from 70% to 90% while specificity remained high (decreasing sightly from 90% to 87%) when compared to relying on a parent questionnaire alone. Including measures of narrative comprehension and production focused on macrostructure into the statistical model further improved sensitivity and specificity to 97%. The following sections will report in more detail on narrative analyses.

1.4.3.2 Narratives

It is well-established that children with DLD obtain lower scores on narrative tasks relative to TD children (Winters et al., 2022). However, existing studies vary widely regarding how narratives are elicited and analysed and with respect to the population of study. We first begin with an overview of narrative development, approaches to analysis, and elicitation strategies. A discussion of the utility of narratives to assess bilingual students will follow.

1.4.3.2.1 Narrative Development

Clinicians using narrative assessments gain rich information about the child's linguistic abilities (Govindarajan & Paradis, 2019). This is because oral narratives require both linguistic and conceptual knowledge (Govindarajan & Paradis, 2019). Storytelling requires the narrator to draw upon general event knowledge or memories and sequence this information into a cohesive plot (Hudson & Shapiro, 1991). Fictional stories that are generally used clinically feature a protagonist's goal and related attempts (or actions) and consequences; this means that the narrator must make it clear how the protagonist's actions are relevant to attainment of the goal (Berman & Slobin, 1994). Storytelling also requires discourse-pragmatic skills such as being explicit about the sequence of events and referring to characters so that the listener can adequately follow the plot of the story (Colozzo & Whitely, 2014; Govindarajan & Paradis, 2022). The narrator must also monitor the goal plan with each successful or unsuccessful attempt of the protagonist in relation to the plan (Berman & Slobin, 1994).

Children as young as 3-years of age have knowledge of various grammatical forms and words to describe individual events, but they show little inclusion of goal-directed behaviours in their narratives (Berman & Slobin, 1994; Trabasso et al., 1992). Reference to goal-directed behaviour is important to relate actions to the overall goal of the story (Berman & Slobin, 1994). By 5 years of age, children have knowledge of narrative structure and become more explicit in sequencing relevant actions related to a goal and an outcome (Berman & Slobin, 1994), although across age-levels, attempts are usually most frequent (Trabasso & Nickels, 1992). Trabasso et al. (1992) found that the use of explicit purposeful attempts (i.e., actions with explicit intentions) increased substantially between ages 3 and 5 years and continued to grow until age 9 years when performance became indistinguishable from that of adults.

Nine-year-olds further differ from younger children by their inclusion of references to frames of mind and thereby the ability to attribute inner states and emotions to characters (Bamberg & Damrad-Frye, 1991; Berman & Slobin, 1994). The inclusion of evaluative comments such as frames of mind, link sequential events together and provides meaning to the story (Bamberg & Damrad-Frye, 1991). While 9-year-olds make more references to inner states and affective responses of the protagonist, these children are still developing this ability when compared adults (Berman & Slobin, 1994). Adult story retells not only refer to inner states more frequently but incorporate these in the beginning of the story (Berman & Slobin, 1994). This sets the scene for the story and provides motivation for the protagonist's actions (Berman & Slobin, 1994). Hence, although narrative abilities develop significantly between preschool ages and late primary grades, producing high-quality narratives is a skill that continues to develop into adulthood (Berman & Slobin, 1994).

1.4.3.2.2 Macrostructure and Microstructure

Researchers have generally separated narrative analyses into macrostructure and microstructure. To simplify, macrostructure reflects the content and ideas of the story while microstructure relates to the word choices to connect them. *Macrostructure* reflects a child's ability to use language to organize ideas into a coherent story; it corresponds to the plot structure. Much of the research on macrostructure is based on story grammar, a narrative model proposed by Stein and Glenn (1979). In the story grammar model, prototypical stories consist of setting information (time, place, characters) and one or more episodes. An episode is a sequence of story elements that represent the event that motivates a character's actions (initiating event), the character's internal responses and external responses (attempts or actions) to the event, the outcome (or consequences) of the character's resulting actions, and reactions to the outcome (Stein & Glenn, 1979). Macrostructure knowledge is deemed transferable across languages and thereby proposed to be less biased against children from different language backgrounds (see below) (Boerma & Blom, 2017; Squires et al., 2014).

Microstructure reflects the child's ability to produce a story that hangs together using lexical and morphosyntactic components (Govindarajan & Paradis, 2019). Microstructure is more closely tied to structural language (or form) and thus relies on specific linguistic knowledge in the target language (Govindarajan & Paradis, 2019). There may nonetheless be utility in including both macrostructure and microstructure measures to gain a more complete view of children's narrative abilities as they gain in proficiency in the LoI.

1.4.3.2.3 Narrative Elicitation

Before a narrative sample can be analyzed, decisions must be made with respect to how the story will be elicited. Story generation and story retells are two options. To generate a story, a child must draw on world knowledge and memories to conceptualize the story schema, and then produce logically organized utterances (Merritt & Liles, 1989). In story retelling, the child must recognize and understand the story schema in the previously heard story, and then reproduce it in their own words (Merritt & Liles, 1989). Merritt and Liles (1989) investigated the clinical utility of story generation and story retelling tasks in groups of children (aged 9;0 to 11;4) with and without DLD. While both types of tasks appear to be an effective measure of narrative ability, the authors conclude that story retelling is more clinically useful since the retold stories were longer, contained more story grammar components, and more complete episodes for both groups of children (Merritt & Liles, 1989).

It can be challenging to elicit language samples from children with language difficulties, yet longer narrative samples may be more representative of a child's use of language in connected discourse (Merritt & Liles, 1989). Therefore, longer samples provide a more thorough picture of a child's language use as well as more utterances to analyze microstructure (Merritt & Liles, 1989). For the child, story retelling may also be simpler since they have the benefit of following a previously presented script (Boerma et al., 2016). With the combination of a script and wordless pictures (which the child can refer to when telling the story), retell may reduce the demands on a child compared to generating new content. Therefore, the processing load may be lower allowing the child to focus on the content and form of their narratives (Charest & Johnston, 2011; Colozzo et al., 2011).

There are also practical benefits to administering story retells over story generation tasks. The children's stories are generally more succinct (i.e., fewer irrelevant details or topic changes), and since the targets are known from the story model, there is minimal inferencing required from the coder (Merritt & Liles, 1989). The latter is particularly important for assessing bilingual students as they may still be mastering the phonological system of the L2 and thus make phonological errors which could make it more difficult to identify whether an utterance fulfills a story grammar element in a generated story. Knowing the target story also increases reliability and the speed at which stories can be transcribed and scored (Merritt & Liles, 1989). Finally, it facilitates comparisons across children (or within child over time) and makes macrostructure analysis more predictable.

Winters et al., (2022) concluded in their systematic review that there is insufficient evidence to determine whether narrative format (generation vs. retell) affects the ability to discriminate between children with DLD from their typical peers. Results from studies that have focused on story grammar analysis in stories of monolingual and bilingual children align with this conclusion (Boerma et al., 2016; Rezzonico et al., 2015). Retell would therefore seem to offer advantages that make it best suited for the current study.

1.4.3.2.4 Existing Elicitation and Scoring Protocols

There is not a single agreed upon way to code for macrostructure or microstructure. Nonetheless, we will review a few existing protocols that have been proposed for research or clinical application.

With respect to macrostructure, Stein and Glenn's story grammar model (Stein & Glenn, 1979) has had wide-ranging influence (e.g., MISL, ENNI, MAIN, see below). Most systems code for element types and episodic structure. The *MISL* (Monitoring Indicators of Scholarly Language) was designed as a progress monitoring tool for narrative proficiency (Gillam et al., 2017). The Macrostructure subscale of the MISL includes seven types of story grammar elements (Character, Setting, Initiating Event, Internal Response, Plan, Action/attempt, and Consequence), each of which is given a rating from zero to three, with higher scores reflecting more complex narrative structures (max. score of 21). To illustrate, Initiating Event(s) are defined as "event(s) that motivate characters to take action". Ratings from 0 to 3 corresponds respectively to "Series of descriptions, no indication of"; "Event stated, does not motivate action"; "One event stated that motivates action"; and "Two or more events that motivate separate actions (complex episode story)" (Gillam et al., 2017, p. 100). The MISL was originally developed for story generation and meant to be applicable to different stories. It has since been adapted to scoring story retells of the wordless picture books *Frog On His Own* and *One Frog Too Many* by Mercer Mayer (Squires et al., 2014). The general structure of the rubric was maintained, but precise words the child had to include in their retell to receive credit were added for each category. For instance, for the Initiating Event category, a rating of 0 is given for "No problem or 'starting' event; a rating of 1, 2 or 3 points is attributed depending on how many (1, 2, or 3) of the following were included: "___The boy <u>saw</u> a box/present." "___The big frog <u>said</u>, 'I don't like you.'" "___The boy <u>wouldn't</u> let the big frog get on the raft." "___The boy was shocked at what he saw, how did the big frog get there and where was the little frog". This rubric has the advantage of being straightforward to apply (once it has been adapted to a

particular story). It nonetheless strays away from the story grammar model by focusing more on specific content and exact words. Furthermore, by providing a rating for a story element category rather than scoring for individual items, the episodic structure of the retell is no longer apparent, as elements from different episodes are considered together.

The Narrative Scoring Scheme (NSS) was designed to incorporate features of story grammar and higher-level narrative skills (Heilmann et al., 2010). The rubric is grouped into seven categories, each scored from 1 (Minimal/Immature) to three (Emerging) to five (Proficient) (max. score of 35). Scores in between are undefined, and raters are expected to use their judgment. Story-specific rubrics (SALT Software, 2016) have been developed for the stories that correspond to the Systematic Analysis of Language Transcripts (SALT) story retell reference databases (SALT Software, 2020) (see Methods for detail). The rubric includes a variety of items that do not all fall within macrostructure (e.g., referencing/listener awareness; cohesion) or that include multiple elements that would correspond to different types (conflict/resolution and event/reaction). The rating scheme is also underspecified which, in our experience, makes it difficult to apply reliably between raters or stories.

The Multilingual Assessment Instrument for Narratives (MAIN) is an instrument for assessing narrative skills in children who are growing up learning one or more languages. The MAIN is designed to evaluate production and comprehension of narratives in children aged 3 to 10 years (Gagarina et al., 2012). The protocol includes four parallel stories that each have six-picture sequences. Each story consists of 17 story elements in total: two settings (time and or place) and three 5-element episodes (i.e., internal state as initiating event, goal, attempt, outcome, and internal state as reaction). The MAIN has scoring criteria for macrostructure

production and comprehension. In terms of production, the MAIN can be used to assess inclusion of story structure elements, and sequence of goal, attempt, and outcome (referred to as structural complexity). MAIN stories are short and quite simple, and thus seem best suited for preschoolers or kindergartners.

The Edmonton Narrative Norms Instrument (ENNI) is another protocol that has been developed to assess narrative ability (Schneider et al., 2005). Simple illustrated stories of different length are presented to the child who is asked to tell the story. Story grammar rubrics have been developed for two of the stories. Scoring sheets clearly lay out the story macrostructure element by element and provide explicit scoring instructions regarding what can be credited for each element (e.g., Character 2—elephant / female / girl (or any type of animal such as cow) [not pronoun], score 0 or 1; Setting—at swimming pool / going swimming / are playing / has/is holding airplane / one asks other to play, score 0 or 1; Initiating Event—G playing with airplane/making airplane fly / G shows/gives E his airplane, score 0 or 2). The ENNI stories seem best suited for preschoolers or kindergartners. The scoring rubric, however, clearly highlights the episodic structure, makes it possible to consider which element types were included in the story and presents with clear, simple instructions—all features that we sought to incorporate when developing rubrics for this study (see Methods).

With respect to microstructure coding, there is considerable variability across studies (see Winters et al., 2022). For instance, the Microstructure subscale of the MISL includes coordinating conjunctions, subordinating conjunctions, mental verbs, linguistic verbs, adverbs, elaborated noun phrases, grammaticality, and tense. Each of the categories is given a rating from zero to three. This can be problematic as the ratings are not length-normalized which

means that it may be difficult to highlight differences between children and to avoid ceiling or floor effects (e.g., grammaticality is rated based on the number of grammatical errors in the story, from three or more to none).

By following standard conventions from SALT and using the software, it is easy to obtain multiple measures that tap productivity (in words or utterance), utterance length, lexical diversity, verbal fluency (mazes, pauses), and grammatical accuracy (Miller & Iglesias, 2020). Many authors have taken this approach (Charest & Skoczylas, 2019; Peña et al., 2023; Rezzonico et al., 2015; Thordardottir & Weismer, 2002).

1.4.3.2.5 Assessment Strategy

In a systematic review and meta-analysis focusing on narrative performance of children with DLD, Winters et al. (2022) considered a wide range of variables grouped into three assessment types: macrostructure (i.e., referencing, cohesion, story grammar, complete episodes, etc.), microstructure (i.e., measures of accuracy, lexical diversity, fluency, length and complexity) and internal state language (i.e., motivational and perceptual verbs). They concluded that story grammar and grammatical accuracy resulted in the greatest differences between children with and without DLD. It should be noted that the results were largely from data collected with monolingual children, but bilingual studies were included.

The next sections will focus on the literature reporting on the narrative performance of children with DLD and their typical peers as it relates to Chinese-English sequential bilinguals.

1.4.3.2.5.1 Macrostructure

Many studies with bilingual children from various linguistic backgrounds and ages have found that children with DLD perform less well than their bilingual TD peers on macrostructure measures (Boerma et al., 2016; Fichman et al., 2017; Govindarajan & Paradis, 2019, 2022; Rezzonico et al., 2015; Squires et al., 2014). Although a few individual studies (e.g., Altman et al., 2016; Tsimpli et al., 2016) did not find differences between groups, these largely focused on goal, attempt, and outcome scores (Altman et al., 2016; Tsimpli et al., 2016); this may have reduced the variability by concentrating on fewer elements and the basic episodic structure (see Boerma et al., 2016 below). More consistent results have been found when looking at the full range of story grammar elements, which Sheng et al. (2020) point out is likely to be more clinically informative.

Boerma et al., (2016), investigated the diagnostic utility of narrative production and comprehension using the MAIN with monolingual (Dutch) and bilingual (mixed L1 and Dutch) 5- and 6-year-old children. They modified the scoring system to distinguish between 1) internal states (i.e., number of internal state terms, comprehension of internal state terms, and production of story grammar elements corresponding to internal states as initiating event or reactions) and 2) comprehension and production of basic episodic structure consisting of goal, attempt, and outcome. The optimal model included only the internal states variables and resulted in sensitivity and specificity levels of 85%. These results suggest that monolingual and bilingual children with DLD may have more difficulties with understanding and expressing the character's feelings and intentions compared to the basic story structure (Boerma et al., 2016).

In the study by Rezzonico et al. (2015), monolingual and bilingual pre-school children (mixed L1 and English) completed story retells of the Renfrew Bus Story at two test points (6 months apart) which were evaluated using an information score (i.e., content measured by inclusion of key events or key words). The authors describe that the information score was sensitive to the presence of a language impairment in the monolingual and bilingual group, but that monolingual and bilingual children with DLD had similar macrostructure scores despite differences in exposure to English (Rezzonico et al., 2015).

To document growth patterns in narrative skills of Spanish-English bilinguals with and without a language disorder, Squires et al. (2014) examined story retells from children in Kindergarten and again in Grade 1. For each time point, the students retold stories from wordless picture books (frog stories) in their L1 and L2. The story retells were analyzed for total macrostructure score and total microstructure score using an adaptation of the MISL (see above). Squires et al. (2014) found that, overall, the TD children outperformed those with a DLD for macrostructure at both time points in either language. While both groups of children improved in macrostructure scores, larger gains were obtained by the TD group. The authors also note that the macrostructure scores in Spanish in Kindergarten correlated with macrostructure scores in English in Grade 1, but a similar effect was not evident for microstructure. The authors conclude that this supports evidence for macrostructure abilities transferring more readily between language. Hence this study and that of Rezzonico et al. (2015) provide supports for including macrostructure analysis when assessing ELLs.

Although there is a considerable body of research that has explored narrative macrostructure among measures to identify bilingual children with DLD, significant gaps

remain. First, much of the research has focused on preschool or school-age children younger than 7 years of age (e.g., Altman et al., 2016; Boerma et al., 2016; Fichman et al., 2017; Govindarajan & Paradis, 2019; Iluz-Cohen & Walters, 2012). Considering that differences between 5-and 9-year-olds have been described and that narrative skills continue to develop into adulthood (Berman & Slobin, 1994), this raises the concern that certain narrative measures in preschool children may not be sensitive when used with older children.

Furthermore, little research has focused specifically on Chinese-English bilinguals although they may have been included in prior studies that investigate DLD in samples of children with diverse L1s (e.g., Boerma et al., 2016; Govindarajan & Paradis, 2019, 2022; Paradis et al., 2013). Some studies did investigate narrative ability across Chinese and English (e.g., Hao et al., 2019) but not potential diagnostic markers for DLD. This is why we turn to evidence from monolingual Chinese-speaking children.

Identification of DLD in monolingual Chinese-speaking children is not as well understood or researched as for English-speakers (Sheng, Yu, et al., 2023). Nonetheless, there is evidence that macrostructure can distinguish between groups of typical children and those at-risk of DLD across studies that have used different approaches to macrostructure analysis (total story structure (Sheng et al., 2020; Torng & Sah, 2020); inclusion of story element types (Hao et al., 2018), and high point analysis (Xue et al., 2022). With respect to individual element types, (Hao et al., 2018) found a difference between Mandarin-speaking TD children and children with DLD (4;3 – 7;11) for character, setting, internal response, action, and consequence (large effect sizes) and initiating event (medium effect). No difference was found for plan which they reasoned might be late acquired.

1.4.3.2.5.2 Microstructure

Turning to microstructure, common measures throughout the literature include accuracy (e.g. percent grammatical utterances), diversity (e.g., number of different words, NDW), fluency (e.g., repetitions, revisions), and length (e.g., mean length of utterance, MLU) (Winters et al., 2022). Grammatical accuracy is commonly proposed as the best differentiating variable to distinguish between typical and disordered language in both monolingual and bilingual populations (Ebert, 2020; Winters et al., 2022). Although grammatical accuracy is an indicator of DLD in many Indo-European languages (e.g., English), this may not apply to children who speak Mandarin (Hao et al., 2018; Sheng et al., 2020) or other languages with similar structure. Hao et al. (2018) found that Mandarin-speaking children with and without a language disorder (mean age 6;2) produced few ungrammatical sentences in a narrative production. The authors argue that since Mandarin has sparse morphosyntactic features (e.g., plural, tense, or aspect markers), there are fewer opportunities to produce utterances with morphological errors (Hao et al., 2018). This finding has also been replicated by Sheng et al. (2020) whose participants had a mean age of 5;8.

Studies with monolingual Chinese samples of children with and without a language disorder report variable results for measures of length and syntactic complexity (e.g., MLU or percentage complex sentences). Studies with younger children (4-to 6-year-olds) have generally found no differences for MLU (Sheng et al., 2020) and percentage of complex sentences (Torng & Sah, 2020), although Hao and collaborators (2018) did. Xue et al. (2022) included children from grades 1 to 5 and found between-group differences and increases with grade for MLU. Several studies (Hao et al., 2018; Sheng et al., 2020; Torng & Sah, 2020; Xue et al., 2022) have

also identified lexical diversity to be sensitive to the presence of DLD. Given the influence of L1 typology on grammatical development, these data from monolinguals may have implications when assessing school-aged Chinese-English bilinguals using microstructure measures. In particular, grammatical accuracy might not be a good marker for DLD in Chinese-English ELLs.

1.5 The Present Study

Bilingual children learning English as a L2 differ along multiple dimensions. First, the structural features of their first language may be more or less similar to those of the LoI (i.e., characteristic related to linguistic typology). Second, the length of exposure to the L2 can be highly variable even among children of similar chronological age or at the same grade level. At school, these children are expected to learn new skills and develop academic knowledge through a language that they are still trying to master. A few of these children may have a language disorder that affects each language the child speaks and understands, but most will show adequate proficiency in their L2 with enough time and exposure. Prior research has shown insufficiencies in the application of monolingual tests and norms to a bilingual population leading to possible misinterpretation of language abilities in bilingual students. In line with research recommending reliance on converging evidence (Boerma & Blom, 2017; Paradis et al., 2013), this study considers multiple data sources (and variables): narratives (macrostructure and microstructure), NWR, a background questionnaire, and English Language Learning (ELL) teacher reports. In addition, the participants will be compared to each other instead of to monolingual norms to establish an expected range of performance.

1.6 Research Questions and Analyses

This study is situated within a larger project that aims to identify language measures that could help SLPs ascertain which Chinese L1 ELLs are at risk of DLD by comparing them to each other (i.e., their true peers). The current study focuses primarily on data obtained from narrative retells, and more specifically on narrative macrostructure. Performance on NWR and narrative microstructure variables will also be considered.

The first aim is to describe the range of performance of Chinese-English bilinguals in Grades 2 and 3 in order to establish the range of *normal* on these tasks. These grade-level norms would then provide a basis of comparison to identify children who fall below the expected range. The second aim is to explore the possible clinical utility of including narrative macrostructure for identifying ELLs whose linguistic profile may be consistent with a diagnosis of DLD.

The following research objectives and questions will be addressed:

1. For Grade 3 and Grade 2 separately, describe the performance range of ELLs for macrostructure on a narrative retell task.
    a. What characterizes a typical narrative from the low- and average-scoring groups and how do they differ in their use of macrostructure?
    b. Which story grammar items could potentially discriminate between typical development and low-language proficiency?
        - We expect low-scoring stories to not only be comprised of fewer story grammar elements but also have less consistent use of certain item types; they will likely

omit story elements that refer to mental states such as internal responses and reactions. Additionally, these stories will infrequently include complete episodes.

- At the other end of the spectrum, we hypothesize that stories with higher scores will include each type of story grammar element and include complete episodes consistently.
- As narrative skills continue to develop into adulthood, we would also expect the overall use of story grammar elements to increase between Grade 2 and 3 due to the children being older and having completed more formal education.

