Consistency among Language Assessment Tools used in Kindergarten

by

Nicole Denise Salvador

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

SPEECH-LANGUAGE PATHOLOGY

Department of Communication Sciences and Disorders

University of Alberta

Abstract

Despite research indicating long-term risks associated with Developmental Language Disorder (DLD; Anderson et al., 2016; Bishop et al., 2016; Winstanley et al., 2018), there is still uncertainty whether children with language difficulties/DLD in the preschool years may outgrow their language difficulties by school entry. The focus of this study is on the course of development and identification of needs at kindergarten age using different language assessment tools in children with a history of struggling with language development in the preschool years.

There is some research that historically reported a substantial rate of recovery occurring in kindergarten (Bishop & Edmundson, 1987). Still others further suggested that recovery around kindergarten age may be illusory (Scarborough & Dobrich, 1990). These studies are several decades old and require further follow up within our current clinical context.

Method. We followed a case series design, looking at 4 5-year-old children who were diagnosed with severe language delays (language difficulties / risk of DLD in current terminology) prior to kindergarten entry. Six different assessments targeting different dimensions of language were administered on each child: a standardized assessment tool, language sample analysis (LSA) measures, parent report, narrative task, emergent literacy tasks, and a sentence repetition task.

Results. The results were analyzed in two ways: (1) how likely each participant would have been flagged for further evaluation given their results, and (2) how frequently each measure indicated concern across the participants. Our findings revealed that when commonly used assessments were examined, half of our participants would not have been put in a range of concern; however, when we looked at a broader range of assessments, only 1 out of the four children was likely to be in a range of *no* concern. When a more conservative set of criteria was adopted (following Bishop & Edmundson, 1987), none of the children met the "recovered" outcome. Several assessments were more likely to indicate ongoing concerns at this age: (1) LSA measures namely MLU, and sentence complexity, and as well, (2) measures of emergent literacy, and (3) sentence repetition.

Conclusion. The results from this study showed that not only is recovery from DLD at kindergarten possible, but there can also be a scenario where a child's recovery in some language skills is illusory when looking at an expanded set of language assessment tools. These findings

emphasize the need for clinicians to re-evaluate their current assessment practices to better identify children with ongoing needs at kindergarten. Several insights on how clinicians, education and policy professionals move forward with continuity of services beyond kindergarten for children with DLD were discussed.

Preface

This thesis is an original work by Nicole Denise Salvador. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name: "Language and Literacy Assessment Measures Sensitivity in Kindergarten", No. 0074679, July 13, 2017.

Acknowledgements

In one of my undergraduate seminars, a psychologist told us that where she ended up was not by choice; rather, a lot of it was influenced by the people she came in contact with. Her philosophy has always struck a chord in me. I like to think that the opportunities I had come across thus far were because of persistence and hard work, but the psychologist was also right: you need a few people in your lifetime to believe in your potential and drive you to become the best version of yourself. In saying this, there are a few souls that I will be forever indebted to for their endless support in completing this manuscript.

I would like to express my utmost gratitude to Dr. Monique Charest for taking me on as one of her students. Thank you for guiding me through the research and writing process with kindness and grit. You have pushed me to think outside boundaries and helped me recalibrate what it means to write for different audiences, and the responsibilities that came with it.

Thank you to my committee members, Dr. Phyllis Schneider, and Dr. Denyse Hayward, for asking the critical questions—they have helped me challenge the way I thought about different aspects of my research and perceive concepts in a new light.

I would like to acknowledge the Social Sciences and Humanities Research Council for funding the first year of my thesis project through the Canada Graduate Scholarships, and the University of Alberta Faculty of Graduate Studies and Research for the Walter Johns Graduate Fellowship Award.

Many thanks to Tegan Hryciw for assisting in collecting and coding data. Seeing her work with our participants inspired me to think more creatively about how I can make sessions more exciting not only for children in our study, but also with children who I will be working with in the future.

I would also like to thank a few people, now lifelong friends, who became my support system throughout this program: Joyce Chan, Jessica Zhang, Stephanie Borle, Mikayla Ho, Youran Lin, Kelly Chan, John Wayne Dela Cruz, and my thesis accountability partners: Deanna Neri, and Kevin Chavez Laxamana.