2. For each grade, identify ELLs who obtain low scores compared to their peers on the language measures: macrostructure, microstructure, and NWR.
    a. Which children score below their peers for each language measure taken separately?
    b. Among those children who stand out from their peers, how do macrostructure, NWR, and microstructure overlap?
3. Examine to what extent the results from the language assessments are consistent with respect to i) ELL teacher evaluations and ii) SLP reports indicating suspected or diagnosed language difficulties. Consider background info on L1 development from parent reports as additional evidence.
    a. What combination of tasks or variables appear most sensitive to identify students with low language proficiency in Grades 2 and 3?

b. Does the information from the language variables identify low language proficiency children who may have been missed based on reports?

c. Are some children with a suspected language disorder not being identified by the language variables?

- The language variables analyzed in this study cover a range of linguistic subdomains. Bilinguals tend to converge on monolingual norms for cognitive-linguistic skills like story grammar and language-processing skills such as NWR earlier than language-specific skills like vocabulary and morphology (Paradis, 2016). Therefore, we expect that, as a group, the bilingual children will exhibit more difficulties in microstructure measures making it difficult to identify low-language proficiency in English using these measures on their own. In contrast, we hypothesize that lower cognitive-linguistic skills as demonstrated on the narrative retell task could be better indicators of low-language proficiency.
- Due to the heterogenous nature of DLD, it is expected that a combination of variables namely macrostructure, parent questionnaire and teacher report will be most sensitive in identifying low-language proficiency ELL children.
- Bilingual children are at risk of being over- and under-represented as having a language disorder and thus there is the possibility that this occurred for the children in this sample. We hypothesize that the scores from the language variables will only imperfectly align with ELL teacher evaluations and SLP reports, and it is deemed more likely that that some children will present with low-language proficiency based on the language measures but not ELL teacher or SLP

report. By Grade 3, the children will have had three-to-four years of English exposure so that concerns from teacher and SLP reports are more likely to be consistent with the results from language assessments.

- As mentioned previously, parent report provides valuable information about the early development in the child's L1 which is not captured by other assessment tasks. Therefore, we predict that the information from the parent questionnaire will provide additional valuable evidence when considering a hypothesis of low-language proficiency for a given child.

Chapter 2: Method

2.1 Design

This study includes data from a longitudinal sample of students who were assessed towards the end of the school year (April-June) at three time points (Grade 1, 2, and 3).[2] The focus is on Grades 2 and 3, supplemented with data from a background questionnaire collected in Grade 1. When the data were collected for the later two grade levels, most students would have had close to three (Grade 2) or four (Grade 3) full years of schooling in English augmented with ELL support. It is expected that differences related to length of exposure will be reduced, thereby making it more likely that differences from peers would be about language-learning difficulties. Additionally, by three to four years of schooling the students would be expected to have developed conversational proficiency and be on their way to developing academic proficiency (Cummins, 2000).

2.2 Participants

Children from 21 schools in the Vancouver School Board participated in the study. Participants were recruited among Chinese-speaking (Cantonese or Mandarin) students who qualified for ELL services in the first grade. In the BC school system, *English Language Learners* are students whose home language is not English and who need additional English language development support to access the curriculum and succeed academically (Province of British

[2] Once child was tested in September of fourth grade as they were absent for the end of the previous school year.

Columbia, 2024). Children with known developmental conditions (e.g., Autism Spectrum Disorder, Down syndrome) generally associated with language disorder were excluded from the study as they are not usually among the children with language learning needs who are difficult to identify.

From the original sample of 109 children at the first time point, 75 children (39 girls, ages 7;4 to 8;5) participated in Grade 2 and 68 children (35 girls, ages 8;4 to 9;4) for whom complete data are available participated in Grade 3.[3] The majority of students began schooling in English in Kindergarten in the Vancouver (*n* = 68) or Richmond school districts (*n* = 2); the other 5 students (4 of whom remained in the Grade 3 sample) began their schooling in a Vancouver public school at the beginning of Grade 1 (see Table 1).

[3] Although 70 children participated in Grade 3, narrative data are missing for two children (and nonword repetition data for one of them) as their recordings were lost due to technical issues.

Table 1

Demographic characteristics of the participants, by grade

Characteristic	Grade 2	Grade 3
Gender (n, %)		
Female	39 (52%)	35 (51%)
Male	36 (48%)	33 (49%)
Age (years; months)		
Mean	7;10	8;10
Range	7;4 to 8;5	8;4 to 9;4
LoE to English (months)		
Mean	51	64
Range	19 to 84	40 to 95

Note. LoE = Length of Exposure to English; this is an approximation as there was variability in how precise parents were in their report of when their child first began learning English.

Data regarding participant demographic characteristics were obtained at the time of initial consent in first grade though a background questionnaire completed by the parents. The consent package and questionnaire were translated into Cantonese and Mandarin so that parents could complete them in their language and writing system of choice (English, Simplified Chinese, or Traditional Chinese). The background questionnaire consisted of ten questions focused on the child's language exposure and language development history (including early rate of learning in the first language and history of speech and language service). Information was also collected regarding languages spoken in the home and mother's education level. See Appendix A and Table 2 for details.

Table 2

Language exposure, language history and maternal education at baseline (N = 75)

Parent Reported Information	
English Spoken at Home by Family (%)	
Mean (*SD*)	23 (21)
Range	0 - 90
English Spoken at Home by Child (%)	
Mean (*SD*)	44 (27)
Range	0 - 100
Age of English acquisition (months)	
Mean (*SD*)	43 (13)
Range	12 - 72
Early Proficiency (*n*, % of sample)	
Later/slower	14 (19%)
Typical	45 (60%)
Earlier/quicker	16 (21%)
History of SLP services (*n*, % of sample)	
Speech	4 (5%)
Language	2 (3%)
Speech and Language	6 (8%)
None	63 (84%)

Parent Reported Information	
Maternal Education (*n*, % of sample)	
Did not finish high school	17 (23%)
High school	20 (27%)
Vocational college	13 (17%)
Undergraduate degree	19 (25%)
Post-graduate degree	6 (8%)

Note: These data were collected from a background questionnaire completed by the parents at the end of Grade 1 (see Appendix A). The data are presented for the sample of children in Grade 2 (n = 75) with missing data for one child for the percentage of English spoken at home by the family. When considering the sample of children for whom there is complete data for both grades (*n* = 68), the statistics are almost identical as for the larger sample for the use of English in the home, age of acquisition, and distribution by levels of maternal education. Most of the children who were no longer in the Grade 3 sample were reported to have typical early proficiency in their first language (*n* = 6) and no history of speech-language services (*n* = 6).

While each child began learning a Chinese language from birth, the participants recruited for this study differ with respect to language experience and language development history. The children largely came from households where the family mainly spoke a Chinese language (Means 77% Chinese and 23% English), although some parents reported that this varied by family member (e.g., more English spoken by siblings than parents). Nineteen families spoke only Chinese at home. In contrast to the parents, the children tended to speak English and Chinese in more equal proportions (Means 56% Chinese and 44% English). While there was large variability for both parents and children, all parents spoke some Chinese at home, but not all children did.

The parents reported that, on average, the children started to learn English at 3;7 (range: 1;0 to 6;0) and many (64%) spoke little English at Kindergarten entry (i.e., few words or few words in sentences). The majority of parents did not report any language delays in Chinese. Nonetheless, 14 children had later or slower language proficiency in Chinese and 12 children had a history of SLP services. Not all children who may have been late talkers had a history of SLP services or vice versa. The parents of nine children reported both later and slower development and a history of SLP. Finally, the children came from various socioeconomic backgrounds as indicated by the variability in maternal education.

2.3 Materials and Procedures

Language assessments readily available to clinicians were chosen so that the assessment protocols could be replicated. Existing scoring protocols or tools were used whenever possible. The intent was to expand on existing practices and consider practical constraints. Data collected in Grade 2 and Grade 3 came from assessment protocols as well as reports from ELL teachers and SLPs. The assessment protocols were administered by six SLPs at the Vancouver School Board who were all experienced in language assessment. Table 3 details the data included for this study by grade level. Testing was completed in one session in Grade 2 but required two sessions in Grade 3.

Table 3

Data included in the current study

Data sources	Time Point	
	Grade 2	Grade 3
Assessment protocol		
Narrative retell	*Pookins Gets Her Way*	*A Porcupine Named Fluffy*
Nonword Repetition (NWR)	Dollaghan & Campbell	TILLS
KBIT Matrices subtest	n/a	X
Hearing screening	n/a	X
Data from ELL teachers and SLPs		
ELL Matrices - Primary	X	X
Report of children on SLP caseloads	n/a	X

Note. Dollaghan and Campbell (1998) Nonword repetition task. TILLS, Test of Integrated Language and Literacy Skills. KBIT-2, Kaufman Brief Intelligence Test – Second Edition. ELL Matrices—Primary for Oral Language, Reading, and Writing.

Narrative retell and NWR tasks were administered in each grade. Additional details regarding the language tasks are provided in later sections. A hearing screening was completed in Grade 3 to ensure hearing status was not implicated in potential language difficulties; all children passed. Additionally, the matrices subtest of the Kaufman Brief Intelligence Test 2nd Edition (KBIT-2) was administered to the participants as a measure of nonverbal cognition (Kaufman & Kaufman, 2004); most children performed within average (*M* = 111.0, *SD* = 19.5), but with large variability (Range: 67 to 153; standardized *M* of 100 and *SD* of 15).

2.3.1 ELL Teacher Evaluations and Information Provided by SLPs

In British Columbia, the Ministry of Education has put forth *ELL Standards* which are meant to be used by educators to describe and monitor students' language proficiency from Kindergarten to Grade 12 (British Columbia Ministry of Education, 2017). As part of these Standards, the ELL Matrices are rubrics that provide detailed descriptors of language proficiency. Separate ELL Matrices exist for Oral Language, Reading, and Writing, and for Primary, Intermediary, and Secondary levels. Each ELL Matrix evaluates multiple aspects using a five-level scale : 1-Beginning (beginning stages of language development), 2-Developing (showing progress in developing their language skills), 3-Expanding (expanding in their language development to further access classroom learning outcomes), 4-Consolidating (consolidating their language skills in the academic environment), and 5-Bridging (bridging to the grade-level language proficiency of their peers). The Primary (K-3) Oral Language ELL Matrix assesses the following aspects: *Meaning* (i.e., vocabulary; word choice; expressing and understanding ideas), *Form* (i.e., grammar; syntax; phonology; fluency), and *Use* (i.e., functions of language; strategies; social/cultural conventions), each of which are evaluated for *Receptive* and *Expressive* proficiency. Although, the Oral Language Matrix includes 6 scales, a single summary score is given that is meant to reflect that child's overall proficiency level, based on where the child's abilities correspond to most of the descriptors. To illustrate, at the Expanding level (score of 3) "the student can participate in a conversation about familiar topics and some academic content" (British Columbia Ministry of Education, 2017, p. 21).

The Primary (1-3) Reading ELL Matrix evaluates *Strategies* (i.e., word attack skills; decoding; fluency), *Comprehension* (i.e., vocabulary; main ideas and details; retelling and

organizing information; locating and recording information; drawing inferences; interpretations and socio-cultural elements; knowledge of genres), and *Response and Analysis* (i.e., opinions and reactions; connections). Finally, the Primary (1-3) Writing ELL Matrix evaluates *Meaning* (i.e., ideas and information; use of details; strategies), *Style* (i.e., word choice; sentence fluency; voice), *Form* (i.e., organization and sequencing; connections and transitions; awareness of forms/genres), and *Convention* (i.e., capitals and punctuation; spelling; grammatical elements and syntax).

ELL Matrices were completed by ELL teachers once or more in a given school year, depending on the school. ELL Matrix data were obtained for Grade 2 and 3. The default was to use Term 3 data since most children had complete data and this also aligned most closely to the examination dates. In cases of missing data, Term 2 data were substituted as needed. Data for all three Matrices were included given that the Reading and Writing Matrices assess skills related to oral language and specifically include items that focus on story retell. For example, the Reading Matrix includes a student's ability to retell the sequence of a story. Relatedly, the Writing Matrix mentions the student's ability to produce a logical narrative with a beginning, middle, and end. Any child who had scores of less than 3 for two or more Matrices was considered low scoring for English language proficiency based on ELL teacher evaluations.

Finally, school SLPs provided information about which children were on their caseloads by the end of Grade 3 due to suspected speech and/or language difficulties.

2.3.2 Nonword Repetition

NWR tasks were administered in both grades. Each child was instructed to listen to a recording of nonwords and then immediately repeat each one after hearing it. The child's productions were audio-recorded for later transcription and scoring. For Grade 2, the instructions and stimuli followed the protocol developed by Dollaghan and Campbell (1998).This NWR task consists of 16 words (1- to 4-syllable length) that were designed to be minimally affected by the child's vocabulary knowledge in English and exclude late developing phonemes and consonant clusters (Dollaghan & Campbell, 1998). The task is freely available and has been widely used in research with monolingual children with and without DLD (e.g., Weismer et al., 2000). Various studies have proposed a cut off of 70% for Percentage Phonemes Correct (PPC) to distinguish between monolingual children with DLD and TD peers, including the original study by Dollaghan and Campbel (1998) and research by Weismer et al. (2002) which included a large sample (*N* = 359) of monolingual English-speaking TD second graders (mean age = 7;11). We opted for Percent Consonants Correct (PCC rather than PPC) in the current study for a few reasons. First and foremost, it is easier and faster to code, as transcribing vowels can be challenging. This decision is supported by results from a study with monolingual that found that PCC distinguished between DLD and TD groups (Vuolo & Goffman, 2020). Furthermore, Lee and Gorman (2013) report that Chinese-English sequential bilingual 7-year-olds produced similar levels of consonant errors as the English monolingual group, but lower accuracy on vowels on a NWR task. Thus, measures of vowel accuracy may be less sensitive to identify low proficiency in Chinese ELLs.

To calculate the PCC, the child's productions of the nonwords were transcribed, and then compared to the targets. For each nonword, the total number of consonants correct was equal to the length of the syllable. The total PCC was then based on the total number of consonants correct divided by the total number of consonants produced. The maximum number of consonants correct was 56.

In Grade 3, the NWR subtest of the Test of Integrated Language and Literacy Skills (TILLS) was administered (Nelson et al., 2016). The TILLS is a norm-referenced assessment battery designed to identify language and literacy disorders (Nelson et al., 2016). The TILLS NWR subtest consists of 24 words, ranging from 1- to 5-syllables in length. This task differs from Dollaghan and Campbell's task by its inclusion of late-developing phonemes and consonant clusters. Despite potentially being more language-specific, the task was chosen since it is more likely to be used by SLPs as it is commercially available and easy to score with correct/incorrect decisions made at the word (rather than the phoneme) level.

This task was scored by transcribing the productions of each nonword and comparing it to the target provided in the TILLS scoring sheet. Since this task was scored at the item-level, a score of 0 or 1 was attributed to each of the 24 nonwords, thereby resulting in a maximum score of 24.

2.3.3 Narrative Retell

The illustrated story books for *Pookins Gets Her Way* (PGHW; Lester, 1987) and *A Porcupine Named Fluffy* (APNF; Lester, 1986) were used to elicit narrative retells in Grade 2 and

Grade 3, respectively. These stories correspond to the stories used in the SALT story retell database for children in these grades (SALT Software, 2020). They have a long history of being used successfully with children of similar ages as those in our samples. Furthermore, they are already known to and used by clinicians who include language sampling and comparisons to existing (monolingual) norms in their clinical practice. Both stories have two main characters and amusing yet complex stories that unfold over multiple episodes and 32 illustrated pages. *Pookins Gets Her Way* is about a mischievous girl who always wants and succeeds to get her way – until she meets a magical gnome. In *A Porcupine Named Fluffy*, a spiky porcupine tries to become fluffy like his name, until he meets another animal with an incongruous name.

We followed the elicitation protocol from the SALT normative database (Appendix B). The children heard a recording of the story while looking at the pictures in the book. After listening to the story, they were asked to retell the story using their own words. The words on the page were covered to avoid children attempting to read the text rather than producing their version of the story. Each child's story was audio-recorded for later transcription.

2.3.3.1 Developing a Macrostructure Rubric

Two parallel rubrics were developed for the current project—one for each story. The rubrics were designed specifically for story retell and followed a story grammar framework to analyse macrostructure. As an initial step, existing rubrics were reviewed, including a version of the MISL adapted for story retell of two frog stories (Squires et al., 2014) and the existing NSS rubrics (Heilmann et al., 2010) (see Introduction for details). In the end, we opted to design a

novel rubric with line-by-line scoring so that the underlying episodic structure remained accessible and that it would be possible to consider which element types were included in a child's narration. Additionally, we wanted scoring to be easy with criteria that were specific enough to ensure reliability, yet not so prescriptive that the scoring relied on the inclusion of specific words rather than ideas representing story elements.

The macrostructure rubrics for APNF and PGHW are based on a story grammar system adapted by Merritt and Liles (1987) and by (Gillam et al., 2017) from the work of Stein and Glenn (1979). A story is potentially comprised of elements from the following *nine categories*: (a) *Settings*: information about the social and physical context, including time and location; (b) *characters*: story main and secondary characters, including their names; (c) *initiating events*: external and internal events that influence and cause a character to respond; (d) *internal responses*: the psychological state that motivates a character to formulate a goal plan, and that are related to the initiating event; (e) *plans*: the plans and intentions to achieve a goal; (f) *attempts* (or actions): the application of the goal plan actions meant to cause or lead to a resolution; (g) *consequences*: the attainment or nonattainment of the character's goal or other changes in the sequence of events caused by a character's actions; (h) *reactions*: a character's feelings about the attainment or nonattainment of a goal; and (i) *resolutions*: the final outcome of the story, which may involve a moral. The rubric for APNF was developed first and later served as a template for creating the PGHW rubric. For this study, the rubrics were applied to the transcribed texts of each participant's story.

2.3.3.1.1 APNF Macrostructure Rubric

2.3.3.1.1.1 Development

As an initial step, the story text for APNF was analyzed for story elements and episodic structure. This resulted in the plot being divided into three episodes, one of which is embedded (see Appendix C.2). The rubric was designed with the goal of clinical applicability. Hence, explanations of how to meet scoring criteria were integrated into the rubric. Furthermore, each story element was assigned 1 point to allow coders to make yes/no judgments for each story element. The general structure of the rubric consists of four columns: *story grammar element* (e.g., Initiating Event-1b), *representation*, which gets at the idea represented by that element (e.g. she meets a magic gnome), *target*, which is closer to the story text (e.g., "she met a magic gnome"), and *considerations for crediting the element* (e.g., Indicates the girl meets or sees someone).

The rubric was developed following an iterative process. Early drafts were applied to the transcriptions of the story retells of three TD monolingual children and then six bilingual children from our sample. Myself, my supervisor (PC) and a third coder (RH) closely involved in the project independently scored the stories, reviewed, discussed, and agreed on changes needed to be made to the rubric until everyone was satisfied.

2.3.3.1.1.2 Trial Process

After the rubric had undergone multiple revisions, it was time to involve novel coders. This would be the starting point to train research assistants (RAs). At this point, we were still trying to ascertain whether the rubric was sufficiently clear to be applied reliably with minimal training. Three RAs and one SLP (CW) were asked to code six transcripts using the rubric. PC and I examined discrepancies among the coders and how they related to my coding. While my scoring was seen as the reference, it was not aways the final score as discussions could lead to changes in how the rubric should be interpreted. The RAs were asked to review discrepancies among coders and the feedback provided to them before initiating another round of coding with the updated rubric. In the subsequent trial phase, the three RAs and PC coded another three transcripts. At this point, the RAs had reached high levels of reliability (above 80%). The discrepancies between each coder and myself were examined and modifications were made to the rubric, as appropriate. The 45-item APNF rubric was judged to be finalized.

2.3.3.1.2 PGHW Macrostructure Rubric

2.3.3.1.2.1 Development

Mirroring the process for creating the APNF Macrostructure Rubric, the story text for PGHW was analyzed for story elements and episodic structure. (Appendix C.1). The plot of this story was also divided into three episodes, one of which is embedded. This step was particularly useful to fill out the *Story Grammar Elements*, *Representation*, and *Target* columns of the rubric. Based on the experience developing the APNF Macrostructure Rubric, some descriptions

for “Considerations for crediting the element” were added to create a draft of the rubric that PC and I applied to a few story texts.

2.3.3.1.2.2 Trial Process

Once the rubric had undergone several revisions, the same three RAs involved in applying the APNF rubric were asked to code five transcripts. PC and I reviewed the transcripts together and came up with consensus scores. My scoring was used as the reference but wasn’t always the final score. The RAs were made aware of changes to the rubric and were asked to review the discrepancies between coders. The same coding and feedback process was repeated for another five transcripts. At this point, the RAs had reached high levels of reliability (each RA was at least 85% in agreement with LB). Minor changes were made to the rubric and the 45-item PGHW rubric was judged to be finalized. Although there are similarities between the story analyses (number of total elements, number of episodes, etc.) the way the items distribute and the weighting of specific elements is different from for the two stories.

2.3.3.2 Transcription and Coding of Narratives

2.3.3.2.1 Transcription of Narratives

Each child’s audio-recorded story retell was transcribed using SALT software (Miller & Iglesias, 2020). Standard SALT conventions were followed for transcription into communication units (C-units: main clause and attached subordinate clauses) and basic coding (i.e., bound

morphemes, errors, omissions, pauses). Following extensive training, each story was transcribed and coded by a first transcriber and then checked by a second person. A master transcriber was consulted for difficult transcripts or passages and also resolved discrepancies between transcribers. The transcripts were then ready to be coded for macrostructure using one of the macrostructure rubrics and for microstructure by extracting data using SALT.

2.3.3.2.2 Coding for the Story Retells

The APNF story retells were scored using the APNF Macrostructure Rubric (see Appendix D.2), which was developed first. The same three RAs that were part of the development process applied the finalized rubric to 58 out of 68 transcripts once they had reached high levels of reliability during training. Coding was completed in pairs with two people assigned to the same transcript. The RAs were instructed to score a block of assigned transcripts independently, and then to compare their scoring and resolve any discrepancies. As the master coder, I reviewed the discrepancies among the coders and provided feedback as appropriate to ensure coding consistency, looking for any patterns of errors, and consulting with my supervisor (PC) as needed to make final coding decisions in cases of uncertainty. Cohen's kappa (κ), a measure of inter-rater reliability, was calculated to determine the reliability between the three coding pairs. Additionally, the inter-rater reliability between the coder's finalized score and my own independent scoring was calculated for about 15% of the sample (n = 9) coded by the RA's. Inter-rater reliability was almost perfect between the coders ($\kappa = 0.845$, $p<0.001$) and between the coders and myself ($\kappa = 0.884$, $p<0.001$) (Landis & Koch, 1977). Any discrepancies between

the coders and myself were resolved in consultation with PC based on the finalized rubric. The other 10 transcripts coded during the development process were reviewed based on the final version of the APNF rubric.

After the trial phase and once they had reached high levels of reliability, the RAs applied the PGHW Macrostructure Rubric (see Appendix D.1) to 65 out of 75 transcripts. The process was identical to APNF. Inter-rater reliability was almost perfect between the coders ($\kappa = 0.874$, $p<0.001$) and between the coders and myself ($\kappa = 0.889$, $p<0.001$) based on 15% ($n = 10$) of the transcripts (Landis & Koch, 1977). The other 10 transcripts that had already been coded during the development process were reviewed based the final version of the PGHW rubric.

2.3.3.2.3 Macrostructure Variables

Once coded using either the APNF or PGHW rubric, story retells were analyzed for the total number of story grammar elements, and the nine aforementioned element types: Character, Setting, Initiating event, Internal Response, Plan, Action, Consequence, Reaction, and Resolution. Due to the low number of certain element types in the original stories, these elements were regrouped into modified categories: *Cognitive Elements* and *Physical Initiating Events*. Cognitive (or psychological) elements refer to any element that makes reference to the internal, mental states of the character. These items include Plan, Internal Response, Reaction, and Resolution items. Cognitive Initiating Events are also included in this grouping. This distinction within Initiating Events seemed important since a Physical Initiating Event is more like an Action (e.g., She meets a gnome), whereas a Cognitive Initiating Event (e.g., Fluffy

doubts he is fluffy) is more like a Plan. The last modified category, *Action-Consequence Pairs* (ACo pairs), was created to capture how well children connect attempts to their outcomes, in some cases resulting in non-attainment of the goal within a given episode. Thus, ACo pairs go beyond how many actions and consequences were included, by considering how often they were linked.

2.3.3.2.4 Microstructure Variables

For the purpose of describing the children's narratives, the transcripts were analyzed using SALT to obtain data for variables that reflect productivity (total number of utterances, TotNumUt), utterance length and complexity (mean length of utterance in words, MLUw; number of verbs per utterance, VPU), lexical diversity (moving-average number of different words, MA-NDW), verbal fluency (words per minute, WPM; percent total pause time, % pause time; percent total maze words, %MzWds), and grammatical accuracy (percent utterances with errors, %Errors). We chose mean length of utterances in words and not morphemes as a better choice to measure utterance length since children who speak a Chinese L1 may take longer to develop English morphology. For more detailed analyses, only four variables were included: MLUw, MA-NDW, %MzWds, and %Errors. Additionally, we considered two or more scores of 1 *SD* from the mean (above or below, depending on the measure) for these four variables as indicative of low microstructure.

2.4 Analyses

Descriptive statistics were obtained for macrostructure, microstructure, and NWR data to provide a thorough report of the range of abilities. A Pearson correlation between the length of exposure to English and the macrostructure total score was calculated to determine the relationship between these two variables. Statistical analyses were completed using SPSS Statistics software (IBM, 2023). To compare to monolingual norms for microstructure variables, means and standard deviations were extracted from the SALT retell normative databases by matching for age and grade to our samples for each story. The monolingual norms were then compared to our bilingual sample using Cohen's *d*. Effect sizes were interpreted as small (0.2), medium (0.5), and large (0.8) (Cohen, 1988).

Chapter 3: Results

This chapter begins with an overview of the range of abilities of the ELL children for the language measures: narrative macrostructure, narrative microstructure, and NWR. Data from ELL Matrices are also reported. Results and analyses focus on the sample of children at a given grade. To wrap up the results, we investigate variables that may account for individual differences (i.e., length of exposure to English), and compare our sample to monolingual norms for microstructure variables and NWR.

3.1 Overview

Across all measures, children showed considerable variability in both grades (Table 4). They were generally able to retell long and coherent stories with numerous story grammar elements. The children were also able to repeat nonwords with good accuracy. Ratings from the ELL Matrices suggest that students were generally at an *expanding* level of English language proficiency. By the end of the study, eight students had been referred for SLP services. Five of these children were identified as needing expressive language support (one speech and language), while the other three reportedly needed support only for speech (e.g., phonology).

Table 4

Overview of children's performance on language measures and ELL Matrices, by grade

Assessment	Grade 2		Grade 3	
	M (*SD*)	Range	*M* (*SD*)	Range
Macrostructure				
Total score (max. 45)	22.6 (7.6)	7 - 41	26.1 (6.3)	12 - 39
Microstructure				
TotNumUt	38.2 (9.2)	17 - 62	49.6 (12.2)	29 - 84
MLUw	8.1 (1.3)	4.5 - 11.2	6.5 (0.9)	4.0 - 8.7
NWR				
PCC	83.4 (9.9)	54 - 98	n/a	n/a
Items correct (max. 24)	n/a	n/a	18.4 (2.8)	10 - 23
ELL Matrix				
Oral Language (max. 5)	3.4 (0.9)	1.5 - 5	3.6 (0.9)	1.5 - 5
Reading (max. 5)	3.4 (1.0)	1 - 5	3.6 (1.0)	1.5 - 5
Writing (max. 5)	2.9 (1.0)	1 - 5	3.3 (1.0)	1.5 - 5

Note. TotNumUt = total number of utterances, MLUw = mean length of utterances in words, PCC = percent consonants correct.

3.2 Grade 2

3.2.1 Macrostructure

Results are first presented by story grammar element types across participants followed by how the participants performed on the rubric. In addition to total scores, we report on the

combination of elements that were included and episodic structure in order to determine expectations for a typical PGHW retell.

3.2.1.1 Item and Episode Difficulty

As illustrated in Figure 1, Episodes 1, 2, and 3 were of uneven length (26, 11, and 8 elements per episode), with the first being the longest. We considered how many children included each of the 45 elements as a proxy for item difficulty. There was considerable variability across items (see Figure 1). Each rubric element was included by at least 9 participants (12%). On the high end, two elements were included by 69 participants (92%). We used the distribution into quartiles to group items into three levels of difficulty: easy, average, and difficult.

Nine elements were identified as easy with 75% or more of participants ($n \geq 57$) including them in their stories. Easy items consist of Initiating Events (items 6, 7, 27, 32), Actions (items 17, 22, 23, 22), and Character (item 1). These Initiating Events are physical occurrences that are clearly represented by the pictures (e.g., item 7; She meets a magic gnome). Similarly, the easy Action items generally corresponded to visible, physical actions clearly illustrated in the book (items 22-24: the gnome puts Pookins in a pot, the gnome dumps soil on Pookins, and the gnome waters Pookins). Lastly, the Character item was met by referring to the little girl by her name, which is repeated numerous times throughout the story. Easy items were dispersed throughout the three episodes but were more concentrated within the first, longest, episode.

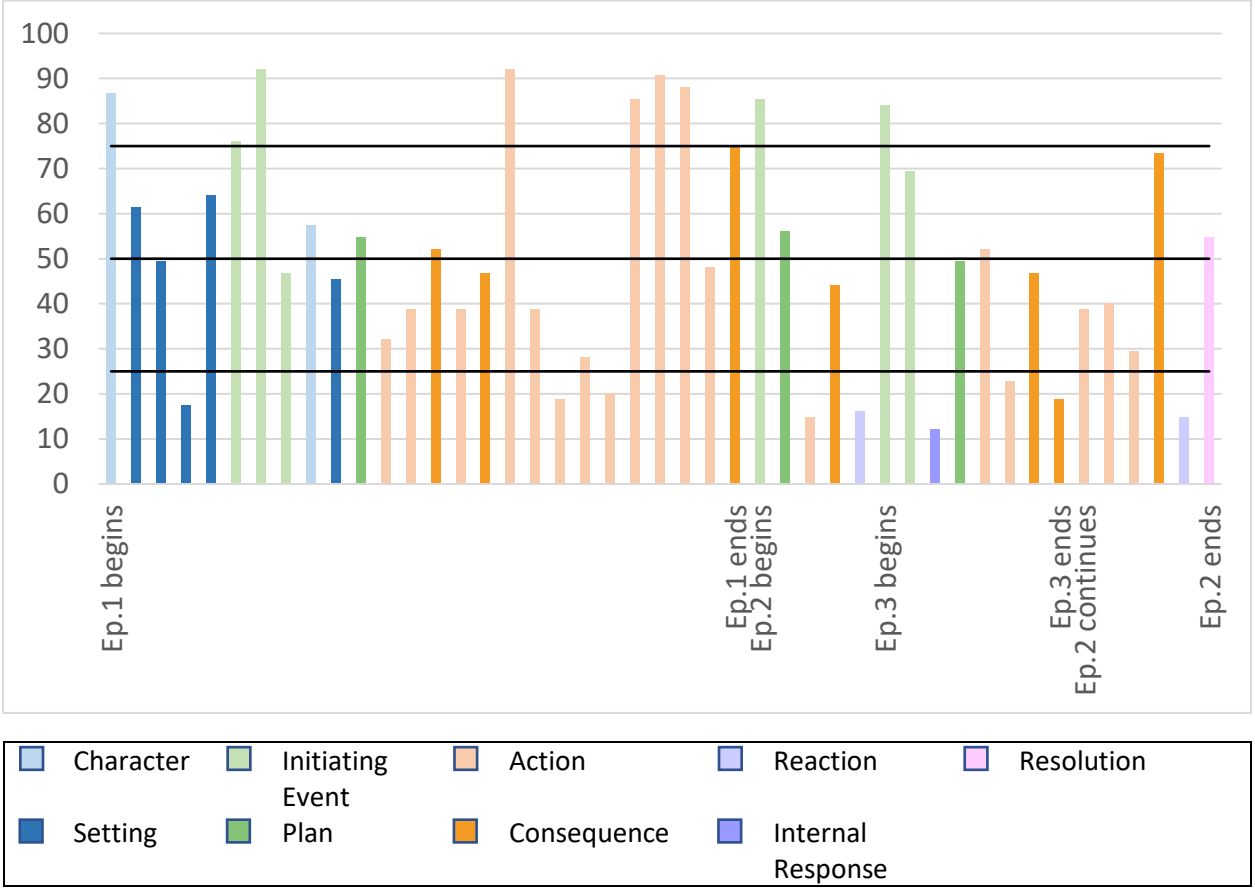
Another nine items were identified as difficult with 25% or fewer students ($n \leq 19$) including them in their stories. These correspond to Actions (items 19, 21, 29, 37), Consequence (item 39), Setting (item 4), Reactions (items 31, 44), and Internal Response (item 34). The difficult items generally refer to goal-directed behaviours, motivations, and internal states. For example, Pookins insists by threatening the gnome (item 19); the gnome changes his mind about helping Pookins because she helped him (item 39) and Pookins did whatever she wanted because nobody wanted her to misbehave (item 4). These items often required including a form of justification or explicitly stated intention and not just a simple description. For example, for Item 4 (Setting) some children would narrate that Pookins did whatever she wanted but failed to mention *why* she could do that. By definition, the Reaction and Internal Responses correspond to internal states (e.g., Pookins is happy). The difficult items were also dispersed throughout the story but were most abundant in Episodes 2 and 3.

As a proxy for episode difficulty, we considered the frequency of complete episodes (i.e., Initiating Event, Action, and Consequence; (Stein & Glenn, 1979). Episodes 1, 2, and 3 were complete in 75%, 11%, and 15% of stories, respectively. Hence, Episode 1 appeared to be considerably easier than the other two episodes. This is consistent with the distribution of easy and more difficult items, and was also likely affected by episode length, which resulted in more opportunities to include required elements in Episode 1 (see Appendix E.1 for further detail regarding scoring criteria for complete episodes).

In summary, students typically had a strong start by introducing the first main character and the problem of the story, but then had difficulties explaining how attempts were resolved and did not tend to refer to characters' mental states.

Figure 1

Proportion of children who included each item in their story retell



3.2.1.2 Performance on the Macrostructure Rubric

There was wide variability in total scores for the PGHW macrostructure rubric, with scores ranging from 7 to 41 (possible maximum of 45), and an average score of 22.6 (*SD* = 7.6). For this and later sections, PGHW story retells are divided into low-scoring (15 or less; below -1*SD*) and average-scoring (16 or more; -1*SD* and above) based on the total macrostructure score. An average-scoring story provides information about what might be expected for an ELL student in

Grade 2. In contrast, low-scoring stories provide information about the narrative profile of children who may have a language disorder. Occasionally, reference will be made to the sub-group of high-scoring stories (31 and higher; above +1 *SD*) as these narratives provide further evidence of the variety in quality of a typical narrative. For each group, story grammar elements (Table 5), and recombined story grammar elements (Table 6) will be considered.

Table 5

Story grammar elements, by group

Story Grammar Element	Group			
	Low (*n* = 15)		Average (*n* = 60)	
	M (*SD*)	Range	*M* (*SD*)	Range
Character (2)	0.9 (0.8)	0 - 2	1.6 (0.6)	0 - 2
Setting (5)	0.9 (0.9)	0 - 2	2.8 (1.5)	0 - 5
Initiating Event (6)	3.7 (1.4)	1 - 6	4.7 (1.0)	2 - 6
Internal Response (1)	0.1 (0.3)	0 - 1	0.1 (0.3)	0 - 1
Plan (3)	0.4 (0.5)	0 - 1	1.9 (0.8)	0 - 3
Action (18)	4.0 (1.4)	1 - 6	9.2 (2.9)	4 - 16
Consequence (7)	1.6 (0.9)	0 - 4	4.1 (1.7)	0 - 7
Reaction (2)	0.0 (0.0)	0	0.4 (0.6)	0 - 2
Resolution (1)	0.3 (0.5)	0 - 1	0.6 (0.5)	0 - 1

Table 6

Recombined story grammar elements, by group

Recombined Story Grammar Element	Group			
	Low (*n* = 15)		Average (*n* = 60)	
	M (*SD*)	Range	*M* (*SD*)	Range
Cognitive Elements (8)	1.2 (1.0)	0 - 3	3.8 (1.3)	0 - 7
Physical Initiating Events (5)	3.3 (1.2)	1 - 5	4.0 (0.9)	2 - 5
Action-Consequence Pairs (6)	0.5 (0.5)	0 - 1	1.9 (1.3)	0 - 5

Note. Cognitive states include all Plan, Reaction, Internal Response, and Resolutions items, and Cognitive Initiating Events. The remaining Initiating Events are physical. Action-Consequence pairs (ACo pairs) reflect the links between Actions and Consequences.

Sixty children produced average-scoring stories. Overall, 92% of these stories included at least the following elements: one Character or Setting, two Initiating Events, one Plan, four Actions, and one Consequence (see Table 5). Furthermore, 88% of the average-scoring stories included a least one Cognitive element (Cognitive Initiating Event, Plan, Reaction, Internal Response, or Resolution) and at least one ACo pair (Table 6). Finally, most average-scoring stories (85%) included at least one complete episode (*M* = 1.1, *SD* = 0.7, range 0 - 3).

Considering the top end, the ten highest scoring stories within this group showed an expansion of Settings, Actions, and Consequences as well as a broadening in element types, with the consistent inclusion of a Resolution. These stories also included multiple Cognitive elements and ACo pairs (at least 3 of each) and most also had two complete episodes (*M* = 1.8, *SD* = 0.4).

Turning now to the 15 participants (one fifth of the sample) who produced low-scoring stories, only four of these students (27%) included the minimum expected elements established based on the performance of the average-scoring group. Plans were particularly infrequent, and Consequences were generally not well related to their respective Actions. Only one-third of the low-scoring stories included at least 1 Cognitive element and 1 ACo pair (*n* = 5). It is worth noting, however, that the children in both groups scored similarly for Physical Initiating Events (see Table 6). With respect to episodic structure, approximately half of the low scoring-stories had one complete episode (*n* = 7; *M* = 0.5, *SD* = 0.5, range 0 - 1).

In combination, these data for low-scoring participants reflect unelaborated stories with scattered story grammar elements and minimal plot development. Largely, the narrators attempted to retell the original PGHW story but missed or modified important elements needed to form complete episodes and to capture the plot. This in turn made it difficult to know whether changes were deliberate, or whether they misunderstood the original story. There was a theme of uncertainty in most of the children's narrations evident by the need for frequent prompting from the experimenter, prosody reflecting doubt, numerous pauses, and children mentioning that they had forgotten parts of the story.

3.2.1.3 Summary of Expectations

The results of the average-scoring group provide benchmarks for the number and types of story grammar elements to expect in a typical story. To recap, a reasonable expectation for PGHW retells in Grade 2 would be for the stories to include at least the following: one

Character or Setting, two Initiating Events, one Plan, four Actions, and one Consequence (Table 7). Alternatively, expectations could focus on the inclusion of at least one Cognitive Element and one ACo pair (Table 7). Either of these criteria could be used in addition to the overall score to identify children who may have low language proficiency. Considering the presence of a complete episode is less likely to be useful, given the limited range of possible scores.

Table 7

Expected number of story grammar elements in a typical story, by type

Story Grammar Element	Number
Character or Setting	1
Initiating Event	2
Internal Response	0
Plan	1
Action	4
Consequence	1
Reaction	0
Resolution	0
Recombined Story Grammar Elements	
Cognitive Elements	1
Action-Consequence pairs	1

3.2.2 Microstructure

The story retells were generally intelligible and quite long (*M* = 38.2 utterances). There was large variability across all measures: story length (TotNumUt), utterance length/complexity (MLUw and VPU), verbal fluency (WPM, %PauseTime, %MzWds), and grammatical accuracy (%Errors) (see Table 8). The high level of disfluency evident by the microstructure measures is congruent with observations made when transcribing the stories.

Table 8

Summary of microstructure variables for PGHW (N=75)

Variable	*M* (*SD*)	Range
TotNumUt	38.2 (9.2)	17 - 62
MLUw	8.1 (1.3)	4.5 - 11.2
VPU	1.7 (0.3)	0.8 - 2.4
MA-NDW	56.6 (4.1)	46.9 - 64.8
WPM	67.8 (20.9)	12.0 - 110.0
%PauseTime	23.5 (13.4)	2.9 - 76.1
%MzWds	12.1 (6.1)	0.4 - 31.7
%Errors	34.8 (16.6)	5.3 - 82.4

Note. TotNumUt = total number of utterances, MLUw = mean length of utterance in words, VPU = verbs per utterance, MA-NDW = moving average number of different words, WPM = words per minute, %PauseTime = percent total pause time, %MzWds = percent total maze words, %Errors = percent utterances with errors.

To determine which students scored low on microstructure, the number of indicators of interest was reduced to four main variables: MLUw, MA-NDW, %MzWds, and %Errors. Scores more than 1 *SD* from the mean (below for MLUw and MA-NDW; above for %MzWds and %Errors) were deemed worthy of further consideration. The 12 participants with flagged scores for two or more indicators were identified as having low microstructure scores. A few patterns stood out. For those with two scores (*n* = 6), one variable was always MLUw (with %MzWds or %Errors) and for those with three (*n* = 6), %Errors and MA-NDW were most common.

3.2.3 Nonword Repetition

The mean for the NWR task was 83.4% (*SD* = 9.9) consonants correct, with a wide range from 54% to 98%. A total of 14 participants scored below -1 *SD* from the sample mean.

3.2.4 Teacher Evaluations: ELL Matrices

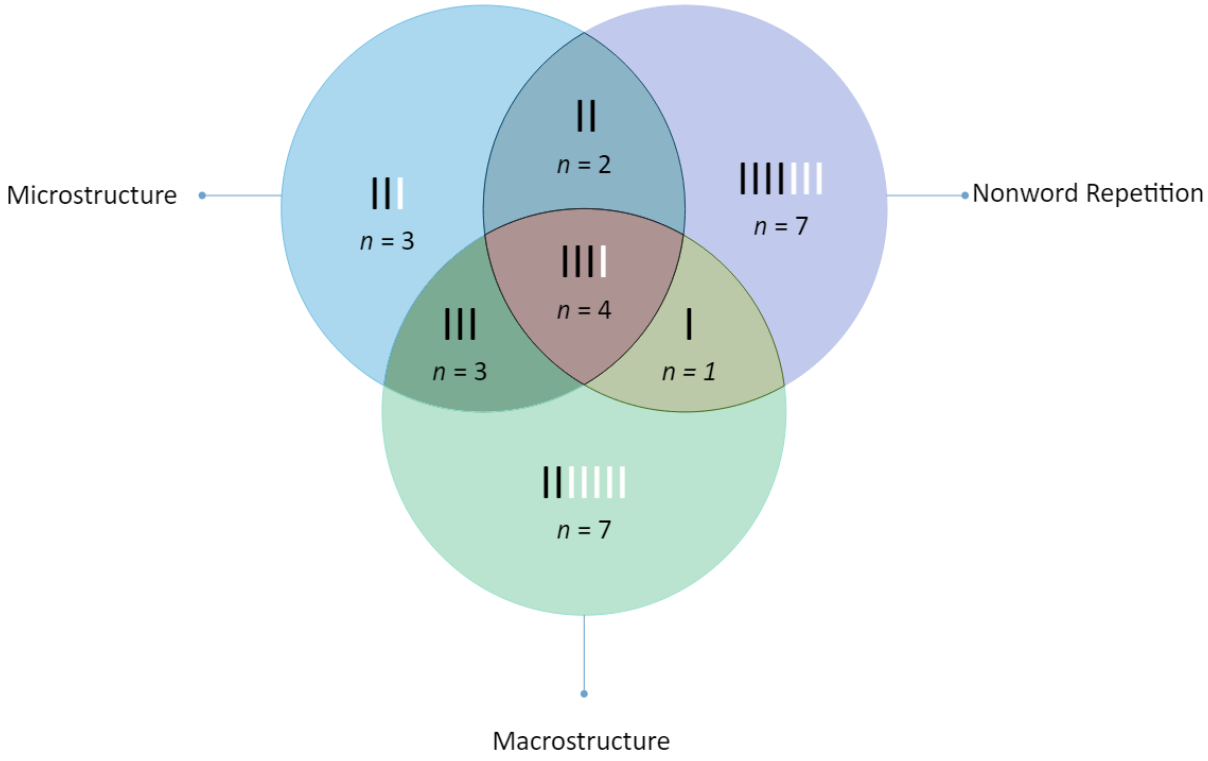
ELL Matrix data were available for 95% (*n* = 71) of the students. The averages were similar for Oral Language and Reading (3.4) and somewhat lower for Writing (2.9), but at the *expanding* level and with similar ranges (1 or 1.5 to 5) across Matrices (see Table 4). Six children were no longer receiving ELL services (i.e., had been delisted) by the end of Term 2. Furthermore, 22 children were considered to have low English proficiency based on having received two or more scores of less than 3 on the Matrices. In most cases, one of the low scores was for Oral Language (Oral Language + Reading + Writing, *n* = 15; Oral Language + Writing, *n* = 3; Reading + Writing, *n* = 4).