Thank you, Mama Jen and Papa Den, if not for the unthinkable sacrifices you have both made to bring our family here in Canada, I would not have been able to even dream of this opportunity. To my little sister, Isabella, thank you for believing that I can even when I thought I could not. *Maraming Salamat.*

To Matthew, I am eternally grateful for your presence. Thank you for reminding me that both successes and setbacks are better shared with someone, and that many aspirations, once thought inconceivable, are within arm's reach when you have the right people around to cheer you on.

I would never have thought to produce a piece of writing like this, but indeed, with a few people who motivate you to be 'better', you can, and you will.

To close, here is a quote from a Filipino author in his book called *Culture and History* about embracing the hardships and challenges that have shaped what we are in the present, and what we can be.

"Identity is the history that has gone into bone and blood and reshaped the flesh. Identity is not what we were but what we have become, what we are at this moment."

(Joaquin, 1988)

Table of Contents

List of Tables

Introduction

Developmental Language Disorder and the Needs that Go Unrecognized

Developmental Language Disorder (DLD), according to the CATALISE consortium, is defined as language difficulties that do not have a known biomedical etiology (Bishop et al. 2017). As the field embraces this terminology to encompass the traits of an invisible disorder that previously went by different names (e.g. language delay, specific language impairment), the struggle continues for many children with DLD. According to McGregor (2020), the needs of children with DLD go unrecognized in many classrooms. A study in the UK found that only 50% of children with DLD are being identified for intervention services at kindergarten age (Norbury et al., 2016). That means in every classroom, only 1 out of 2 children who will experience DLD will be identified. Children with DLD are at risk for literacy, academic and socioemotional difficulties as they progress in their school years (Bishop et al., 2017; Dockrell & Hurry, 2018; Eadie et al. 2018; Stothard et al. 1990). Given the potential long-term negative outcomes, it is crucial that we provide appropriate support, and this begins with the accurate identification of children's needs.

Is Recovery Possible in Kindergarten?

While some children with language needs may arrive at kindergarten never having had a language assessment, others may arrive having had prior access to preschool services. The focus in this study is on the course of development and identification of needs at kindergarten age in children with a history of difficulty with language development in the preschool years. In the literature, there is still uncertainty whether children with language difficulties/ DLD in the preschool years may outgrow their language difficulties by school entry. For example, Scarborough and Dobrich (1990) identified a research literature that simultaneously seemed to

suggest a high rate of recovery from early language difficulties, as well as persistent difficulties in studies looking at long-term outcomes. They also proposed that recovery around kindergarten age may often be illusory. While there is some research that points to the possibility of recovery from early language difficulties, these studies are now several decades old and need more follow-up. Whether recovery in kindergarten is true or illusory, the implications for children with DLD are real and far-reaching; thus, further examination in this area is needed.

Bishop and colleagues (2016) suggested that the trajectories of children with DLD differ in many ways depending on age and severity of the disorder. For example, some children whose difficulties are solely in expressive phonology seem to have better prognoses, whereas those who have impairments in many areas of language functioning do not (Bishop et al., 2016). In a recent paper reviewing the existing evidence on recovery from early language difficulties during kindergarten entry, Charest et al. (2019a) explained that to understand outcomes, it is necessary to look at the different assessment measures used in kindergarten. For example, measures looking at syntactic complexity might imply something different than measures of information processing, or predictors of literacy outcomes (Charest et al., 2019a). To that end, the authors also suggested that some of these measures might be more sensitive at detecting persisting language difficulties that otherwise would have gone unnoticed in other forms of assessment. If so, these are important considerations for clinicians who may be working with large caseloads, and with limited time. This paper will look at the consistency among assessment tools used in kindergarten to ensure that ongoing needs are not missed.