3.2.5 Low Language Proficiency

Next, the results from the language measures—macrostructure, NDW, and microstructure—will be combined to identify students who did not perform as well as their peers, which could indicate lower proficiency in English. In line with the aims of this study, the results from the macrostructure analysis will receive greater attention. Data provided by ELL teachers and SLPs will provide supplementary information against which to compare results from the language measures.

Figure 2

Participants who scored more than 1 SD from the mean for macrostructure, microstructure (two or more indicators), or NWR.



Note. Black bars identify participants who had low scores on two or more ELL Matrices.

3.2.5.1 To What Extent Do Potential Indicators of Low Language Proficiency Overlap?

We now turn to describe the degree of overlap for the language measures. We also considered whether a child had been evaluated at early stages of English proficiency by their ELL teacher.

Figure 2 illustrates how the 27 children with low scores on one (*n* = 17) or more (*n* = 10) of the language measures distributed in Grade 2. The children who received low scores on multiple measures, had variable profiles, with 4 children obtaining low scores for all three. The remaining 17 children scored low on single measures – most commonly macrostructure or NWR. Low ELL Matrices scores (identified with black font in Figure 2) best aligned with the language variables when considering multiple measures (9 of 10, 90%) compared to single measures (8 of 17, 47%).

Focusing more specifically on macrostructure, there is moderate to good overlap for low scores and low ratings on ELL Matrices (9 of 15 children, 60%), with better overlap with ELL Matrices when a child scored low on macrostructure and NWR (4 of 5 children, 80%) or macrostructure and microstructure (6 of 7 children, 86%). However, this relationship does not hold when macrostructure is the only low score: five of the seven students presented with no other indicators of low language proficiency; in fact, one child had been delisted from ELL service. Hence, macrostructure on its own may not be generally informative for identification of low language proficiency. The same is likely the case for microstructure and NWR on their own.

3.2.5.2 How Do the ELL Matrices Relate to the Language Measures?

Most of the children (17 of 22, 77%) who obtained low scores on two ELL Matrices (Oral Language, Reading, Writing) also obtained low scores for macrostructure, microstructure, and/or NWR. Five students did not obtain low scores for any language measures. For three of the five, only the Reading and Writing Matrices obtained ratings below 3. Thereby, it is plausible that these children did not have difficulty with spoken English but rather literacy. Hence the Oral Language ELL Matrix data may be most informative from an SLP perspective.

3.2.5.3 How Does SLP Report Relate to the Language Measures and the ELL Matrices?

Focusing now only on the five children who were identified with language difficulties by an SLP, all but one scored low on multiple language measures and they all received low ratings on ELL Matrices suggesting early stages of proficiency. Macrostructure and/or NWR scores best accounted for the linguistic profiles of the children. Among the three children identified with speech only concerns, one child scored in the low range for both macrostructure and microstructure, and another on NWR, whereas the third did not stand out from peers for any indicators (language measures or teacher evaluations).

3.3 Grade 3

3.3.1 Macrostructure

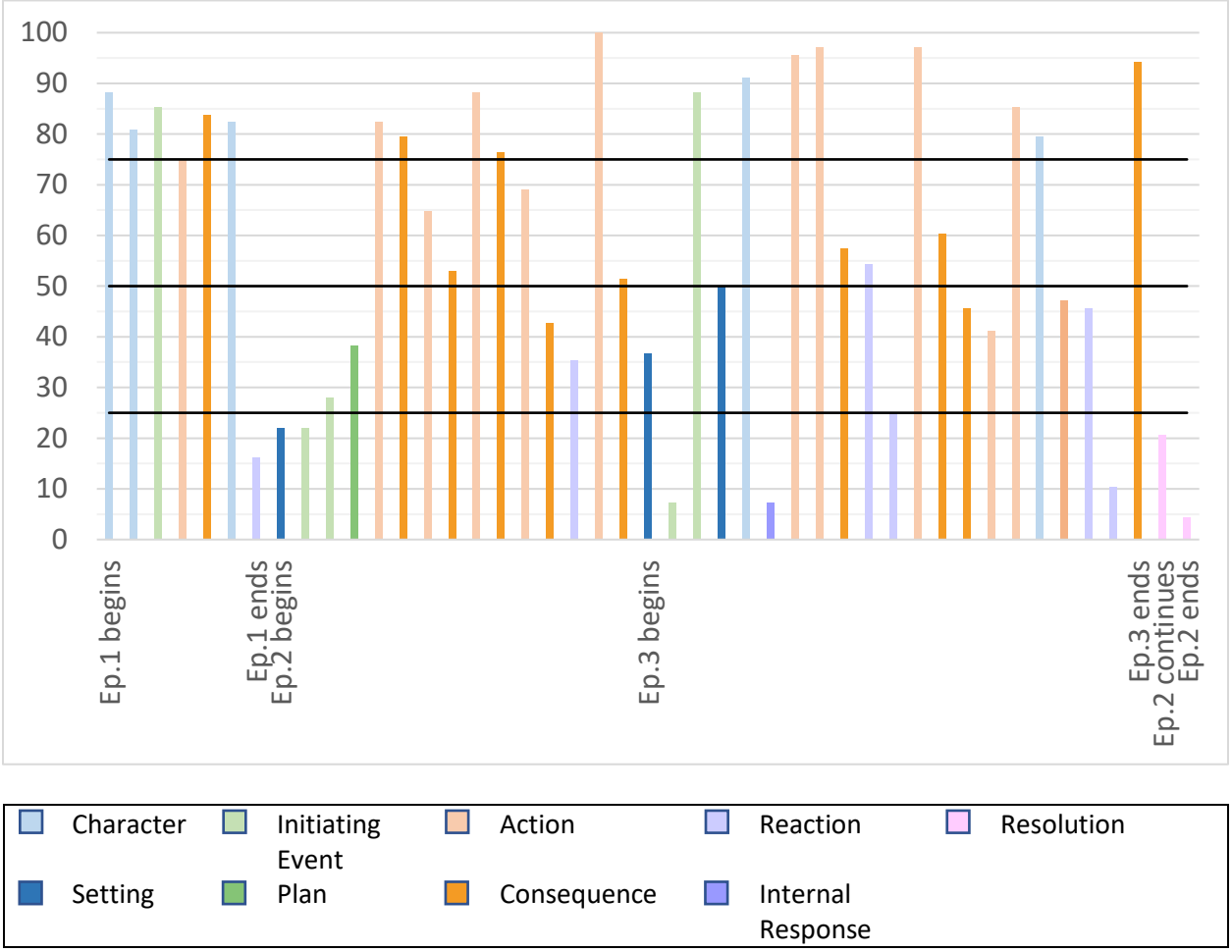
Turning now to Grade 3, results are again presented by story grammar element types across participants followed by how the participants performed on the rubric, concluding with expectations for a typical APNF story retell.

3.3.1.1 Item and Episode Difficulty

The model story for the retell task was coded into 45 story elements distributed across three episodes (see Figure 3). Once again, Episode 1, 2, and 3 differed in length (7, 17, and 21 items per episode), with the first being the shortest and the other two being similar in length. As a proxy for item difficulty, we considered how many children included each element. As is evident in Figure 3, there was considerable variability across items. Each rubric element was included by at least 3 participants (4%), and one element was included by all (100%). By dividing the distribution into quartiles, we grouped the items into three levels of difficulty: easy, average, and difficult.

Figure 3

Proportion of children who included each item in their story retell.



Nineteen items were included in the stories of 75% or more of the children (n ≥ 51) and were therefore identified as easy. Action elements (items 4, 12, 16, 21, 29, 30, 34, 38) dominated, followed by Characters (items 1, 2, 6, 27, 39), Consequences (items 5, 13, 17, 43), and Initiating Events (items 3, 25). Some Action elements are made salient by the picture book (e.g., item 21, He tries soaking in a bath), while others are based on common social exchanges, like asking a person's name (e.g., item 29, The rhinoceros asks Fluffy his name). The general pattern of Actions and Consequences (i.e., related to attempting to become fluffier but not

succeeding) is repeated throughout the second episode, which gives the child ample opportunities to retain and retell these sections. Interestingly, one of the easy Initiating Events was cognitive (item 3, the baby/child needs a name). The easy items were dispersed throughout the story.

At the other end of the distribution, 9 items were judged to be difficult, as they were included in the stories of 25% or fewer of the children ($n \leq 17$). These inconsistently produced items fell into the following categories: Reactions (items 7, 33, 42), Initiating Events (items 9, 24), Resolutions (items 44, 45), Internal Response (item 28), and Setting (item 8). All but one of these difficult items (the Setting) were cognitive elements. Cognitive Initiating Events (e.g., item 9, Fluffy doubts he is fluffy) were challenging since they required the child to describe the character's internal state, something that is not easily gleaned from the illustration. Additionally, the Resolution items (e.g., item 44 and 45, Fluffy accepts his name even if his appearance doesn't match it), once again required the child to extend beyond the immediate consequence of the episode (e.g., becoming friends) to summarize the overall consequence of the story by linking back to the overarching goal. The difficult items were more concentrated in Episodes 2 and 3.

As an indicator of episode difficulty, we considered the frequency of complete episodes (see detailed criteria in Appendix E.2). Episode 1 was most often complete (65%), followed by Episode 2 (43%), and then Episode 3 (29%). Therefore, the shorter Episode 1 seems to have been easiest, and Episode 3 most challenging. This is generally consistent with the distribution of easy and difficult items, with many cognitive elements appearing in the final episode.

In summary, the following storytelling pattern emerges from Figure 3: most children had a strong start, completing the first episode successfully, but then had difficulties transitioning between episodes and concluding their story.

3.3.1.2 Performance on the Macrostructure Rubric

Once again, there was wide variability (range 12 to 39) of total macrostructure scores for APNF (maximum 45). The mean score was 26.1 (*SD* = 6.3); hence, on average, children included 60% of the elements in their stories. To make comparisons within this sample and to establish general expectations, the APNF story retells are divided into low-scoring (19 or less; below –1 *SD*), and average-scoring (20 or more; -1 *SD* and above) based on total macrostructure score. For each group, story grammar elements (Table 9), and recombined story grammar elements (Table 10) will be considered. Further information about how ACo pairs and complete episodes were coded can be found in Appendix E.2.

Table 9

Story grammar elements, by group

Story Grammar Element	Group			
	Low (*n* = 9)		Average (*n* = 59)	
	M (*SD*)	Range	*M* (*SD*)	Range
Character (5)	3.0 (1.1)	1 - 5	4.4 (0.8)	1 - 5
Setting (3)	0.6 (0.7)	0 - 2	1.2 (0.9)	0 - 3
Initiating Event (5)	1.3 (0.7)	1 - 3	2.5 (0.7)	1 - 4
Internal Response (1)	0.1 (0.3)	0 - 1	0.1 (0.3)	0 - 1
Plan (1)	0.1 (0.3)	0 - 1	0.4 (0.5)	0 - 1
Action (11)	6.0 (2.1)	2 - 8	9.4 (1.3)	7 - 11
Consequence (11)	3.0 (1.6)	1 - 5	7.5 (2.0)	2 - 11
Reaction (6)	0.9 (0.9)	0 - 2	2.0 (1.3)	0 - 5
Resolution (2)	0.0 (0.0)	0	0.3 (0.6)	0 - 2

Table 10

Recombined story grammar elements, by group

Recombined Story Grammar Element	Group			
	Low (*n* = 9)		Average (*n* = 59)	
	M (*SD*)	Range	*M* (*SD*)	Range
Cognitive State (14)	1.7 (1.3)	0 - 4	4.4 (2.0)	2 - 10
Physical Initiating Event (1)	0.8 (0.4)	0 - 1	0.9 (0.3)	0 - 1
Action-Consequence Pairs (9)	1.2 (1.1)	0 - 3	5.3 (1.9)	1 - 9

Note. Cognitive states include all Plan, Reaction, Internal Response, and Resolutions items, and Cognitive Initiating Events. The remaining Initiating Events are physical. Action-Consequence pairs (ACo pairs) reflect the links between Actions and Consequences.

Fifty-nine participants produced average-scoring stories. Overall, 90% of these stories included at least the following elements: three Character/Setting items, one Initiating Event, seven Actions, three Consequences, and one Reaction. Furthermore, when considering regrouped elements, 95% of average-scoring stories included at least two Cognitive elements (commonly one Initiating Event plus a Plan or a Reaction) and at least three ACo pairs (Table 10). Additionally, most average-scoring stories (88%) included at least one complete episode (*M* = 1.5, *SD* = 0.8, range 0 – 3). The ten highest scoring stories within this group expanded the number of elements across types and included Setting, Plan, and Resolution elements more consistently. These stories included multiple Cognitive elements and ACo pairs (at least 5 of each) and most also had two complete episodes (*M* = 2.1, *SD* = 0.6).

Looking now at the other group, none of the nine low-scoring stories met the stated expectations of an average-scoring story. The inclusion of a Plan was rare, and Setting was

infrequent (Table 9). Overall, Cognitive elements were much reduced. Also, despite including multiple Actions and Consequences, these were often not well-linked as evident in the low number of ACo Pairs (Table 10). As an example, consider this excerpt[4]. It includes two Actions, one (item 16) that is not followed by a Consequence, the other (item 18) by an ambiguous Consequence: *"Now Gerald was taking a bubble bath and washing his hair…After the shower, he decided to put whipped cream all over his spikes and his body. He didn't like that idea."* Furthermore, less than a quarter of children included at least 2 Cognitive elements and 3 ACo pairs. In contrast, the frequency of Physical Initiating Events was similar across groups. Among the low-scoring stories, only a third of the children included one complete episode (*M* = 0.4, *SD* = 0.7, range = 0-2).

A few patterns emerged within the low scoring stories. Four of the students modified important parts of the story. It appears that the participants either decided to tell a different story or didn't understand the key plot points (e.g., wanting to become fluffier, laughing about each other's names). One child produced an attempt at a parallel storyline (about an animal named Cavey who wants to become dark and meets an animal named Prickles that is smooth). The other four followed the original storyline but omitted too many story grammar elements to create well-developed stories.

Among the students who modified portions of the story, one showed many signs of doubt in their storytelling (e.g., frequent long pauses and gentle prompting from the examiner). Another made a comment suggesting limited comprehension (e.g., "but for some reason they

[4] SALT transcription codes omitted for clarity.

became the best of friends for some reason"). Importantly, these apparent signs of low self-confidence in storytelling were not exclusive to the low-scoring stories, nor did they occur in all of them.

3.3.1.3 Summary of Expectations

Based on the results of the average-scoring group, we can formulate expectations about typical stories. To recap, a reasonable expectation for APNF retells in Grade 3 would be for the stories to include at least the following: three Character/Setting items, one Initiating Event, seven Actions, three Consequences, and one Reaction (Table 11). This benchmark was met by none of the low-scoring stories, suggesting that it could be useful to identify children with low language proficiency and possibly more informative than overall score. Alternatively, expectations could focus on the inclusion of two Cognitive Elements and three ACo pairs. Once again, the absence of a complete episode is unlikely to be a good indicator given the limited range of possible scores.

Table 11

Expected number of story grammar elements in a typical story, by type

Story Grammar Element	Number
Character or Setting	3
Initiating Event	1
Internal Response	0
Plan	0
Action	7
Consequence	3
Reaction	1
Resolution	0
Recombined Story Grammar Element	
Cognitive Elements	2
Action-Consequence pairs	3

3.3.2 Microstructure

There was considerable variability in the microstructure measures (see Table 12 for details). This is most evident when considering story length (TotNumUt), verbal fluency (WPM, %PauseTime, and %MzWds), and grammatical accuracy (%Errors). However, all story retells were generally highly intelligible and consisted of utterances with variable complexity and a variety of different words.

Considering the four main microstructure variables (MLUw, MA-NDW, %MzWds, and %Errors), 10 children obtained low scores based on two or more of these indicators being more

than 1 *SD* from the mean (below for MLUw and MA-NDW; above for %MzWds and %Errors). For those with two scores (*n* = 7), two combinations emerged: MLUw with %MzWds or MA-NDW with %Errors. For those with three (*n* = 3), MA-NDW and %Errors were most common (as had been the case for Grade 2).

Table 12

Summary of microstructure variables for APNF (*N* = 68)

Variable	*M* (*SD*)	Range
TotNumUt	49.6 (12.2)	29 – 84
MLUw	6.5 (0.9)	4.0 – 8.7
VPU	1.5 (0.2)	1.0 – 1.9
MA-NDW	55.2 (4.0)	46.8 – 62.1
WPM	73.1 (22.7)	34.7 – 135.8
%PauseTime	22.8 (14.7)	0.0 – 59.9
%MzWds	12.8 (6.5)	1.7 – 30.8
%Errors	22.4 (13.1)	3.0 – 57.9

Note. TotNumUt = total number of utterances, MLUw = mean length of utterance in words, VPU = verbs per utterance, MA-NDW = moving average number of different words, WPM = words per minute, %PauseTime = percent total pause time, %MzWds = percent total maze words, %Errors = percent utterances with errors.

3.3.3 Nonword Repetition

The mean for the NWR task was 18.3 (*SD* = 2.8) items correct (max. 24) with a wide range from 10 to 23. Ten participants scored below – 1 *SD* which corresponds to a score of 15 or lower.

3.3.4 Teacher Evaluations: ELL Matrices

ELL Matrix data were available for 88% (*n* = 60) of participants. The averages were similar for Oral Language and Reading (3.7 and 3.6) and somewhat lower for Writing (3.3), with identical ranges (1.5 to 5) (see Table 4). Therefore, the participants were generally rated between *expanding* and *consolidating* levels of English proficiency by their ELL teachers. By the end of Grade 3 Term 3, 11 students had been delisted and thereby where given ratings of 5 in all three areas. Another child who was still receiving ELL services received ratings of 5 in all skill domains. Based on two or more scores below 3 on the Matrices, 13 students were considered to have low English proficiency. Similarly to the Grade 2 data, in most cases, one of the low ratings was for Oral Language (Oral Language + Reading + Writing, *n* = 9; Oral Language + Writing, *n* = 1; Reading + Writing, *n* = 3).

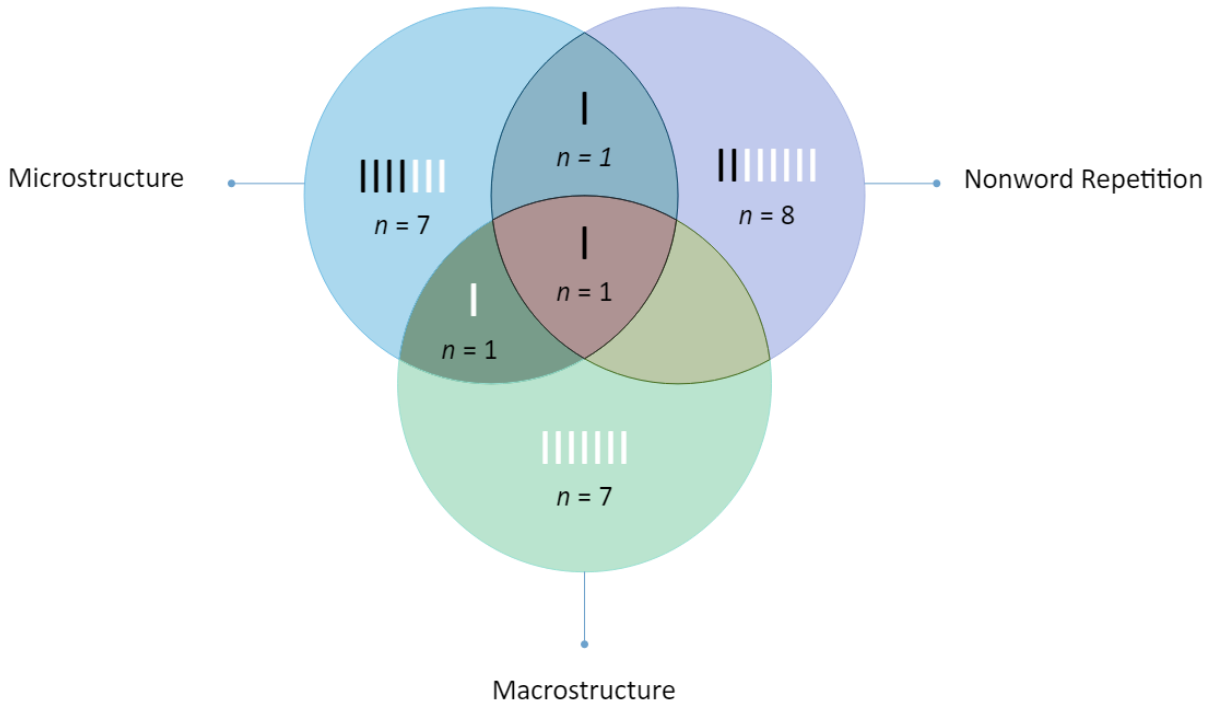
3.3.5 Low Language Proficiency

We now turn to describing the combined results for macrostructure, microstructure, and NDW in combination to identify students who did not perform as well as their peers, which

could indicate lower proficiency in English. The results from the macrostructure analysis will once again receive greater attention. Analyses will be supplemented by the data provided by ELL teachers and SLPs.