Differentiating True or Illusory Recovery

Scarborough and Dobrich (1990), in summarizing the existing evidence at the time on outcomes in children with language disorders, observed that a substantial number of children

with previously identified language difficulties appear to have resolved their difficulties by kindergarten age. However, Scarborough and Dobrich (1990) cautioned that for many, this recovery may, in fact, be illusory. This group of children in kindergarten, they suggested, are likely to have unresolved underlying difficulties, and miss out on continued support or services from educational professionals and speech-language pathologists (S-LPs). This proposal is consistent with the finding from the studies in the literature at the time, that showed that language disorders tend to be long-lasting.

In discussing the potential illusory recovery phenomenon, Scarborough and Dobrich (1990) proposed that children, both typically developing (TD) and those with early language difficulties, undergo periods of both growth and plateau in their language development. In a period of plateau such as at in the late preschool years, Scarborough and Dobrich (1990) argued that where TD children level off in some of their language abilities, children with previously identified language disorder, in essence, would have an opportunity "catch up" to their TD peers in some skills. In a longitudinal sample following 12 TD children and 4 children with identified severe language delays, Scarborough and Dobrich (1990) found evidence in support of this proposal. In the four measures of language production that they examined, namely pronunciation accuracy, lexical diversity, mean length of utterance (MLU) and Index of Productive Syntax (IPSyn; grammatical complexity), they found that in their control group, the rate of progress started to plateau after a sharp increase at earlier time points. They saw this levelling off begin at 36 months for lexical diversity, and at 42 months for the other measures (Scarborough and Dobrich, 1990). These plateaus, as argued by the authors, provided opportunities for the four children with language difficulties to "catch up" and score in the average range in all of these measures of language production before kindergarten entry. However, these gains were transient

over the long term. At Grade 2, three out of the four children were identified as having severe reading disabilities.

Interestingly, the children studied by Scarborough and Dobrich (1990) were found to have less severe forms of expressive language delay as they aged: from a more global language difficulty, these children had selective deficits over time (i.e. residual phonological and/or syntactic difficulties). However, this trend did not hold true for their receptive language skills: these four children were reported to have scores that were one standard deviation (*SD*) below the mean, and severity was greater when their expressive language skills caught up. Scarborough & Dobrich (1990) cautioned, however, that the receptive language measures were not used at all time points in their study, and not consistently at similar ages, therefore, conclusions about receptive skills are somewhat limited.

The data reported by Scarborough and Dobrich (1990) advanced the hypothesis that having difficulties with language early on was related to challenges with literacy in the school age years—an outcome that has since been well documented (e.g., Bishop, & Adams, 1990; Catts et al., 2016). Scarborough and Dobrich (1990) cautioned that the apparent resolution of language difficulties at kindergarten age does not suggest that language skills stop developing overall as children master them in these areas of plateau. Rather, they cautioned that the measures that were used in their study "may not adequately tap changes in language proficiency that may be occurring" (Scarborough & Dobrich, 1990, p. 12). In other words, the measures used in the study may have not fully captured the different language skills that their participants had at the time. With a limited sample size of 4 children with language difficulties, conclusions may be limited in terms of its generalization; nevertheless, this paper has prompted questions about the

tools we use to evaluate the language abilities of children previously identified with language difficulties/DLD, and which skills they were designed to measure.

Although there is limited research directly addressing the Illusory Recovery hypothesis, a series of papers by Bishop and Edmundson (1987) is relevant to this discussion. In the first of a series of studies to look at recovery from language difficulties, Bishop and Edmundson (1987) found that recovery is possible such that a proportion of children with DLD (i.e. "the good outcome group") who scored in the range of language difficulty at age 4 (described at the time as having SLI) obtained scores that were indistinguishable from their TD peers at age 5 1/2. Of 68 children identified with DLD at age 4 in this study, 44% obtained test scores above the 10th percentile in most verbal language measures, placing them in the average range at age 5 ½ according to the outcome criteria set by the authors.

Bishop and Edmundson (1987) proposed several reasons to explain the high rate of recovery at age 5 ½. One plausible possibility was that many early language disorders do in fact resolve. Another was that speech-language intervention between age 4 and 5 ½ during the study made a difference. Because this variable was not controlled for, it was difficult for the authors to infer how much intervention had helped improve the participants' performance versus how much improvement was spontaneous. Lastly, an intriguing proposal was that the apparent recovery may have been influenced by the measures used. The authors noted that the measures used in the study mostly looked at skills that were mastered earlier in development by TD children, meaning that a proportion of those children with DLD had a chance to catch up on these abilities and "recover." Bishop and Edmundson (1987) suggested that ongoing difficulties may have been more apparent if instead they had used more sensitive measures.

On the note of prognosis, Bishop and Edmundson (1987) found that children's performance at age 4 predicted their outcome at age 5 ½. They cautioned here that this prediction was highly dependent on the measures used prior. For example, phonological impairment at age 4 was not different between the good and poor-outcome groups of children with DLD. On the other hand, measures of verbal semantic ability, such as the Renfrew Bus Story that looked at how much information they were able to generate based on a story, were reliable predictors of challenges later on. In particular, children who scored low on the Bus Story task at age 4 were more likely to have poor outcomes at age 5 ½. In discussing their results, the authors emphasized that their study provided evidence that recovery is possible depending on the severity of language impairment from the outset. That is, an isolated difficulty in one specific skill area was more likely to be associated with a good outcome or some form of recovery than were more broad-based difficulties such as children with an underlying receptive impairment as well (Bishop & Edmundson, 1987).

Stothard and colleagues (1998) reported the age-15 outcomes of the children studied by Bishop and Edmundson (1987). They found that 52% of the children who were in the recovery (i.e., "good outcome") group at age 5 1/2 experienced literacy challenges (i.e. reading comprehension, accuracy and spelling) at age 15, compared to only 22% of children in the TD group. In addition, even when the children in the recovered group were deemed to have satisfactory speech and language skills (i.e. no score in the impaired range, and no more than one score below satisfactory range), they still performed significantly worse than the TD group in complex language processing tasks that involved verbal short-term memory and phonological awareness skills (e.g. sentence repetition, nonword repetition, spoonerism tasks). These results were contrary to their earlier data in Grade 2 (see Bishop & Adams, 1990), which showed that

children who had been in the resolved group at age 5½ did not have issues with reading or spelling at age 8.

Stothard et al. (1998) discussed the role of assessment measures in distinguishing children who have recovered or changed in their 'impairment' status (i.e. recovered to impaired or vice versa). That is, measures that looked at verbal short-term memory revealed difficulties for children in the resolved group at 15 years of age (i.e. sentence repetition, nonword repetition and spoonerism tasks). In addition, the narrative task, Renfrew Bus Story, at age 5;6 distinguished children who maintained their recovered status versus the ones who had impaired status later on. The latter corroborates with the results by Bishop and Edmundson (1987), highlighting that narratives reveal ongoing language difficulties that are not evident in other measures.

In regard to recovery, Stothard and colleagues (1998) pointed out that children who had recovered status at 5 ½ years still performed within the typical range for most of the measures of oral language abilities at age 15 but had less than satisfactory standing in terms of literacy skills. The apparent re-emergence of difficulties in skills related to language support the notion of illusory recovery (Scarborough, & Dobrich, 1990; Stothard et al., 1998).

On the whole, the studies reviewed in this section thus far suggest that illusory recovery may occur: some children who seem to have caught up in their preschool and kindergarten years are still at risk for negative long-term outcomes and need more follow up. However, it is important to note here that these studies are several decades old, and that many assessment tools and approaches have changed since. Despite these limitations, concerns about illusory recovery as well as how and what we should be assessing at kindergarten-age remain relevant to S-LPs in current practice according to a recent survey of S-LPs in Alberta (Charest et al., 2019b).

Differentiating Language Skills in Kindergarten

In the previous section, several conclusions emerged from the literature looking at recovery around kindergarten and beyond. First, it was necessary to look at different time points in a child's language development to understand how certain skills undergo periods of spurts, and plateaus (Bishop & Edmundson, 1987; Scarborough & Dobrich, 1990). Second, it was imperative to look at the kinds of assessments being used in kindergarten as they might indicate something different regarding a child's language abilities. Undoubtedly, these two points are intertwined; as Bishop and Edmundson (1987) noted, it is critical that the measures being used should be appropriate to the age range of interest.