Figure 4

Participants who scored more than 1 SD from the mean for macrostructure, microstructure (two or more indicators), or NWR.



Note. Black bars identify participants who had low scores on two or more ELL Matrices.

3.3.5.1 To What Extent Do Potential Indicators of Low Language Proficiency Overlap?

In Grade 3, 25 children obtained low scores for one (*n* = 22) or more (*n* = 3) language measures (see Figure 4). The children with single low scores were generally equally distributed across macrostructure (*n* = 7), NWR (*n* = 8) and microstructure (*n* = 7). We also considered whether a child had low scores on the ELL Matrices (indicated with black font in Figure 4). As was the case for Grade 2, ELL Matrices best aligned with language measures when a child received multiple (2 of 3, 67%) compared to single (6 of 22, 27%) low scores. However, given that very few children obtained low scores on multiple measures, it is difficult to draw any conclusions for this comparison for this grade.

Centering on the macrostructure results, there is poor overlap between low ELL Matrices (1 of 9 children, 11%), and none when macrostructure was the only low language measure (0 of 7, 0%). Although ELL Matrix data is missing for two children, two other participants were delisted from ELL services. In this data set, single low scores for microstructure (4 of 7, 57%) and NWR (2 of 8, 25%) appear better aligned with low ratings for ELL Matrices, but no solid patterns emerged for any measure.

3.3.5.2 How Do the ELL Matrices Relate to the Language Measures?

Overall, 8 out of 13 children (62%) with low ELL Matrix ratings obtained low scores for one or more language measures. Most commonly (6 of 8, 75%), grammatical accuracy was low which indicates that the ELL teachers may find this to be most salient when completing their

evaluations. Interestingly, among the five students who were not captured by any of the language measures, three performed below peers for a single microstructure variable. The other two students may have had difficulties with literacy rather than oral language: one received support for language-based reading difficulties and the other obtained low ELL Matrices ratings for Reading and Writing. Hence there was generally good overlap between the various language measures and Oral Language ELL Matrix ratings.

3.3.5.3 How Does SLP Report Relate to the Language Measures and the ELL Matrices?

A total of eight children were identified to have speech or language concerns that go beyond typical second language acquisition by SLPs. Among the five children identified with language difficulties, only three were captured by our language measure (macrostructure, microstructure, and NWR in combination; NWR only; microstructure only). Another child, performed below peers for only one microstructure variable (grammatical accuracy), hence, they are not included in Figure 4. In fact, all four of these students had poor grammatical accuracy.

The child who was not picked up by our measures yet had identified language difficulties based on SLP report produced a strong story retell of APNF in terms of macrostructure (score = 32/45). However, the story had low-average MLUw and low-average grammatical accuracy as well as poor lexical accuracy—which was not one of the included measures. This child also made comments (e.g., “I don’t know”) suggesting that they may not have understood parts of the story. Finally, this child was receiving language-based reading supports.

Among the three students with speech-related concerns, one child stood out due to low scores for both macrostructure and microstructure (MA-NDW, %MzWds, %Errors) pointing towards communication difficulties beyond speech. This observation was also supported by relistening to the child's story retell.

3.4 Additional Analyses

The next section is divided into two parts: 1) exploring whether length of exposure to English could have contributed to differences in macrostructure scores (i.e., a possible confound) and 2) examining how the performance of the sample of ELL children compares to that of monolingual peers for microstructure and NWR.

3.4.1 Years of Exposure and Total Macrostructure Score

There were no significant relationships between Years of Exposure to English and Total Macrostructure scores in Grade 2 or Grade 3. This was confirmed by visual inspection of scatter plots and Pearson correlations (Grade 2: r = .145, p = .215; Grade 3: r = -.041, p = .742).

3.4.2 Monolingual Comparisons

We chose to compare our sample's results to monolingual data to gauge how they would fare given these expectations and to explore any potential for misdiagnosis. The bilingual group

was compared to monolingual data for microstructure and NWR. As macrostructure was analyzed using a novel tool, no monolingual data were available for comparison.

3.4.2.1 Microstructure

The difference between our results and the monolingual data was calculated using Cohen's *d* to obtain effect sizes for Grade 2 (Table 13) and Grade 3 (Table 14). The monolingual story retells used for comparison came from the SALT databases and were matched based on grade and the mean ages of the ELL sample for each story (Miller & Iglesias, 2020). A total of 74 PGHW retells and 39 APNF retells met the criteria. Differences between the monolingual and bilingual samples indicate large effect sizes for WPM, %PauseTime, and %Errors for Grades 2 and 3 (Table 13, 14). A large effect size for MLUw is also evident only for Grade 3. Importantly, ELLS in Grade 2 and 3 performed similarly to their monolingual peers in terms of story length (e.g., TotNumUt), lexical diversity (MA-NDW) in addition to select measures of syntactic complexity (e.g., VPU) and verbal fluency (e.g., %MzWds).

Table 13

Performance of monolingual (age and grade matched) and bilingual children for microstructure variables obtained from PGHW story retells.

Variable	Monolingual		Bilingual		Effect size
	M (*SD*)	Range	*M* (*SD*)	Range	*d*
TotNumUt	40.7 (11.5)	24 - 75	38.2 (9.2)	17 - 62	-0.24
MLUw	8.3 (0.9)	6.1 - 10.6	8.1 (1.3)	4.5 - 11.2	-0.18
VPU	1.7 (0.2)	1.1 - 2.1	1.7 (0.3)	0.8 - 2.4	0.00
MA-NDW	56.8 (4.4)	41 - 65	56.6 (4.1)	46.9 - 64.8	-0.05
WPM	91.7 (22.9)	40.3 - 144.3	67.8 (20.9)	12.0 - 110.0	-1.09
%PauseTime	13.9 (9.9)	0 - 36.0	23.5 (13.4)	2.9 - 76.1	0.81
%MzWds	11.7 (6.0)	3.3 - 28.8	12.1 (6.1)	0.4 - 31.7	0.07
%Errors	15.6 (8.9)	0 - 45.5	34.8 (16.6)	5.3 - 82.4	1.44

Note. Effect sizes (*d*) were calculated between the bilingual (*n* =75) and monolingual (*n* = 74) groups. TotNumUt = total number of utterances, MLUw = mean length of utterance in words, VPU = verbs per utterance, MA-NDW = moving average number of different words, WPM = words per minute, %PauseTime = percent total pause time, %MzWds = percent total maze words, %Errors = percent utterances with errors.

Table 14

Performance of monolingual (age and grade matched) and bilingual children for microstructure variables obtained from APNF story retells.

Variable	Monolingual		Bilingual		Effect size
	M (*SD*)	Range	*M* (*SD*)	Range	*d*
TotNumUt	46.9 (11.0)	30 - 75	49.6 (12.2)	29 – 84	0.23
MLUw	7.4 (1.1)	5.3 - 9.4	6.5 (0.9)	4.0 – 8.7	-0.93
VPU	1.6 (0.2)	1.1 - 2.1	1.5 (0.2)	1.0 – 1.9	-0.47
MA-NDW	54.6 (2.6)	49 - 60	55.2 (4.0)	46.8 – 62.1	0.17
WPM	101.1 (27.4)	51.3 - 190.5	73.1 (22.7)	34.7 – 135.8	-1.14
%PauseTime	11.3 (8.4)	1.1 - 32.4	22.8 (14.7)	0.0 – 59.9	0.90
%MzWds	13.0 (6.1)	0.5 - 28.9	12.8 (6.5)	1.7 – 30.8	-0.03
%Errors	11.6 (7.2)	0 - 29.4	22.4 (13.1)	3.0 – 57.9	0.95

Note. Effect sizes (*d*) were calculated between the bilingual (*n* = 68) and monolingual (*n* = 39) groups. TotNumUt = total number of utterances, MLUw = mean length of utterance in words, VPU = verbs per utterance, MA-NDW = moving average number of different words, WPM = words per minute, %PauseTime = percent total pause time, %MzWds = percent total maze words, %Errors = percent utterances with errors.

3.4.2.2 Nonword Repetition

In Grade 2, the average PCCs for the ELL sample (*M* = 83.4; *SD* = 9.9) for the Dollaghan and Campbell NWR task was higher than the average (*M* = 72.3, *SD* = 13.4) for a monolingual TD sample of children (aged 4;0 to 6;3 years) reported by Vuolo and Goffman (2020). The children in this sample are older, and thus it is to be expected that there would be some age-related differences, making it difficult to draw strong conclusions.

In Grade 3, the average for the ELL sample was 18.3 (*SD* = 2.8) items correct for the NWR subtest of the TILLS. Based on the norming sample for the test, a standard score of 7 or less corresponds to more than 1 *SD* below the mean (Nelson et al., 2016) and the average standard score of a monolingual child (*M* age 10;7 months) with a language and literacy disorder is 7.05 (*SD* = 4.22). Using a standard score of 7 or less, 14 ELL children would fall below this cutoff, the 10 who were identified by comparing to their ELL peers (within sample) and an additional four children. Hence, 79% (54 of 68) of the students would have obtained scores within the normal range.

Chapter 4: Discussion

The first aim of this study is to describe the range of abilities of the bilingual students in Grade 2 and 3. Our primary focus is on the potential clinical utility of narrative macrostructure. By understanding typical performance, we can extrapolate to describe expectations which could potentially serve to identify children with language disorder. We can also consider how this group of students fares compared to their monolingual peers.

In this chapter, we first summarize the similarities and differences in difficulty levels across the PGHW and APNF macrostructure rubrics, and then turn to expectations for a typical story in each Grade. We then discuss how children in this study align with the research on typical narrative development and identify story grammar elements that may aid in the diagnostic process. We discuss the extent to which our results are consistent with ELL teacher evaluations and SLP reports and the possibility of misdiagnosis. Next, the relationships between macrostructure and length of exposure to English and the results of the monolingual comparisons are examined. Finally, we conclude on the utility of narrative macrostructure for the purpose of identifying children with a language disorder.

4.1 Rubric Difficulty

PGHW and APNF are multi-episode stories that each provides children with multiple opportunities to hear and retell a complete episode. There was considerable variability in the apparent difficulty of rubric items and in the range of scores for both PGHW and APNF.

The items that were relatively easy for children to include in their PGHW and APNF retells largely belonged to the categories of Character, Initiating Event, Action, and, more specific to

APNF, Consequences. The easy items were likely included by most children due to their saliency in the picture book, repetition of names or patterns, and their focus on physical states. There was also overlap in what made items more challenging among the two stories, with the most difficult items being those that made reference to internal states or that related to goal-directed behaviour and motivations (i.e., Cognitive Initiating Event, Internal Response, Reaction, and Resolution).

A few caveats should be noted. While each story was divided into 45 story grammar elements, the numbers for each type of element differed across stories. For example, APNF has more Cognitive elements. Additionally, the motivation and intentionality evident in Cognitive elements in APNF (e.g., Reactions), was also required for difficult non-cognitive elements in PGHW (e.g., Actions). Therefore, we conclude that challenging elements for students in both grades correspond to: (1) cognitive elements and (2) elements relating goal-directed behaviour. Nonetheless, the average-scoring stories show emerging or developing use of these elements.

This is consistent with prior research where younger children (4- and 5-year-olds) rarely made reference to mental states, but when they did, it was in relation to what was visible in the pictures (Berman & Slobin, 1994). Therefore, the ability to make inferences to mental states not directly evident in pictures is a skill beyond preschoolers but is emerging in school-aged children (Berman & Slobin, 1994). The children in this study are 7- and 8- years old (Grade 2) and 8- to 9-years old (Grade 3), which places them towards the upper range for the developmental trends described by Berman and Slobin (1994).

4.2 Typical Narratives

Based on our sample of mostly TD ELLs, the following expectations emerged for a typical PGHW retell in Grade 2: (1) inclusion of these story grammar elements: one Character or Setting, two Initiating Events, one Plan, four Actions, and one Consequence. Additionally, we should expect (2) at least one ACo pair and at least one Cognitive element. Stories often, but not always, included at least one complete episode. Nonetheless, the typical story retell includes reference to goal-directed behaviour, as evident by ACo pairs and emerging use of Cognitive elements. The typical story retell of APNF in Grade 3 includes at least the following: (1) three Character/Setting elements, one Initiating Event, seven Actions, three Consequences, and one Reaction; as well as (2) two Cognitive elements and three ACo pairs.

Given the higher proportion of cognitive elements in APNF compared to PGHW, we can conclude that the plot in APNF is more challenging than PGHW. Despite this difference, as hypothesized, there is evidence of growth between Grade 2 and 3. Overall, we see a higher total macrostructure score between Grades 2 and 3 which means more story grammar items being included in the story retells. We also see more complete episodes in Grade 3 compared to Grade 2.

At first glance, it may appear that Plans are not mentioned as consistently in APNF as in PGHW. This is partly because PGHW has explicit goals for each of its three episodes (e.g., Pookins decides that she doesn't want to be a flower anymore; see Appendix C.1). In contrast, there is only one explicitly stated goal in APNF (Fluffy decides that he wants to become fluffier) with two of the episodes having implicit goals (e.g., Fluffy and Hippo’s goal to learn each other's

names; see Appendix C.2) which are made salient by reference to goal-directed behaviour (i.e., purposeful actions). Therefore, it's not that goal-directed behaviour isn't present as consistently in the APNF retells, but that it is represented differently in the story. In addition, there is more stability in making connections between Actions and Consequence (as evident by the number of ACo pairs) in Grade 3. This is similar to the observations of Trabasso et al. (1992) that older children can be distinguished from younger by their use of purposeful attempts. Connecting actions to goals or purposes that align with a higher-order goal leads to a more coherent tale which is indicative of a more mature narrator (Trabasso & Nickels, 1992). Additionally, the more frequent inclusion of Cognitive elements in a typical narrative for APNF compared to PGHW may also be due to age since this is a skill that continues to develop past adolescence. For instance, studies with older participants (ages 9;0 to 11;4) also suggest that Internal Responses and Reactions are infrequently included in story retells (Merritt & Liles, 1987). These children also implement Cognitive elements in different ways whether through an Initiating Event, a Plan (more common in PGHW), and/or a Reaction (more common in APNF). Thereby, there is variability in how children convey elements that are cognitive or relate to goal-directed behaviours, but there is evidence for their emergence or, perhaps even, stable use. Therefore, the results are consistent with developmental expectations: we would not expect children, even in Grade 3, to include these types of elements frequently.

In summary, the story retells produced in Grades 2 and 3 largely focus on physical events. The students can sequence multiple (not always complete) episodes into a story that involves one or more characters, a problem, and goal-directed behaviour (i.e., an explicit goal or implicitly by reference to purposeful actions). The children supplement physical elements with

reference to mental states, particularly in the form of Reactions. It should be noted that the expectations outlined above also demonstrate the minimum or low-end of expectations for a typical story. Moreover, evidence from high-scoring stories shows that these bilingual children can surpass these expectations substantially. Therefore, clinicians should set high expectations for this group of students while remaining cautious of over-interpreting the absence of late-developing story grammar elements.

4.3 Low-Scoring Stories

As expected, the low-scoring stories not only differed in the total number of story grammar elements, but also in types. Generally, low-scoring stories focused on basic story grammar elements like Initiating Event, Action, and Consequence. However, even if children included multiple relevant Actions and Consequences – they may not be well linked as indicated by the low frequency of ACo pairs. As expected, reference to cognitive states was infrequent, particularly in the form of a Plan, Internal Response, or Resolution. If we assume that some of the children who produced low-scoring stories have a language disorder, then our findings align with previous research. Findings from Boerma et al. (2016) suggest children with a language disorder have more difficulty expressing the feelings and intentions of characters compared to the basic episode structure (Boerma et al., 2016).

Similarly to Merritt and Liles (1987), differences were also evident in the number of complete episodes across grades and groups. The average-scoring stories in Grade 3 were more consistent in including at least one episode (*M* =1.5, *SD* = 0.8) compared to Grade 2 (*M* = 1.1, *SD*

= 0.7), so grade-level may also at least be a factor to consider. While low-scoring stories in Grade 2 (*M* = 0.5, *SD* = 0.5) and Grade 3 (*M*= 0.4, *SD* = 0.7) often had no complete episodes, the range of possible complete episodes is small (0 to 3) and variability exists within scoring groups and between them. Therefore, this points to the utility of analyzing the number of complete episodes in addition to total story grammar structure, but not on its own.

4.4 Identifying Developmental Language Disorder

Based on the evidence for which story grammar elements are more challenging as well as grade-level expectations, we now turn to our other research question and make some inferences about how this information can aid when differentiating between low-proficiency language skills that may be consistent with a diagnosis of DLD or typical bilingual development. The distribution of total scores for both PGHW and APNF shows the potential of using total macrostructure scores in combination with the minimum expectations of element types described for PGHW and APNF. Based on the easy items, children should make consistent reference to story grammar items that are illustrated in the stories (e.g., Initiating Events, Actions). Another potential avenue is to look into ACo pairs and Cognitive elements. For the latter, these would be more apparent in the forms of Initiating Events, Plan, or Reactions, and not Internal Responses or Resolutions. Although we don't expect maximum scores for these recombined elements, children in Grades 2 and 3 should be making *occasional* use of them in their narratives. Furthermore, while some items were generally more difficult than others, we

can conclude that a given child may include one difficult item but not another and another child may do the opposite– therefore they show variability, but also capability.

4.4.1 Are There Some Children Who We Should Be Concerned About Based on the Language Measures Who Are Not Being Picked Up by the Reports?

In Grade 2, one participant obtained low scores for macrostructure, microstructure, and NWR. Unfortunately, the audio files for this child were lost in Grade 3, so that no information for the language measures is available for that year. This child also had a history of SLP support, as per the parent report, but received average scores on the ELL Matrices for both grades. This is an example where language variables and parent data converge, but the child may have been missed by the school-based team and was not referred for SLP services.

Similarly, there is another child whose linguistic profile points to a need for monitoring. Based on parent report, this child was slower or later to start talking in their L1 and had seen an SLP for language-related concerns prior to school entry. In Grade 2, this participant did not stand out for any language measures. However, in Grade 3, the child obtained low scores for macrostructure and one microstructure variable (MA-NDW). In this case, the evidence appears stronger for language difficulties in Grade 3. However, reviewing the transcripts for each year shows similarities, with the child having significant difficulties with recalling the character names, and making comments suggesting limited comprehension (PGHW: “I don’t get this. I don’t get the bottom part. I just get the top”; APNF: “I don’t even know how to explain those”).

This child scored below average on the Writing ELL Matrix for both years, and their Matrix ratings for Oral Language and Reading decreased from Grade 2 to 3.

Finally, the results suggest that one child's communication difficulties are only partially reflected in their diagnosis. This student obtained low scores for macrostructure and microstructure for both years, but low ELL Matrix ratings only in Grade 2. This child was assessed by an SLP and received a diagnosis related to speech difficulties. Results from the language measures and review of the audio files strongly point to suspected language difficulties in addition to any past speech difficulties.

These examples demonstrate the potential underdiagnosis of ELL children, but also the utility of converging evidence, whether obtained through assessments and reports, to inform diagnosis.

4.4.2 Are the Language Measures Missing Some Children Who Have a Diagnosis?

As hypothesized, the language variables mainly aligned with SLP report for Grades 2 and 3. There were two exceptions in Grade 3. Recall that one child scored within typical limits across all measures and even scored high-average for lexical diversity. However, this child had difficulties with recalling character names and choosing the proper nouns and verbs. Some of these errors were reflected in grammatical accuracy (which was generally lower than the monolingual norms), but others were not. There is evidence to suggest that lexical errors versus lexical diversity better reflect the functional difficulties with word use for children with DLD. Charest and Skoczylas (2019) compared lexical errors and lexical diversity in narrative samples

of 6- and 7-year-old children with and without DLD. They found that the two groups differed in lexical errors but not lexical diversity. The authors suggest that children with DLD may use a comparable variety of words relative to their TD peers, but they may not be using those words to convey the target meaning. This may have been the case for some children in our sample as there were obvious difficulties in word-finding and labeling of character names despite hearing these words many times during the initial storytelling.

Children with DLD can present with various strengths and weaknesses (Bishop et al., 2017). It may be that this child's difficulties are in the realms of semantics or word-retrieval – which were not adequately captured by the language variables in this study. Additionally, this child's language difficulties have been reported in relation to reading. Challenges in oral language often lead to difficulties in reading, but there are instances where the opposite may not be true (Catts et al., 2005; Catts et al., 2002; Snowling et al., 2019; Tomblin et al., 2000). This child did obtain low scores for macrostructure and microstructure in Grade 2, which points to possible underlying oral language difficulties.

The other child obtained only one low microstructure score (%Errors) and therefore was not captured by our criteria. While grammatical accuracy was part of the picture for some of children, it was the only outstanding difference for this child.

4.5 Relationship Between ELL Matrices and Language Measures

The children who had low ratings on the ELL Matrices generally had one or more low language scores. The close relationship between low scores for macrostructure and/or NWR

and ELL Matrices was particularly prominent in Grade 2. This suggests that ELL Matrix data may be an effective way to screen for children who may have language difficulties beyond typical L2 language acquisition. Despite variability in which language measures were low, grammatical accuracy was common among many students – suggesting a relationship between ELL teacher ratings and grammatical accuracy. Further studies are needed to support on what basis ELL teachers make their judgements.

4.6 Relationship between SLP Report and Language Measures

The children identified by SLPs to have a language disorder had various linguistic profiles. This is not surprising given that children with DLD are a heterogenous group whereby each child with the same diagnosis can have various strengths and weaknesses (Bishop et al., 2017). Specific to narratives, children with DLD may exhibit trade-offs between clear plot development (macrostructure) and grammatical accuracy (microstructure) (Colozzo et al., 2011). A common thread among these children was that they often scored low on microstructure but there was variability regarding the specific microstructure variable. Nonetheless, focusing on the children with a language disorder only, 4 out of 5 children had high error rates in Grades 2 or Grade 3 (three of those children scored low each year). Does that indicate a potential usefulness for grammatical accuracy? Yes and No. Recall that the bilingual children in this sample, on average, had higher error rates than their monolingual peers. Additionally, there were a number of children who had a high error rate but did not have a diagnosis or low ELL Matrix ratings. In the case that grammatical accuracy is useful for this group, a more rigid cutoff may have to be set.