The Ascendancy Hypothesis

Looking into the skills that better predict later literacy especially in children with DLD, Scarborough (2009) discussed the "Ascendancy Hypothesis," which proposes that while we see plateaus in some language skills for TD children, we also see continuous growth in other areas. Scarborough referred to these areas of continued growth as "ascendant" skills and argued that focusing on this group of skills will likely reveal ongoing challenges in children with DLD. In essence, this hypothesis furthers the "Illusory Recovery" hypothesis discussed earlier and incorporates the importance of using tasks that tap into ascendant skills when attempting to determine whether or not language difficulties have been resolved.

From the limited number of longitudinal studies that we have discussed, we have seen these differences among measures emerge. One is that kindergarten children in the resolved group who scored in the average range in measures of verbal production (i.e. MLU, grammatical complexity, percent consonants correct) performed poorly in the Bus Story narrative task (see Bishop & Edmundson, 1987). Furthermore, at 15 years, while these same children scored in the

average range in oral language tasks (as measured by the WISC-III Vocabulary subtest;

Wechsler, 1992), they did poorly in activities that required verbal short-term memory and

phonological processing (see Stothard et al., 1998).

Charest et al. (2019a) provided a detailed discussion of measures that may be more

sensitive at revealing the ongoing difficulties of children with DLD during kindergarten entry,

based on the existing literature. Of these, the following emerged from the review as measures

that are known to be sensitive indicators of language difficulties around age 5, and are thus

considered as tapping ascendant skills: (1) narrative comprehension and production (Bishop &

Edmundson, 1987; Schneider et al., 2006), (2) sentence repetition (Catts et al., 2001; Stothard et

al., 1998), (3) complex sentence production (Guo et al. 2020), (4) literacy predictors such as

rapid automatized naming, measures of phonological awareness as well as knowledge of letters,

and print (Bishop & Edmundson, 1987; Scarborough, 2009; Stothard et al., 1998). On the

contrary, while language sample measures such as MLU, sentence complexity, and lexical

diversity are informative in comparing the child's performance to TD children during and

beyond kindergarten, they may not be as sensitive to later development as measures of ascendant

skills depending on how the language sample analysis was done, and which variables were

considered (Charest et al., 2019a; Scarborough & Dobrich, 1990).

In the next section, we turn our attention to what is currently being used clinically to

identify whether children previously diagnosed with DLD still need further assessment or

monitoring, and how these are related to the earlier-developing and ascendant skills that

Scarborough (2009) discussed.

Clinician's Toolbox of Kindergarten-Age Assessments

In reporting the results of a survey of clinicians in Alberta, Charest et al. (2019b) stressed that clinicians use multiple perspectives to inform assessment and diagnosis of DLD. In contrast to research studies, which typically determine outcomes of children with DLD using mostly test results, Charest et al. (2019b) found that it is common practice for clinicians to triangulate information from three sources: (1) standardized tests, (2) their clinical observations, and (3) input from parents as well as teachers regarding their thoughts on a child's language skills (Charest et al., 2019b). In this section, we discuss the variety of tools that are available to clinicians when making both diagnostic and prognostic decisions regarding children who have previously been diagnosed with DLD.

Omnibus Standardized Assessments

Standardized assessments are useful tools in gauging a child's language performance relative to same-age peers. In addition, many jurisdictions mandate that standard scores from these assessments be included in order for a child to be considered for funding (e.g., see Government of Alberta, 2013). While some standardized assessments measure a wide range of oral language abilities, these tests are generally not designed to capture the functional use of language in communicative contexts. Paul, Norbury, and Gosse (2018) emphasized that although these measures are valid, it is possible that a child could perform within the typical range of scores in these assessments, and still have difficulty in their daily functioning. Conversely, another child could also perform poorly on an assessment, but would otherwise have no difficulty with their language abilities outside testing. Indeed, standardized measures give clinicians some idea in terms of future progressions of DLD in children; however, we must also consider other measures to paint a wider picture of a child's language abilities. In the survey of

clinicians previously mentioned, Charest et al. (2019b) reported that none of the clinicians made diagnostic decisions relying solely on a standardized assessment. This result prompts us to look at a collection of evidence from multiple lenses in order to make informed decisions about the continuity of services for children with DLD.

Measures of Functional Language Use

In the next section, we discuss two measures used by clinicians (see Charest et al., 2019b) that look at more contextualized communication.

Insight from Norm-Referenced Parent Report Questionnaires

Even though there is movement towards a more evidence-based and family-centered approach in clinical practice, there is scant literature looking at the role of assessment tools like parent report in terms of its use in language assessment (Boudreau, 2005; Hall, & Segara, 2007). In this section, we explore the merit in considering parent perspectives, specifically norm-referenced questionnaires, when identifying children with DLD.

Parent report has been considered to play an important role in language assessment for infants and young children. From a clinical perspective, parent-report measures provide a more comprehensive account of a child's environment and behaviours that do not occur in contrived contexts as in standardized testing situations (Bishop, & McDonald, 2009; Boudreau, 2005; Dale, 1996). In addition, existing research has indicated that including parent report measures, particularly those that provide normative results, give a better index of clinically significant communication needs (Bishop & McDonald, 2009; Dockrell & Hurry, 2018; Eadie et al., 2018).

In a study looking at factors predicting the likelihood of teachers identifying the speech and language needs of children at age 7 and 11, Dockrell and Hurry (2018) found that parental identification of concern accounted for most of the variance in their study across their different

models. Parental concerns (i.e. asking parents whether they were concerned about their child's language skills or not) expressed at ages 3 and 5 predicted the scores of children with speech and language needs at age 7 (Dockrell, & Hurry, 2018). Although they did not use a norm-referenced parent questionnaire, Dockrell and Hurry (2018) stressed the importance of parental perspectives when assessing children with DLD.

Given that there is less and less contact between parents and clinicians as a child moves from kindergarten to elementary (see Dockrell & Hurry, 2018), it might be worthwhile to consider using parent reports with normative scores to be able collect information regarding a child's communicative context in a systematic, and cost-effective manner. An example is the Children's Communication Checklist - Second Edition (CCC-2; Bishop, 2006). Bishop and McDonald (2009) found that the children who were flagged in a standardized test (e.g. naming task, Test of Word Reading Efficiency, Woodcock-Johnson III), and the CCC-2 were more likely to get referred than those who scored in a range of concern in one or the other type of tool. This finding suggests that parent-report measures of language and communication may be sensitive to functional communication challenges that together with test scores may pick up on children with greatest need.

Insights from Narratives

With respect to the previous studies looking at recovery and persisting difficulties in children with DLD, narratives have revealed the challenges that children with DLD face in oral language (Bishop, & Adams, 1990; Bishop, & Edmundson, 1987; Stothard et al., 1998). Several studies in that period and since have strengthened the notion that narratives are a robust predictor of difficulties in using language in discourse in both TD children and children with learning disabilities or those at risk for language difficulties, respectively (Fazio et al., 1996; Griffin et al.,

2004; Schneider et al., 2006). However, consistency in the context being presented in narrative

tasks has to be ensured to confidently compare samples of children with DLD to a normative

sample (Schneider et al., 2006).

Schneider et al. (2006) reported that Story Grammar scores on the Edmonton Narrative

Norms Instrument (ENNI) differentiated performance of TD children and children with DLD in

that children with DLD scored lower than the TD group overall. This means that children with

DLD gave significantly fewer pieces of information than those in the TD group. In addition,

there are recent studies as well that report differences in performance between children with

DLD and TD children in grammatical error and clausal density measures of ENNI narrative

samples (Guo et al., 2020; Guo et al., 2019).

Language Sample Analysis Measures: Microstructure

Clinicians also examine spontaneous language production of children at the level of

microstructure, in play, conversation or narrative (Charest et al., 2019b). The following is a

breakdown of measures typically considered when collecting language samples.