However, not all children with a language disorder had a high error rate, and thus we may have to look at the types of errors that are being made. Since we do not have a confirmatory sample, we cannot conclude to what extent grammatical accuracy is or isn't diagnostically useful.

In addition, the children in this study speak Chinese and English – two typologically distinct languages. While data from monolingual Chinese students indicate grammatical accuracy is not useful for diagnosing DLD (Hao et al., 2018), the data for monolingual English-speaking children suggests the opposite (Winters et al., 2022). The unit of measurement should also be kept in mind. In our analysis, grammatical accuracy is calculated not only on based on over-generalization and subject-verb agreement, but also other word-level (e.g., incorrect pronoun use, incorrect word choices) and utterance-level (e.g., word order) errors. Research on the characteristics of Chinese-English bilingual children with and without a language disorder is sparse, but there is some evidence to suggest that certain grammatical markers may be more informative compared to others (Sheng, Yang, et al., 2023). Nonetheless, we can conclude that comparing bilingual children to monolingual norms, in terms of grammatical accuracy, could lead to misdiagnosis. This echoes cautions expressed in prior research (Paradis et al., 2013). Similar to the ELL teacher ratings, grammatical accuracy may be more salient for the SLPs to base their diagnostic decisions since common language assessments focus on syntax and morphology. Moreover, findings from the monolingual English research point towards grammatical accuracy being a strong indicator of DLD. The tendency of ELL teachers and SLPs to attribute particular importance to grammatical accuracy would require further investigation.

4.7 Considering Length of Exposure to English

The children in this sample form a heterogenous group as marked by variability in length of exposure to English. In Grade 2 and 3, total scores on the macrostructure rubric did not correlate with years of exposure to English. The results of this study are in support of macrostructure being a linguistic skill that is not specific to one language (Paradis, 2016; Squires et al., 2014). Current research has mixed and limited results since prior studies have usually focused on a length of exposure that is either shorter or with younger children in general. Govindarajan and Paradis (2019) concluded that length of L2 exposure predicted better narrative abilities for TD children, but not children with DLD. In contrast, Bohnacker et al. (2022) found length of exposure did not correlate with story grammar scores for typically developing children. As Bohnacker et al. (2022) point out, this may be due to differences in the length of exposure. The children in Govindarajan and Paradis (2019) study had a mean of 24 months of L2 exposure while the children in Bohnacker et al. (2022) study had 14 to 94 months. The children in this sample have between 19 to 84 months of exposure for Grade 2 and 40 to 95 months for Grade 3. Interestingly, in Grade 2, the child with the lowest length of exposure (approximately 21 months) obtained low scores for macrostructure, microstructure (MA-NDW, %MzWds, and %Errors), NWR and ELL Matrices, but there was no parent reported language delay in L1 and no referral for SLP service. Unfortunately, this child was not part of our Grade 3 sample, but this finding has two possible interpretations. One, the converging language variables correctly indicate an unidentified language disorder, thereby showing the utility of macrostructure despite the child still attaining conversational proficiency in English. Two, the child does not have a language disorder and simply requires more time to learn English. Further

evidence with children with variable lengths of exposure is needed to have greater confidence in either explanation.

4.8 Monolingual Comparisons

Comparing the means of the ELL samples in this study to monolingual norms in Grades 2 and 3 further supported the idea that linguistic development is staggered and that bilingual children reach monolingual norms at different times (Paradis et al., 2013). By Grade 2, the bilingual children scored similarly to the monolingual database for story length (e.g., TotNumUt), syntax (e.g., MLUw, VPU), semantics (e.g., MA-NDW), and verbal fluency (%MzWds). This result was almost identical for Grade 3, with the exception of MLUw. This may be due to increased academic expectations from Grade 2 to 3 that bilingual children have not had enough time to meet. Furthermore, monolingual norms are a moving target since monolingual children continue to increase their own knowledge (Cummins, 2000). However, this can only partly explain this result since the average MLUw for both the monolingual and bilingual groups in Grade 3 was lower than Grade 2, suggesting a story effect.

Additionally, the story retell for APNF is more challenging, and therefore there may be a trade-off between story complexity and utterance length. Grammatical accuracy (%Errors), and verbal fluency (WPM, %PauseTime) were statistically different between the monolingual and bilingual groups for each year. Since grammatical accuracy is a well-established clinical marker of DLD for monolingual children who speak an Indo-European language, the application of monolingual norms to this bilingual population would misdiagnose many children with a

language disorder (Ebert, 2020). A rather surprising, and perhaps less discussed, observation, was the frequency of hesitations evident in the bilingual narratives – across both grades. Although mazing (%MzWds) was similar to the monolingual sample, other markers of verbal fluency were not. A study by Arslan et al. (2023) investigated the disfluency rates in monolingual and bilingual children during a narrative task and found that bilingual children were overall more disfluent. In particular, there was a group difference for both silent pauses and filled pauses (Arslan et al., 2023). While we did not isolate filled pauses, these would be captured in mazing. The difference in speaking rate is also supported by previous research (e.g., Michalik et al., 2018). Increased errors and hesitations in children who are still gaining academic proficiency in the LoI may be a sign of linguistic demands exceeding processing demands (Charest & Johnston, 2011).

Turning to NWR, the bulk of the children obtained scores within the normal range of a standardized assessments by Grade 3. The majority who scored below monolingual expectations also stood out with respect to their ELL peers. This points to the potential usefulness of NWR in the assessment process for bilingual children.

4.9 The Clinical Utility of Macrostructure

The participants who obtained low scores for macrostructure do not perfectly align with the five participants who had been referred to SLP services for language-related concerns. This is expected for a variety of reasons. An assessment process that includes more than one variable has proven to be more useful in diagnosis (Boerma et al., 2016). Additionally, not all

children are as likely to be referred to the SLP. It appears that the children who received a diagnosis did not all do poorly on macrostructure and so these children may have been referred for structural deficits. Nonetheless, considering that 3 out of 5 children diagnosed with a language disorder scored low for macrostructure (in combination or on its own) in Grade 2 supports the use of macrostructure in the diagnostic process.

In Grade 3 only one of the children with an identified language disorder scored low on macrostructure and there was generally little overlap between macrostructure and the ELL Matrices. However, how or why children scored low for macrostructure is important to consider. As an example, consider that in Grade 3, a few children told stories with modified or parallel story structures. If these children created novel stories because of choice, they would score low on the macrostructure total, but they would have done better on a measure that is not story specific. Therefore, in this scenario, these children may have good knowledge of story grammar. However, if, instead, these children created novel stories because they did not understand the story, this would indicate poor comprehension and/or ability to apply story grammar. Unfortunately, we are unable to disentangle between these two scenarios. Regardless, including children who may have good knowledge of story grammar could have shifted the distribution and therefore lowered the cut-off score for low versus average stories. For example, two children identified with a language disorder received scores of 21 and 23, which is only slightly above the cut-off of 19.

The next logical question then is – can we rely on macrostructure to identify children who may have a language disorder? Data from Grade 2 was more supportive, whereas no clear patterns for any measure emerged for Grade 3. Overall, our results generally support the utility

of using multiple sources of converging information (Boerma et al., 2016), although confirmatory data is necessary. Finally, we identified low performance on macrostructure based on total scores. Yet, our detailed analyses of how students constructed their stories indicates that it may be more fruitful to consider the combination of expected elements for each grade (and story), or the combination of ACo pairs and Cognitive elements.

4.10 Clinical Implications

The results of this study support the utility of using multiple sources of converging information when assessing bilingual children. The importance of local norms has been highlighted in the literature and to the author's knowledge, this study is the first to provide local norms for Grades 2 and 3 Chinese-English bilinguals in Canada. In addition, this study also expanded on the tools available to assess narrative macrostructure in school-aged children.

The macrostructure rubrics for APNF and PGHW were newly created for this project. At least 4% of children produced each story grammar element for APNF and 12% for PGHW indicating that each point was attainable. Additionally, inter-rater reliability for the rubrics was almost perfect, which is promising for further clinical use. For each grade, scores were widely dispersed, meaning that there were no floor or ceiling effects. This is important in a diagnostic context since variability is needed to pinpoint below average, average, and above average performance. The story retell format was also useful for this population as some children showed a lot of hesitations or spoke quickly, which required multiple replays to understand despite the transcribers' knowledge of the target stories. Moreover, the children used different

names or characteristics of the characters which would have been very difficult to comprehend without a model story. Scoring story retells with the macrostructure rubrics also only took 15 to 20 minutes, which is an efficient way to gain knowledge about a child's narrative abilities. Future work could consider analyzing the difference between scoring stories using an audio recording only instead of relying on transcriptions. This would ideally save time on the assessment process, although losing the ability to quickly obtain microstructure indicators would be a trade-off.

4.11 Limitations

Recall that the first aim of the present study was to describe the range of performance of Chinese-English bilinguals in Grades 2 and 3 in order to establish the range of *normal* on narrative microstructure; grade-level norms could then provide a basis of comparison to identify children who fall below the expected range. Given this, the recruited sample consisted mostly of typical ELL children. While the study includes some children who did eventually receive a DLD diagnosis from a SLP, this group of children does not constitute a confirmatory sample. It is unclear what assessments were used in making these clinical decisions and the process for referral which means there could be considerable variability in how and why children were or weren't assessed. While this study helps to identify the variation in typical performance on a variety of language tasks, the inclusion of a confirmatory sample consisting of two large groups of children deemed to be either TD or to have a DLD would be needed to validate some of the results from the current study.

Some of the low-scoring stories, particularly in Grade 3, modified the details of the retell to a point that it became unclear whether the child understood the original story. While choosing a story retell elicitation method was ideal for this population, an additional comprehension task would be a welcome addition to future research. This would allow the examiner to distinguish between a child who does not understand the main concepts of the story or doesn't know how to narrate a story successfully, from a child who chooses to modify the original story. Comprehension tasks supplement story retells by providing a more complete representation of the child's abilities (Merritt & Liles, 1987).

It may also be the case that story modifications were influenced by expressive language difficulties – though these may not be due to macrostructure difficulties but instead, microstructure. As much as the narrative macrostructure rubric was designed to avoid being influenced by specific word choices, there is a limit to how much the coders can infer from incomplete or ambiguous utterances.

There was also missing data for certain tasks or reports which led to an incomplete picture of the child's abilities.

4.12 Future Directions

There are a few different ways this research can be expanded and validated. A sample of monolingual children could be recruited in order to determine whether differences (if any) between monolingual and bilingual children on macrostructure are due to language background or disorder. In addition, by including bilingual children with previously identified

language concerns we can provide better comparisons between typical and low language proficiency specific to this group. The macrostructure rubric was relatively quick and easy to use, and there was high inter-rater reliability between coders. However, we need to validate its clinical applicability with a large group of Speech-Language Pathologists.

Chapter 5: Conclusion

In Grade 2 and 3, the children proved to be a very diverse group – both in background characteristics and language abilities. In terms of linguistic skills, there was a range of scores for macrostructure, microstructure, and NWR. Generally, children told stories that align with expected narrative retell skills for their age. The overall impression was one of ability.

Considering the differences between low- and average-scoring stories, we suggest that clinicians not only consider total scores for macrostructure, but also the inclusion of expected story-specific story grammar elements, Cognitive elements, and Action-Consequence pairs. Macrostructure scores were not related to length of English exposure indicating that this measure is not specific to knowledge of English. While macrostructure appears to have clinical utility to pick up on low language proficiency, results tentatively suggest that it is best paired with other language measures or reports to increase confidence in accurately identifying children with a possible language disorder. Furthermore, the addition of a comprehension task could serve to differentiate between children who do not understand the story from those who choose to tell a novel story.

This study has also furthered the discussion on possible misdiagnosis of Chinese-English ELLs. Some children who obtained low scores for macrostructure, in addition to other indicators, would have benefited from additional monitoring. Similarly, as a group, Chinese-English bilingual children had decreased verbal fluency and grammatical accuracy compared to monolingual age- and grade-matched peers. This adds to our knowledge that comparison to monolingual norms may overidentify bilingual children with a language disorder. Furthermore,

there may be utility in using grammatical accuracy measures in this group of children, but the appropriate reference and perhaps types of errors should be considered.

References

Abutbul-Oz, H., & Armon-Lotem, S. (2022). Parent Questionnaires in Screening for Developmental Language Disorder Among Bilingual Children in Speech and Language Clinics. *Frontiers in education (Lausanne)*, *7*. https://doi.org/10.3389/feduc.2022.846111

Altman, C., Armon-Lotem, S., Fichman, S., & Walters, J. (2016). Macrostructure, microstructure, and mental state terms in the narratives of English–Hebrew bilingual preschool children with and without specific language impairment. *Applied psycholinguistics*, *37*(1), 165-193. https://doi.org/10.1017/S0142716415000466

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). American Psychiatric Publishing.

Arslan, B., Aktan-Erciyes, A., & Göksun, T. (2023). Multimodal language in bilingual and monolingual children: Gesture production and speech disfluency. *Bilingualism (Cambridge, England)*, *26*(5), 971-983. https://doi.org/10.1017/S1366728923000196

Bamberg, M., & 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. https://doi.org/10.1017/S0305000900011314

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: a crosslinguistic developmental study* (1 ed.). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203773512

Bishop, D. V. M. (2020). Developmental Language Disorder : The Term Is Not Confined to Monolingual Children. *Perspectives of the ASHA special interest groups*, *5*(3), 572-572. https://doi.org/10.1044/2020_PERSP-20-00061

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., Adams, C., Archibald, L., Baird, G., Bauer, A., Bellair, J., Boyle, C., Brownlie, E., Carter, G., Clark, B., Clegg, J., Cohen, N., Conti-Ramsden, G., Dockrell, J., Dunn, J., Ebbels, S., . . . and the, C.-c. (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of child psychology and psychiatry*, *58*(10), 1068-1080. https://doi.org/10.1111/jcpp.12721

Boerma, T., & Blom, E. (2017). Assessment of bilingual children: What if testing both languages is not possible? *Journal of communication disorders*, *66*, 65-76. https://doi.org/10.1016/j.jcomdis.2017.04.001

Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: implications for clinical practice. *International journal of language & communication disorders*, *51*(6), 626-638. https://doi.org/10.1111/1460-6984.12234

Bohnacker, U., Lindgren, J., & Öztekin, B. (2022). Storytelling in bilingual Turkish-Swedish children: Effects of language, age and exposure on narrative macrostructure. *Linguistic approaches to bilingualism*, *12*(4), 413-445. https://doi.org/10.1075/lab.20057.boh

Bonifacci, P., Atti, E., Casamenti, M., Piani, B., Porrelli, M., & Mari, R. (2020). Which Measures Better Discriminate Language Minority Bilingual Children With and Without Developmental Language Disorder? A Study Testing a Combined Protocol of First and Second Language Assessment. *Journal of speech, language, and hearing research*, *63*(6), 1898-1915. https://doi.org/10.1044/2020_JSLHR-19-00100

Botting, N. (2020). Language, literacy and cognitive skills of young adults with developmental language disorder (DLD). *International journal of language & communication disorders*, *55*(2), 255-265. https://doi.org/10.1111/1460-6984.12518

British Columbia Ministry of Education. (2017). *English Language Learning (ELL) Standards*. https://www2.gov.bc.ca/assets/gov/education/kindergarten-to-grade-12/teach/pdfs/ell/ell-standards-full.pdf

Catts, H. W., Adlof, S. M., Hogan, T. P., & Weismer, S. E. (2005). Are Specific Language Impairment and Dyslexia Distinct Disorders? *Journal of speech, language, and hearing research*, *48*(6), 1378-1396. https://doi.org/10.1044/1092-4388(2005/096)

Catts, H. W., Fey, M. E., Tomblin, J. B., & Zhang, X. (2002). A Longitudinal Investigation of Reading Outcomes in Children With Language Impairments. *Journal of speech, language, and hearing research*, *45*(6), 1142-1157. https://doi.org/10.1044/1092-4388(2002/093)

Charest, M., Borger, P., Chan, C., Sanders, K., Yip, B., McFarlane, L.-A., & Schneider, P. (2019). Assessment, Diagnosis, and Recovery From Language Disorder at Kindergarten Age: Research Review and Clinical Discussion. *Canadian journal of speech-language pathology and audiology*, *43*(1), 49. https://go.exlibris.link/DCGQVkYB

Charest, M., & Johnston, J. R. (2011). Processing Load in Children's Language Production: A Clinically Oriented Review of Research. *Canadian journal of speech-language pathology and audiology*, *35*(1), 18-31. https://go.exlibris.link/gPPd5HkK

Charest, M., & Skoczylas, M. J. (2019). Lexical Diversity Versus Lexical Error in the Language Transcripts of Children With Developmental Language Disorder: Different Conclusions About Lexical Ability. *American journal of speech-language pathology*, *28*(3), 1275-1282. https://doi.org/10.1044/2019_AJSLP-18-0143

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates. https://doi.org/10.4324/9780203771587

Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of speech, language, and hearing research*, *54*(6), 1609-1627. https://doi.org/10.1044/1092-4388(2011/10-0247)

Colozzo, P., & Whitely, C. (2014). Keeping track of characters: Factors affecting referential adequacy in children's narratives. *First language*, *34*(2), 155-177. https://doi.org/10.1177/0142723714522164

Cummins, J. (2000). *Language, Power and Pedagogy: Bilingual Children in the Crossfire* (Vol. 23). Multilingual Matters. https://doi.org/10.21832/9781853596773

Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of Early Language Delay: I. Predicting Persistent and Transient Language Difficulties at 3 and 4 Years. *Journal of speech, language, and hearing research*, *46*(3), 544-560. https://doi.org/10.1044/1092-4388(2003/044)

Dollaghan, C., & Campbell, T. F. (1998). Nonword Repetition and Child Language Impairment. *Journal of speech, language, and hearing research*, *41*(5), 1136-1146. https://doi.org/10.1044/jslhr.4105.1136

Ebert, K. D. (2020). Language Sample Analysis With Bilingual Children: Translating Research to Practice. *Topics in language disorders*, *40*(2), 182-201. https://doi.org/10.1097/TLD.0000000000000209

Fichman, S., Altman, C., Voloskovich, A., Armon-Lotem, S., & Walters, J. (2017). Story grammar elements and causal relations in the narratives of Russian-Hebrew bilingual children with SLI and typical language development. *Journal of communication disorders*, *69*, 72-93. https://doi.org/10.1016/j.jcomdis.2017.08.001

Gagarina, N. j. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2012). MAIN: multilingual assessment instrument for narratives. *ZAS Papers in Linguistics*, *56*, 155. https://doi.org/10.21248/zaspil.56.2019.414

Genesee, F., Paradis, J., & Crago, M. B. (2021). *Dual language development & disorders: a handbook on bilingualism and second language learning* (Third ed.). Paul H. Brookes Publishing Co.

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., & Segura, H. (2017). Monitoring Indicators of Scholarly Language: A Progress-Monitoring Instrument for Measuring Narrative Discourse Skills. *Communication disorders quarterly*, *38*(2), 96-106. https://doi.org/10.1177/1525740116651442

Government of British Columbia. (2023). *Education by the numbers*. https://news.gov.bc.ca/factsheets/education-by-the-numbers

Govindarajan, K., & Paradis, J. (2019). Narrative abilities of bilingual children with and without Developmental Language Disorder (SLI): Differentiation and the role of age and input factors. *Journal of communication disorders*, *77*, 1-16. https://doi.org/10.1016/j.jcomdis.2018.10.001

Govindarajan, K., & Paradis, J. (2022). Narrative macrostructure and microstructure profiles of bilingual children with autism spectrum disorder: differentiation from bilingual children with developmental language disorder and typical development. *Applied psycholinguistics*, *43*(6), 1359-1390. https://doi.org/10.1017/S0142716422000406

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. https://doi.org/10.1007/s12187-013-9230-6

Hao, Y., Bedore, L. M., Sheng, L., & Peña, E. D. (2019). Narrative skills in two languages of Mandarin-English bilingual children. *International journal of speech language pathology*, *21*(4), 325-335. https://doi.org/10.1080/17549507.2018.1444092

Hao, Y., Sheng, L., Zhang, Y., Jiang, F., Villiers, J. d., Lee, W., & Liu, X. L. (2018). A Narrative Evaluation of Mandarin-Speaking Children with Language Impairment. *Journal of speech, language, and hearing research*, *61*(2), 345-359. https://doi.org/10.1044/2017_JSLHR-L-16-0367

Heilmann, J., Miller, J. F., Nockerts, A., & Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *American journal of speech-language pathology*, *19*(2), 154-166. https://doi.org/10.1044/1058-0360(2009/08-0024)

Hudson, J. A., & Shapiro, L. R. (1991). From knowing to telling: The development of children's scripts, stories, and personal narratives. In *Developing narrative structure.* (pp. 89-136). Lawrence Erlbaum Associates, Inc.

IBM. (2023). *SPSS Statistics*. In (Version 29.1.0.1) https://www.ibm.com/spss

Iluz-Cohen, P., & Walters, J. (2012). Telling stories in two languages: Narratives of bilingual preschool children with typical and impaired language. *Bilingualism (Cambridge, England)*, *15*(1), 58-74. https://doi.org/10.1017/S1366728911000538

Kaufman, S. A., & Kaufman, L. N. (2004). *Kaufman Brief Intelligence Test - Second Edition (KBIT-2)*. Pearson.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of communication disorders*, *43*(6), 456-473. https://doi.org/10.1016/j.jcomdis.2010.02.002

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174. https://doi.org/10.2307/2529310

Lee, S. A. S., & Gorman, B. K. (2013). Nonword repetition performance and related factors in children representing four linguistic groups. *The international journal of bilingualism : cross-disciplinary, cross-linguistic studies of language behavior*, *17*(4), 479-495. https://doi.org/10.1177/1367006912438303

Lester, H. (1986). *A Porcupine Named Fluffy (L. Munsinger, Illus.)*. Houghton Mifflin Harcourt.

Lester, H. (1987). *Pookins Gets Her Way (L. Munsinger, Illus.)*. Clarion Books.

Merritt, D. D., & Liles, B. Z. (1987). Story grammar ability in children with and without language disorder: story generation, story retelling, and story comprehension. *Journal of speech and hearing research*, *30*(4), 539. https://go.exlibris.link/gVtBjjBz

Merritt, D. D., & Liles, B. Z. (1989). Narrative analysis: clinical applications of story generation and story retelling. *The Journal of speech and hearing disorders*, *54*(3), 438. https://doi.org/10.1044/jshd.5403.438

Michalik, M., Czaplewska, E., Solak, A., & Szkotak, A. (2018). Articulation and speaking rates of Polish-French bilingual children. *Health psychology report*, *6*(4), 330-338. https://doi.org/10.5114/hpr.2018.76180

Miller, J., & Iglesias, A. (2020). *Systematic Analysis of Language Transcripts (SALT)*. In (Version 20)

Nelson, N., Plante, E., Helm-Estabrooks, N., & Hotz, G. (2016). *Test of Integrated Language and Literacy Skills*. Brookes Publishing.

Paradis, J. (2005). Grammatical Morphology in Children Learning English as a Second Language: Implications of Similarities With Specific Language Impairment. *Language, speech & hearing services in schools*, *36*(3), 172-187. https://doi.org/10.1044/0161-1461(2005/019)

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. https://doi.org/10.1075/lab.1.3.01par

Paradis, J. (2016). The Development of English as a Second Language With and Without Specific Language Impairment: Clinical Implications. *Journal of speech, language, and hearing research*, *59*(1), 171-182. https://doi.org/10.1044/2015_JSLHR-L-15-0008

Paradis, J., Emmerzael, K., & Duncan, T. S. (2010). Assessment of English language learners: Using parent report on first language development. *Journal of communication disorders*, *43*(6), 474-497. https://doi.org/10.1016/j.jcomdis.2010.01.002

Paradis, J., Schneider, P., & Duncan, T. S. (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. https://doi.org/10.1044/1092-4388(2012/12-0050)

Paradis, J., Tulpar, Y., & Arppe, A. (2016). Chinese L1 children's English L2 verb morphology over time: individual variation in long-term outcomes. *Journal of child language*, *43*(3), 553-580. https://doi.org/10.1017/S0305000915000562

Peña, E. D., Bedore, L. M., & Vargas, A. G. (2023). Exploring Assumptions of the Bilingual Delay in Children With and Without Developmental Language Disorder. *Journal of speech, language, and hearing research*. https://doi.org/doi:10.1044/2023_JSLHR-23-00117

Peña, E. D. P., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B. A. G., & Bedore, L. M. (2018). *Bilingual English-Spanish Assessment (BESA)*. Brookes Publishing.

Pratt, A. S., Adams, A. M., Peña, E. D., & Bedore, L. M. (2022). Exploring the Use of Parent and Teacher Questionnaires to Screen for Language Disorders in Bilingual Children. *Topics in early childhood special education*, *42*(1), 77-90. https://doi.org/10.1177/0271121420942308

Province of British Columbia. (2024, January 22). *Legislation and Policy Glossary*. https://www2.gov.bc.ca/gov/content/education-training/k-12/administration/legislation-policy/glossary#english_language_learners_ell_students

Rezzonico, S., Chen, X., Cleave, P. L., Greenberg, J., Hipfner-Boucher, K., Johnson, C. J., Milburn, T., Pelletier, J., Weitzman, E., & Girolametto, L. (2015). Oral narratives in monolingual and bilingual preschoolers with SLI. *International journal of language & communication disorders*, *50*(6), 830-841. https://doi.org/10.1111/1460-6984.12179

SALT Software. (2016). *Coding Aids*. https://www.saltsoftware.com/resources/codeaids

SALT Software. (2020). *Database Descriptions and Protocols*. https://www.saltsoftware.com/resources/databases

Schneider, P., Dubé, R. V., & Hayward, D. (2005). *Edmonton Narrative Norms Instrument (ENNI)*. Retrieved from University of Alberta Faculty of Rehabilitation Medicine website: www.rehabresearch.ualberta.ca/enni.

Schwob, S., Edde, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using Nonword Repetition to Identify Developmental Language Disorder in Monolingual and Bilingual Children: A Systematic Review and Meta-Analysis. *Journal of speech, language, and hearing research*, *64*(9), 3578-3593. https://doi.org/10.1044/2021_JSLHR-20-00552

Sheng, L., Shi, H., Wang, D., Hao, Y., & Zheng, L. (2020). Narrative Production in Mandarin-Speaking Children: Effects of Language Ability and Elicitation Method. *Journal of speech, language, and hearing research*, *63*(3), 774-792. https://doi.org/10.1044/2019_JSLHR-19-00087

Sheng, L., Yang, M., Du, Y., Peña, E., & Bedore, L. (2023). Grammatical profiles of Mandarin-English bilingual children at risk for developmental language disorder. In W. Han & C. Brebner (Eds.), *Typical and Atypical Language Development in Cultural and Linguistic Diversity*. Taylor & Francis Group. https://doi.org/10.4324/9781003251194-10

Sheng, L., Yu, J., Su, P. L., Wang, D., Lu, T.-H., Shen, L., Hao, Y., & Lam, B. P. W. (2023). Developmental language disorder in Chinese children: A systematic review of research from 1997 to 2022. *Brain and language*, *241*, 105268-105268. https://doi.org/10.1016/j.bandl.2023.105268

Snowling, M. J., Nash, H. M., Gooch, D. C., Hayiou-Thomas, M. E., Hulme, C., Wellcome, L., & Reading Project, T. (2019). Developmental Outcomes for Children at High Risk of Dyslexia and Children With Developmental Language Disorder. *Child development*, *90*(5), e548-e564. https://doi.org/10.1111/cdev.13216

Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International journal of language & communication disorders*, *49*(1), 60-74. https://doi.org/10.1111/1460-6984.12044

Statistics Canada. (2022). *While English and French are still the main languages spoken in Canada, the country's linguistic diversity continues to grow*. https://www150.statcan.gc.ca/n1/daily-quotidien/220817/dq220817a-eng.htm

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New Directions in Discourse Processing: Advances in Discourse Processing* (Vol. 2). ABLEX.

Stothard, S. E., Snowling, M. J., Bishop, D. V. M., Chipchase, B. B., & Kaplan, C. A. (1998). Language-Impaired Preschoolers: A Follow-Up Into Adolescence. *Journal of speech, language, and hearing research*, *41*(2), 407-418. https://doi.org/10.1044/jslhr.4102.407

Thordardottir, E. T., & Weismer, S. E. (2002). Content Mazes and Filled Pauses in Narrative Language Samples of Children with Specific Language Impairment. *Brain and cognition*, *48*(2-3), 587-592. https://doi.org/10.1006/brcg.2001.1422

Tomblin, J. B., Zhang, X., Buckwalter, P., & Catts, H. (2000). The Association of Reading Disability, Behavioral Disorders, and Language Impairment among Second-grade Children. *Journal of child psychology and psychiatry*, *41*(4), 473-482. https://doi.org/10.1111/1469-7610.00632

Torng, P.-C., & Sah, W.-H. (2020). Narrative abilities of Mandarin-speaking children with and without specific language impairment: macrostructure and microstructure. *Clinical linguistics & phonetics*, *34*(5), 453-478. https://doi.org/10.1080/02699206.2019.1655097

Trabasso, T., & Nickels, M. (1992). The Development of Goal Plans of Action in the Narration of a Picture Story. *Discourse processes*, *15*(3), 249-275. https://doi.org/10.1080/01638539209544812

Trabasso, T., Stein, N. L., Rodkin, P. C., Park Munger, M., & Baughn, C. R. (1992). Knowledge of goals and plans in the on-line narration of events. *Cognitive development*, *7*(2), 133-170. https://doi.org/10.1016/0885-2014(92)90009-G

Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied psycholinguistics*, *37*(1), 195-216. https://doi.org/10.1017/S0142716415000478

Vancouver Board of Education. (2016). *Cultural Diversity in our schools: Home Languages and Countries of Birth, Year 2016-2017** [Internal VSB document].

Vancouver School Board. (2023). *Our District*. https://www.vsb.bc.ca/page/4775/our-district

Vuolo, J., & Goffman, L. (2020). Vowel Accuracy and Segmental Variability Differentiate Children with Developmental Language Disorder in Nonword Repetition. *Journal of speech, language, and hearing research*, *63*(12), 3945-3960. https://doi.org/10.1044/2020_JSLHR-20-00166

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

Whiteside, K. E., & Norbury, C. F. (2017). The Persistence and Functional Impact of English Language Difficulties Experienced by Children Learning English as an Additional Language and Monolingual Peers. *Journal of speech, language, and hearing research*, *60*(7), 2014-2030. https://doi.org/10.1044/2017_JSLHR-L-16-0318

Winters, K. L., Jasso, J., Pustejovsky, J. E., & Byrd, C. T. (2022). Investigating Narrative Performance in Children With Developmental Language Disorder: A Systematic Review and Meta-Analysis. *Journal of speech, language, and hearing research*, *65*(10), 3908-3929. https://doi.org/10.1044/2022_JSLHR-22-00017

Xue, J., Zhuo, J., Li, P., Liu, J., & Zhao, J. (2022). Characterizing macro- and micro-structures of narrative skills for Mandarin-speaking school-age children with specific language impairment. *Journal of communication disorders*, *96*, 1-16. https://doi.org/10.1016/j.jcomdis.2022.106199

Ziegenfusz, S., Paynter, J., Flückiger, B., & Westerveld, M. F. (2022). A systematic review of the academic achievement of primary and secondary school-aged students with developmental language disorder. In (Vol. 7, pp. 1-33). London, England: SAGE Publications.

Appendices

Appendix A: Parent Questionnaire

Parent Questionnaire: Language Use at Home

Please answer the following questions to give us a better idea of your grade 1 child's language use.

Name of grade 1 child: _____Date of birth: _____

1. What language(s) do <u>you and your family</u> use when you talk to your grade 1 child at home?

_____

   a) If you and your family use more than one language, please give a rough estimate of how much you use each language (e.g. 70% Cantonese/Mandarin, 30% English)

   _____

2. What language(s) does <u>your child</u> usually use when s/he talks to you and your family at home?

_____

3. If your child uses more than one language, please give a rough estimate of how much s/he uses each language (e.g. 60% Cantonese/Mandarin, 40% English)

_____

4. When your child first started to talk, what language(s) did s/he use? _____

5. When your child first started to talk, how well did s/he speak Cantonese/Mandarin?
   a) Later and slower than most children of the same age
   b) Just as well as most children of the same age
   c) Better than other children of the same age

6. Has your child ever received speech and language therapy? Yes / No
   a) If yes, please provide further details (e.g. Was it speech and language, speech only, language only?) _____

7. At what age did your child start to learn English? _____

8. Where did s/he start learning English?_____

9. How much English did your child speak when s/he first went to school in kindergarten in Canada?
   a) Only a few words
   b) Could put a few words together to say simple things about activities, food or TV programmes
   c) Could talk about most topics but made lots of mistakes and didn't always know the words s/he needed
   d) Knew a lot of English and used it every day

What level of education does your child's mother have?
   a) Did not finish high school
   b) Graduated from high school
   c) Graduated from vocational college
   d) Graduated from university
   e) Post graduate degree

Appendix B: Elicitation Protocol for *Pookins Gets Her Way* and *A Porcupine Named Fluffy*

Use 2 copies of the book, one with the text covered. Have the speaker seated next to you with the book (text visible) in front of you.

Show the book to the speaker, and say **"I would like to find out how you tell stories. First, I will read this story to you while you follow along. Then I'm going to ask you to tell the story using your own words."** Read the story. Make sure the speaker is looking at the book.

After reading the story, prepare the recorder. Give the speaker the copy of the book with the text covered and say, **"Now I would like you to tell the story. Notice that the words are covered up. That's because I want you to use your own words to tell the story."**

Turn to the first page with pictures and start recording. Say, **"Do the best that you can. Now you tell me the story."**

Appendix C: Macrostructure Analysis

C.1 Pookins Gets Her Way

Episode 1

Character Primary-1a: Pookins.	One of the main characters is introduced.
Setting-1a	Social context for the story.
Setting-1a1: Pookins always gets her own way.	
Setting-1a2: If Pookins did not get her way, she would misbehave.	
Setting-1a3: Pookins did whatever she wanted because nobody wanted her to misbehave.	
Setting-1b: At a later point in time	Temporal context for the story.
Initiating Event-1	Sequence of events that influence Pookins to formulate a plan.
Initiating Event-1a: Pookins goes outside.	
Initiating Event-1b: She meets a magic gnome.	
Initiating Event-1c: The gnome asks Pookins how he can help her.	
Character Primary-1b: Gnome	Second main character is introduced.
Setting-1c: The gnome has magical powers	Social context for the story.
Plan-1: Pookins wants three wishes.	Pookins' goal is to have three wishes granted. Explicit goal.
Action-1a:	Pookins attempts to meet her goal by demanding her first wish.
Action-1a1: Pookins demands cowboy boots.	
Action-1a2: The gnome invokes magical powers.	
Consequence-1a: Pookins gets cowboy boots.	Partial attainment of goal.
Action-1b: Pookins demands a queen hat.	Pookins attempts to meet her goal by demanding her second wish.
Consequence-1b: Pookins gets a queen hat.	Partial attainment of goal.
Action-1c:	Pookins attempts to meet her goal by demanding her third wish. The gnome is reluctant to fulfill Pookins' wish, but eventually completes all necessary steps to turn Pookins into a flower.
Action-1c1: Pookins asks to become a flower.	
Action-1c2: The gnome questions Pookins' request.	

Action-1c3: Pookins insists by threatening the gnome.	
Action-1c4: The gnome concedes.	
Action-1c5**:** The gnome tells Pookins it's not easy to become a flower.	
Action-1c6: The gnome puts Pookins in a pot.	
Action-1c7: The gnome dumps soil on Pookins.	
Action-1c8: The gnome waters Pookins.	
Action-1c9: The gnome invokes magical powers.	
Consequence-1c: Pookins is a growing flower.	Pookins is slowly turning into a flower. Complete attainment of goal. All three wishes are granted.

Episode 2

Initiating Event-2: Pookins is a flower.	Event that invokes feelings of regret and leads Pookins to reconsider her choices.
Plan-2: Pookins decides that she doesn't want to be a flower anymore.	Implied goal of Pookins deciding she doesn't want to be a flower anymore. Explicit goal.
Action-2a: Pookins demands to be released from the pot.	Pookins' attempts to turn back into a girl.
Consequence-2a: The gnome refuses.	Non-attainment of goal.
Reaction-2a: Pookins is sad.	Pookins feelings towards turning into a flower.

Episode 3 (embedded)

Initiating Event-3:	The rain and the gnome's worries leads to a response from Pookins.
Initiating Event-3a: It begins to rain.	
Initiating Event-3b: The gnome is worried he will lose his powers.	
Internal Response-3: Pookins feels sorry for the gnome.	Pookins' feellings cause her to formulate a plan.
Plan-3: Pookins tells the gnome she will keep him dry.	Pookins' goal is to help the gnome keep his powers. Explicit goal.
Action-3:	Pookins' attempts to help the gnome.
Action-3a: Pookins tells the gnome to get under her petals.	
Action-3b: The gnome goes under Pookins' petals.	
Consequence-3:	Attainment of goal.
Consequence-3a: The gnome is safe from the rain.	

Consequence-3b: The gnome changes his mind about helping Pookins because she helped him.	

Episode 2 continued:

Action-2b:	The gnome agrees to help Pookins' turn back into a girl.
Action-2b1: The gnome agrees to help Pookins only if she acts nicely.	
Action-2b2: Pookins agrees.	
Action-2b3: The gnome invokes magical powers.	
Consequence-2b: Pookins is no longer a flower.	Attainment of goal.
Reaction-2b: Pookins is happy.	Pookins' feelings towards attainment of goal.
Resolution-2: Pookins kept one apple - just in case.	The overarching consequence of the story is that Pookins became nicer, but still gets her own way.

C.2 A Porcupine Named Fluffy

Episode 1

Character Secondary-1: Parents	A minor character is introduced.
Character Primary-1a: Baby/child porcupine	Partial introduction of one of the main characters.
Initiating Event-1: The baby needs a name.	External event that leads to a response from the parents. Implicit goal is to find a name for their child.
Action-1: The parents consider different names for their child.	Parents' response to the external event.
Consequence-1: The parents name their child.	Consequence due to the parents' response. Full goal attainment.
Character Primary -1b: Baby/child's name is Fluffy	Complete introduction of one of the main characters.
Reaction-1: The parents like the name Fluffy.	Parents' feelings towards finding a name.

Episode 2

Setting-2: Time has passed...	Temporal context for the story.
Initiating Event-2	Internal events that leads to a response from Fluffy.
Initiating Event-2a: Fluffy doubts he is fluffy.	
Initiating Event-2b: Fluffy realizes he is not fluffy.	
Plan-2: Fluffy decides that he wants to become fluffier.	Fluffy's goal is to become fluffier. Explicit and overarching goal of the story.
Action-2a: Fluffy tries to be a cloud...	Fluffy's first attempt at becoming fluffier.
Consequence-2a: ...but he can't stay up.	Non-attainment of goal.
Action-2b: Fluffy tries to become a pillow...	Fluffy's second attempt at becoming fluffier.
Consequence-2b: ...but he is not comfortable to sit on.	Non-attainment of goal.
Action-2c: He tries soaking in a bath...	Fluffy's third attempt at becoming fluffier.
Consequence-2c: ...but he does not become fluffy.	Non-attainment of goal.
Action-2d: Fluffy tries to put whipped cream on his quills...	Fluffy's fourth attempt at becoming fluffier.
Consequence-2d: …but it didn't make Fluffy fluffy.	Non-attainment of goal.
Reaction-2d: Fluffy feels discouraged after all these unsuccessful attempts.	Fluffy's feelings due to the non-attainment of his goal.
Action-2e: Fluffy keeps on doing things to become fluffier.	Fluffy keeps on attempting to become fluffier.
Consequence-2e: ...but none of that worked.	Non-attainment of goal.

Episode 3 (embedded)

Setting-3a: At a later point in time	Temporal context for the story.
Initiating Event-3:	
Initiating Event-3a: Fluffy goes for a walk to think of ways to become fluffy.	
Initiating Event-3b: Fluffy runs into an animal.	Encounter motivates the characters to act. Implicit two-part goal to learn each other's names.
Setting-3b: The animal is mean or threatening towards Fluffy.	Provides social context for the story by describing the character.
Character Secondary-3a: Animal is a rhinoceros.	Partial introduction of one of the main characters.
Internal Response-3: Fluffy is worried about the grumpy/intimidating rhinoceros.	Fluffy's feelings in response to meeting the grumpy rhinoceros.
Action-3a	The rhinoceros' attempt to learn Fluffy's name.
Action-3a1: The rhinoceros asks Fluffy his name.	
Action-3a2: Fluffy responds.	
Consequence-3a: The rhinoceros now knows Fluffy's name.	Attainment of rhinoceros' goal.
Reaction-3a:	The rhinoceros' reaction to knowing the porcupine's name, which also leads to a response from the porcupine.
Reaction-3a1: The rhinoceros howls with laughter at Fluffy's name.	
Reaction-3a2: Fluffy was embarrassed.	
Action-3b: Fluffy asks the rhinoceros his name.	The porcupine's attempt to learn the rhinoceros' name.
Consequence-3b:	Non-attainment of the porcupine's goal.
Consequence-3b1: The rhinoceros is trying to say his name but is unable to tell his name...	
Consequence-3b2: ...because he is laughing too hard.	
Action-3c:	The porcupine's second attempt to learn the rhinoceros' name.
Action-3c1: Fluffy tries to guess the rhinoceros' name.	
Action-3c2: The rhinoceros responds that his name is Hippo.	
Character Secondary-3b: Rhinoceros' name is Hippo.	Complete introduction of one of the main characters.
Consequence-3c: Fluffy now knows Hippo's name.	Attainment of the porcupine's goal.

Reaction-3c: Fluffy laughs at Hippo's name.	The porcupine's reaction to knowing the rhinoceros' name.
Reaction-3: Fluffy and Hippo laugh together because of their (incongruous) names.	The porcupine's and the rhinoceros' responses to knowing each other's names.
Consequence-3: They became the best of friends.	The overall result of the porcupine and rhinoceros meeting.

Episode 2 (continued)

Resolution-2	There is overall non-attainment of Fluffy's original goal, but Fluffy learns to accept himself.
Resolution-2a: Fluffy accepts his name...	
Resolution-2b: ...even if his appearance doesn't match it.	

Appendix D: Macrostructure Rubrics

D.1 Pookins Gets Her Way

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
1	Character Primary-1a	Pookins	"Pookins"	Refers to the girl as "Pookins" Phonological differences are accepted.
2	Setting-1a1	Pookins always gets her own way	"Pookins was used to getting her own way"	Makes it clear that the girl gets her own way. Accept want or likes her own way.
3	Setting-1a2	If Pookins did not get her way, she would misbehave	"If Pookins did not get her own way, she would make faces, throw apples, and yell very loudly"	Mentions the girl misbehaving **if** she does not get her way. Misbehaving can be expressed in general terms or by stating 1 or more bad behaviours (e.g., making faces, etc.).
4	Setting-1a3	Pookins did whatever she wanted because nobody wanted her to misbehave	"And because nobody wanted her to make faces, Pookins got her own way. She had ice cream for breakfast. She never ate her vegetables. She did not pick up her clothes, and she got all the toys she ever asked for. She roller-skated in the living room. And she went to bed very late, sometimes even after the owls."	Mentions the girl does whatever she wants or gets her own way **because** nobody wants her to misbehave. Doing whatever she wants can be expressed in general terms or by stating 1 or more examples (e.g., eating ice cream for breakfast, etc.).
5	Setting-1b	At a later point in time	"One day..."	Indicates a later point in time or that time has passed before the girl meets someone.
6	Initiating Event-1a	Pookins goes outside	"One day Pookins went out for a skip"	Mentions the child going outside or an activity like walking, or skipping.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
7	Initiating Event-1b	She meets a magic gnome	"She met a magic gnome"	Indicates the girl meets or sees someone.
8	Initiating Event-1c	The gnome asks Pookins how he can help her	"The gnome asked Pookins how he can help her"	Mentions the creature asking how he can help the girl or what she wants.
9	Character Primary-1b	Gnome	"Gnome"	Refers to the creature as a gnome. Phonological differences are accepted. If the child uses multiple names for the same character, score based on first mention.
10	Setting-1c	The gnome has magical powers	"Magic gnome"	Mentions that the creature has magical powers. This may be in the form of a descriptor (e.g., magic) or by referring to the creature using a term that suggests magical powers (e.g., wizard, elf).
11	Plan-1	Pookins wants three wishes	"I want three wishes"	Mentions the girl wanting or demanding three things. Accept even if mentioned later in the story.
12	Action-1a1	Pookins demands cowboy boots	"I want a new pair of cowboy boots, or else I'll make faces, throw apples, and yell very loudly"	Mentions the girl demanding cowboy boots **and** threatening the creature. A general statement or dialogue are acceptable. Any way of referring to footwear is acceptable (e.g., cowboy shoes, magic boots, boots, etc.). The threat can be expressed in general terms, by stating 1 or more explicit threats (e.g., throwing apples, etc.), or when demanding three wishes.
13	Action-1a2	The gnome invokes magical powers	"The gnome rubbed his magic hat"	Mentions the creature touching his hat or using magic. This can happen with the cowboy boots or the queen hat.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
14	Consequence-1a	Pookins gets cowboy boots	"Pookins got her cowboy boots"	
15	Action-1b	Pookins demands a queen hat	"I want a beautiful queen hat or else I'll make faces, throw apples, and yell very loudly"	Mentions the girl demanding a queen hat **and** threatening the creature. A general statement or dialogue is accepted. Any way of referring to headwear is acceptable (e.g., princess hat, crown, hat, etc.). The threat can be expressed in general terms, by stating 1 or more explicit threats (e.g., throwing apples, etc.), or when demanding three wishes.
16	Consequence-1b	Pookins gets a queen hat	"Pookins got her queen hat"	
17	Action-1c1	Pookins asks to become a flower	"I want to become a flower"	Mentions the girl wanting to be a flower. A general statement or dialogue is accepted.
18	Action-1c2	The gnome questions Pookins' request	"Are you sure you want to become a flower?"	Mentions the creature questioning what the girl wants.
19	Action-1c3	Pookins insists by threatening the gnome	"If you don't let me become a flower, I'll make faces, I'll throw..."	Mentions the girl threatening to do something unpleasant (e.g., making faces) **if** the creature doesn't grant her wish.
20	Action-1c4	The gnome concedes	"Never mind"	Mentions the creature dropping the argument or simply accepting to turn her into a flower.
21	Action-1c5	The gnome tells Pookins it's not easy to become a flower	"It's not easy to become a flower"	
22	Action-1c6	The gnome puts Pookins in a pot.	"The gnome puts Pookins in a pot"	Mentions the creature putting the girl in a pot. This can be accomplished by describing the creature's action (e.g., putting the girl in a pot) or by the creature telling the steps to the girl (e.g., you need to be in a pot).

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
23	Action-1c7	The gnome dumps soil on Pookins.	"The gnome adds some dirt into the pot"	Mentions that the creature put dirt or something similar (e.g., soil) in a pot, or around the girl. This can be accomplished by describing the creature's action (e.g., putting dirt in the pot) or by the creature telling the steps to the girl (e.g., you need some dirt).
24	Action-1c8	The gnome waters Pookins.	"The gnome watered Pookins"	Mentions the creature watering or pouring water over the girl. This can be accomplished by describing the creature's action (e.g., watering the girl) or by the creature telling the steps to the girl (e.g., you need lots of water).
25	Action-1c9	The gnome invokes magical powers	"The gnome rubbed his magic hat"	Mentions the creature touching his hat or using magic.
26	Consequence-1c	Pookins is a growing flower	"Pookins stood under the sun for hours and hours"	Mentions that the girl is standing under the sun for a long time or that she is growing.
27	Initiating Event-2	Pookins is a flower.	"Pookins was a flower"	Mentions the girl being a flower.
28	Plan-2	Pookins decides that she doesn't want to be a flower anymore	"Pookins decided that getting her own way wasn't so much fun after all"	Uses a cognitive verb (e.g., decide, want) to make it clear that the girl *decides* to be turned back to herself. This could be expressed by stating that the girl *doesn't want* her own way anymore or that she *doesn't want* to be a flower.
29	Action-2a	Pookins demands to be released from the pot	"Let me out of this pot, or else..."	Mentions that the girl wants to get out of the pot or be turned back into a girl **and** threatens the creature.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
30	Consequence-2a	The gnome refuses	"You wanted your own way and I gave it to you"	Mentions that the creature refuses to turn the girl back into herself. This can also be expressed by stating that he tells the girl he gave her what she wanted or that her threats won't work (e.g., flowers can't throw apples).
31	Reaction-2a	Pookins is sad	"Pookins felt sorry for herself"	Mentions how the girl feels about being a flower. Do not credit "crying" or "tears" without reference to the girl's feelings.
32	Initiating Event-3a	It begins to rain	"It began to rain"	Mentions the occurrence of rain or a rainstorm.
33	Initiating Event-3b	The gnome is worried he will lose his powers	"My hat will shrink and I will lose my powers"	Mentions that the creature could lose his powers or that his hat would no longer be magical. This can be expressed in dialogue. No need to mention that the hat will shrink.
34	Internal Response-3	Pookins feels sorry for the gnome	"Pookins felt sorry for the gnome"	Mentions how the girl now feels towards the creature (e.g., sad, sorry for him, etc.). The reason behind her feelings does not need to be stated.
35	Plan-3	Pookins tells the gnome she will keep him dry	"I'll keep you dry."	Mentions the girl deciding to protect the creature or his powers. This can happen in dialogue where Pookins tells the creature she will keep him dry.
36	Action-3a	Pookins tells the gnome to get under her petals	"Get under my petals"	Mentions the girl telling the creature to take cover. Reference to any flower structure (e.g, flower, leaves, petals, etc.) is accepted and can occur for this or the next element (Action-4b).

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
37	Action-3b	The gnome goes under Pookins' petals	"The gnome went under Pookins' petals"	Mentions the creature hiding or taking cover. Reference to any flower structure (e.g, flower, leaves, petals, etc.) is accepted and can occur for this or the preceding element (Action-4a).
38	Consequence-3a	The gnome is safe from the rain	"The gnome stayed under Pookins' petals until the rain stopped"	Mentions that the creature is safe. This can be expressed by stating that the creature stays under cover until the rain stops, or has been protected or helped by the girl.
39	Consequence-3b	The gnome changes his mind about helping Pookins because she helped him	"You helped me so I will help you"	Mentions the creature deciding to help the girl **because** she helped him.
40	Action-2b1	The gnome agrees to help Pookins only if she acts nicely	"I will help you if you put all of your bad faces, loud yells, and apples into my magic hat forever"	Mentions the creature helping the child only **if** she stops all her bad behaviours. No longer misbehaving can be expressed in general terms or referring to 1 or more bad behaviours (e.g., making faces, etc.).
41	Action-2b2	Pookins agrees	"Pookins agreed"	Mentions the girl agreeing to behave nicely. This can be expressed by a general statement or by the girl putting things in the creature's hat.
42	Action-2b3	The gnome invokes magical powers	"The gnome rubbed his hat"	Mentions the creature touching his hat or using magic.
43	Consequence-2b	Pookins is no longer a flower	"Pookins was no longer a flower"	Mentions the girl turning back to herself or not being a flower anymore.
44	Reaction-2b	Pookins is happy	"Pookins was happy"	

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
45	Resolution-2	Pookins kept one apple - just in case	"Pookins kept one apple, just in case"	Mentions that the girl kept one apple "(just) in case". Alternatively, indicates that the girl wasn't supposed to keep an apple or that she still got her way. Do not credit "kept one apple" on its own.

D.2 A Porcupine Named Fluffy

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
1	Character Secondary-1	Parents	"Mr. and Mrs. Porcupine"	Mentions that there are two parents, **or** a mom and dad.
2	Character Primary-1a	Baby/child porcupine	"Mr. and Mrs. Porcupine had their first child"	Mentions that there is a baby **and** that it is a porcupine. This might be made clear when talking about the parents.
3	Initiating Event-1	The baby/child needs a name	"The baby needed a name"	Mentions that the parents need a name for their child, **or** do not know what to name the baby.
4	Action-1	The parents consider different names for their child	"Should they call him Spike? No Should they call him Lance? No"	Makes clear that the parents considered different names for their child. Names do not need to be exactly the same as mentioned in the story. If names are listed, two names (including the name chosen for the child) are sufficient.
5	Consequence-1	The parents name their child	"Let's call him Fluffy"	Mentions the parents naming the child. Any name is accepted.
6	Character Primary-1b	Baby/child's name is Fluffy	"Let's call him Fluffy"	Names the porcupine Fluffy. Must be clear that the name 'Fluffy' is attributed to the child porcupine (not another or unspecified animal/character). If the child uses multiple names for the same character, score based on first mention.
7	Reaction-1	The parents like the name Fluffy	"Fluffy is such a pretty name!"	Expresses the parent's feelings about the child's name **or** about having found a name.
8	Setting-2	Time has passed...	"Some time later...", "Eventually..."	Indicates passage of time between young (e.g., being named) and older porcupine (e.g., doubting he is fluffy).

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
9	Initiating Event-2a	Fluffy doubts he is fluffy	"Fluffy began to doubt that he was fluffy"	Mentions a cognitive or thinking verb to refer to state of doubt (e.g., "doubted" or "wondered").
10	Initiating Event-2b	Fluffy realizes he is not fluffy	"Fluffy stuck to the door, and poked holes in the mattress. He realized he was not fluffy."	States explicitly that the porcupine realizes he is not fluffy by using a cognitive verb. This must follow or precede mention of 1 or more unfluffy incidents (stuck to door; poked holes in the mattress; poked holes in umbrella) or a statement referring to events leading up to the realization.
11	Plan-2	Fluffy decides that he wants to become fluffier	"Fluffy decided that he wants to become fluffier"	States explicitly the **overarching** intention to become fluffier (or fluffy) using a cognitive verb (e.g., decide, try to, want to, etc.). Accept element even if mentioned later in the story as long as it refers to the general intention.
12	Action-2a	Fluffy tries to be a cloud...	"I'll be a cloud"	Mentions being a cloud **or** being near/with the clouds.
13	Consequence-2a	...but he can't stay up	"Fluffy couldn't stay up" *"But it did not work"	Indicates failure at being a cloud or being near/with the clouds (e.g., falling, not staying up, etc.). *Accept "but it did not work" or something equivalent, if it is clear that the **immediately** previous action (being a cloud) is done to become fluffier **or** Plan-2 (item 11) has been met.
14	Action-2b	Fluffy tries to become a pillow...	"I'll be a pillow"	Mentions being a pillow or being fluffy like a pillow.
15	Consequence-2b	.... but he is not comfortable to sit on	"But when his mom sat on him, she was not pleased" *"But it did not work"	Mentions that the mom sat on the porcupine **and** that she is not pleased or not comfortable. *Accept "but it did not work" **or** something equivalent, if it is clear that the **immediately** previous action (being a pillow) is done to become fluffier or Plan-2 (item 11) has been met.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
16	Action-2c	He tries soaking in a bath	"He tried soaking in a bubble bath"	Mentions the porcupine taking a bath. Action must be intentional on part of the porcupine.
17	Consequence-2c	...but he does not become fluffy	"He became soggy" *"But it did not work"	Mentions the porcupine is not successful in becoming fluffy. This failure (at becoming fluffy) can be expressed using descriptions such as being soggy, soaked, wet, or gooey. *Accept "but it did not work" or something equivalent, if it is clear that the immediately previous action (taking a bath) is done to become fluffier or Plan-2 (item 11) has been met.
18	Action-2d	Fluffy tries to put whipped cream on his quills...	"He put some whipped cream on each quill"	Mentions the porcupine putting whipped cream **on** (parts of) his body. Whipped cream, shaving cream, and cream are acceptable word choices.
19	Consequence-2d	... but it didn't make Fluffy fluffy	"But it did not make him fluffy" *"But it did not work"	Indicates the porcupine is not successful in becoming fluffy. This failure (at becoming fluffy) can be expressed using descriptions such as being gooey, stickier, etc. Accept "but it did not work", **or** something equivalent if it is clear that the **immediately** previous action (putting whipped cream) is done to become fluffier or Plan-2 (item 11) has been met.
20	Reaction-2d	Fluffy feels discouraged after all these unsuccessful attempts	"They should have named me Gooey" "Fluffy sighed"	Expresses how the porcupine feels about not becoming fluffier (e.g., discouraged, sad, upset). This can happen in self-talk (e.g., "Maybe they should call me Gooey").

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
21	Action-2e	Fluffy keeps on doing things to become fluffier	"Fluffy eats fluffy marshmallows, rolls in shaving cream and feathers, and tries to be a bunny."	Mentions other attempts at becoming fluffier such as eating fluffy marshmallows, rolling in shaving cream and feathers, and trying to be a bunny. A general statement **or** only one of the above examples is acceptable. The reason behind the attempts (to become fluffier) does not need to be included.
22	Consequence-2e	...but none of that worked	"Fluffy still wasn't fluffy" *"But nothing would work"	Mentions explicitly that the porcupine still wasn't fluffy. *Accept "but it did not work" or something equivalent if it is clear that the immediately previous action(s) (e.g., eating marshmallows) is/are done to become fluffier or Plan-2 (item 11) has been met.
23	Setting-3a	At a later point in time	"One day", "one afternoon"	Indicates a later point in time or that time has passed before the porcupine goes for a walk. Do not generally accept 'when' or 'once' on their own.
24	Initiating Event-3a	Fluffy goes for a walk to think of ways to become fluffy	"Fluffy goes for a walk, trying to think of ways to become fluffy"	Mentions that the porcupine goes for a walk **and** includes a motivation for the action (e.g., to become fluffy). The specific action can vary (e.g., goes for a walk, outside, or to the forest, etc.).
25	Initiating Event-3b	Fluffy runs into an animal	"Fluffy meets a rhinoceros"	Indicates the porcupine meets or sees an animal. No need to specify type of animal.
26	Setting-3b	The animal is mean or threatening towards Fluffy	"The rhinoceros wanted to give Fluffy a rough time"	Indicates in some way that the animal is mean or is threatening **to** the porcupine.
27	Character Secondary-3a	Animal is a rhinoceros	"Rhinoceros"	Refers to the animal as a rhinoceros. Phonological differences or *rhino* are accepted.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
28	Internal Response-3	Fluffy is worried about the grumpy/intimidating rhinoceros	Fluffy didn't know what a rough time was but he didn't like the sound of it"	Mentions how the porcupine reacts to the intimidating rhinoceros. Must mention an internal state.
29	Action-3a1	The rhinoceros asks Fluffy his name	"What's your name?"	
30	Action-3a2	Fluffy responds	"Fluffy"	Mentions that the porcupine tells the rhinoceros his name. The exact name will depend on the child's story **but** should match the name given for Character Primary-1b (item 6).
31	Consequence-3a	The rhinoceros now knows Fluffy's name	"A porcupine named Fluffy!"	Mentions that the rhinoceros acknowledges the porcupine's name. This will often happen in dialogue. The rhinoceros can either repeat the porcupine's name **or** comment that the name is silly or odd.
32	Reaction-3a1	The rhinoceros howls with laughter at Fluffy's name	"'A porcupine named Fluffy!' howled the rhinoceros"	Explicitly refers to the rhinoceros' reaction **to** the porcupine's name. Laughter can occur before **or** after the acknowledgement of the porcupine's name.
33	Reaction-3a2	Fluffy was embarrassed	"Fluffy was embarrassed"	Mentions feelings or thoughts that reflect the porcupine's reaction to the rhinoceros' laughter. Do not credit "trying to be polite" on its own.
34	Action-3b	Fluffy asks the rhinoceros his name	"What's your name?"	
35	Consequence-3b1	The rhinoceros is trying to say his name but is unable to tell his name	"I just can't say it"	Makes it clear that the rhinoceros is unable to tell his name--not unwilling. 'Trying' is accepted as it implies not being able.
36	Consequence-3b2	because he is laughing too hard	"I'm laughing so hard"	Mentions laughing as the reason why the rhinoceros is unable to say his name. Do not accept laughing sounds on their own.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
				The reason for laughing does not need to be mentioned.
37	Action-3c1	Fluffy tries to guess the rhinoceros's name	"Is it Hubert?"	Mentions the porcupine trying to guess the rhinoceros' name. No specific name is required.
38	Action-3c2	The rhinoceros responds that his name is Hippo	"Hippo"	Mentions that the rhinoceros tells the porcupine his name. The exact name will depend on the child's story.
39	Character Secondary-3b	Rhinoceros' name is Hippo	"Hippo"	Names the rhinoceros "Hippo". If the child uses multiple names for the same character, score based on first mention.
40	Consequence-3c	Fluffy now knows Hippo's name	"A rhinoceros named Hippo!"	Mentions that the porcupine acknowledges the rhinoceros' name. This will often happen in dialogue. The porcupine can either repeat the rhinoceros' name **or** comment that the name is silly or odd.
41	Reaction-3c	Fluffy laughs at Hippo's name	"'A rhinoceros named Hippo!', Fluffy giggled"	Explicitly refers to the porcupine's reaction **to** the rhinoceros' name. Laughter can occur before or after the acknowledgement of the rhinoceros' name.
42	Reaction-3	Fluffy and Hippo laugh together because of their (incongruous) names	"A porcupine named Fluffy! A rhinoceros named Hippo! They laughed together."	Mentions the porcupine and rhinoceros are laughing together **because** of their names. The fact that they are both laughing because of their names can again be expressed in dialogue, by repeating each other's names or by commenting that they both have names that are silly or odd. Laughing can be represented by a verb (laughing, giggling, howling with laughter) or by laughing sounds.

Item #	Story Grammar Element	Representation	Target	Considerations for crediting the element
43	Consequence-3	They become the best of friends	"They become best friends"	Mentions becoming friends.
44	Resolution-2a	Fluffy accepts his name...	"Fluffy didn't mind being Fluffy anymore"	Mentions the porcupine not minding either being named Fluffy **or** not being fluffy.
45	Resolution-2b	...even if his appearance doesn't match it	"Fluffy didn't mind being Fluffy anymore even though he wasn't fluffy"	Acknowledges the mismatch between the porcupine's name and appearance.

Appendix E: Identifying Complete Episodes and Action-Consequence Pairs

E.1 Pookins Gets Her Way

Episode 1:

IE (one of):

- Pookins meets a magic gnome (Initiating Event-1b).
- The gnome asks Pookins how he can help her (Initiating Event-1c).

A/Co (one pair among):

- Pookins demands cowboy boots/Pookins gets cowboy boots (Action-1a1/Consequence-1a).
- Pookins demands a queen hat/Pookins gets a queen hat (Action-1b/Consequence-1b).
- Pookins asks to become a flower/Pookins is a growing flower (Action-1c1/Consequence-1c).

Episode 2:

IE: Pookins is a flower (Initiating Event-2).

A: Pookins demands to be released from the pot (Action-2a).

Co (one of):

- The gnome refuses (Consequence-2a).
- Pookins is no longer a flower (Consequence-2b).

Episode 3:

IE (one of):

- It begins to rain (Initiating Event-3a).
- The gnome is worried he will lose his powers (Initiating Event-3b).

A: The gnome goes under Pookins' petals (Action-3b).

Co: The gnome is safe from the rain (Consequence-3a).

ACo Pairs:

1. Pookins demands cowboy boots (Action-1a1) + Pookins gets cowboy boots (Consequence-1a)
2. Pookins demands a queen hat (Action-1b) + Pookins gets a queen hat (Consequence-1b)
3. Pookins asks to become a flower (Action-1c1) + Pookins is a growing flower (Consequence-1c)
4. Pookins demands to be released from the pot (Action-2a) + The gnome refuses (Consequence-2a)
5. The gnome goes under Pookins' petals (Action-3b) + The gnome is safe from the rain (Consequence-3a).
6. The gnome agrees to help Pookins only if she acts nicely (Action-2b1) + Pookins is no longer a flower (Consequence-2b).

Note. IE = Initiating Event, A = Action, Co = Consequence.

E.2: A Porcupine Named Fluffy

Episode 1:

IE: The baby needs a name.

A: The parents consider different names for their child.

Co: The parents name their child.

Episode 2:

IE (one of):

- Fluffy doubts he is fluffy
- Fluffy realizes he is not fluffy.

A/Co (one pair among):

- Fluffy tries to be a cloud/but he can't stay up
- Fluffy tries to become a pillow/but he is not comfortable to sit on
- He tries soaking in a bath/but he does not become fluffy
- Fluffy tries to put whipped cream on his quills/but it didn't make Fluffy fluffy
- Fluffy keeps on doing things to become fluffier/but none of that worked

Episode 3:

IE: Fluffy runs into an animal

A/Co (BOTH lines are required):

- The rhinoceros asks Fluffy his name (Action-3a1)/The rhinoceros now knows Fluffy's name (Consequence-3a)
- Fluffy asks the rhinoceros his name (Action-3b)/ Fluffy now knows Hippo's name (Consequence-3c).

ACo Pairs:

1. The parents consider different names for their child (Action-1) + The parents name their child (Consequence-1).
2. Fluffy tries to be a cloud (Action-2a) + but he can't stay up (Consequence-2a).
3. Fluffy tries to become a pillow (Action-2b) + but he is not comfortable to sit on (Consequence-2b).
4. He tries soaking in a bath (Action-2c) + but he does not become fluffy (Consequence-2c).
5. Fluffy tries to put whipped cream on his quills (Action-2d) + but it didn't make Fluffy fluffy (Consequence-2d).
6. Fluffy keeps on doing things to become fluffier (Action-2e) + but none of that worked (Consequence-2e).
7. The rhinoceros asks Fluffy his name (Action-3a1) + The rhinoceros now knows Fluffy's name (Consequence-3a).
8. Fluffy asks the rhinoceros his name (Action-3b) + the rhinoceros is trying to say his name but is unable to tell his name (Consequence-3b1).
9. Fluffy tries to guess the rhinoceros' name (Action-3c1) + Fluffy now knows Hippo's name (Consequence-3c).

Note. IE = Initiating Event, A = Action, Co = Consequence.