Mean Length of Utterance

Mean Length of Utterance (MLU) collected from a conversation- or play-based language

sample has been a widely used clinical measure especially for children in the early years

(Goffman & Leonard, 2000). However, some previous studies have reported plateaus in MLU

and have found that this measure is less likely to differentiate between children with and without

language difficulties beyond the preschool years (Bishop & Edmundson, 1987; Scarborough &

Dobrich, 1990). Depending on the severity of a child's language disorder, it is still possible that

MLU could point to oral language difficulties in conversation for children with DLD. Some

studies have also shown that when MLU is combined with other LSA measures (such as lexical

diversity) to form an index, the ability to differentiate between disorder and no disorder is higher

(Klee et al., 2004; Wong et al., 2010; Shahmahmood et al., 2016).

Lexical Diversity

Lexical diversity is a measure of the number of different or unique words (NDW) that a

child uttered in a sample. Children with DLD tend to use non-specific words, thereby making

their diversity in words significantly different from TD children (Paul et al., 2018). When

measuring this construct, it is important to control for length of sample or consider specific

measures that are less sensitive to difference in sample lengths (Charest et al., 2020; Heilmann et

al., 2010). Some studies have urged clinicians to interpret this measure with caution as it might

not always report concern for children with DLD (Charest & Skoczylas, 2019).

Syntactic Complexity and Grammatical Accuracy

In typical language development, it is evident that children continue to learn and use

complex syntactic structures well beyond the preschool years (Paul et al., 2018). For children

with DLD, existing research has pointed out that their skills are not on par with TD children at

age 8-9 as well as at 15 years of age; studies have shown that children with DLD score

significantly lower in measures of syntactic complexity in language samples and produce more

grammatical errors, especially when looking at grammatical morphemes (Domsch et al., 2012;

Manhardt & Rescorla, 2002; Rescorla, 2009). In a study done by Domsch et al. (2012), they

found that at 8-9 years old, while children with DLD score in the average range on a

standardized narrative task, they scored significantly lower in syntactic complexity in a 5-minute

spontaneous language sample. This finding strengthens the possibility raised by Bishop and

Edmundson (1987) in their discussion that certain syntactic measures might be able to detect

ongoing difficulties whereas other measures may not.

Looking at the single word level, children with DLD experience difficulties in using grammatical morphemes particularly auxiliary verbs and tense markers (Leonard, 2014). As previously mentioned, when grammatical accuracy is combined along with other measures, this provides a better index of identifying disorder or no disorder (Shahmahmood et al., 2016). Note here that these LSA measures have been reported to become less sensitive over the course of the preschool years, particularly if collected in play/conversation (Bishop & Edmundson, 1987; Scarborough & Dobrich, 1990), but that narrative likely provides a more challenging context to allow continued sensitivity. In addition, some measures have not been extensively explored, and may continue to be sensitive (e.g. grammatical errors and syntactic complexity) (Eisenberg & Guo, 2013).

Measures of Emergent Literacy Skills

As noted previously, measures looking at emergent literacy skills (i.e. print concept, letter knowledge, phonological awareness) tap into ascendant skills that may be sensitive to ongoing difficulties around kindergarten age. Although it has been suggested that indices of these skills are better predictors for kindergarten and in the school years, many researchers are still working on the best combination and criteria for these indices to have better predictive power over longer periods of time (Milburn et al., 2017; Scarborough, 2009). In addition, while there are assessment tools that currently examine these skills in kindergarten, clinicians are more likely to use some of these measures with school-age children (Charest et al., 2019b).

Sentence Repetition

Stothard et al. (1998) found that children who had recovered status from DLD at 15 years had residual difficulties in sentence repetition, verbal short-term memory and language processing tasks, and performed worse than the TD group— even though they scored in the

average range in measures of oral language abilities. Since the results from Stothard et al. (1998)

were released, more and more studies looking at preschool, and school-aged children have

shown that sentence repetition, a measure of verbal short-term memory, is a reliable indicator of

persisting language difficulties (Archibald & Joanisse, 2009; Everitt et al., 2013; Conti-Ramsden

et al., 2001).

Summary

The previous section highlighted the variety of tools that clinicians utilize when they

make diagnostic and prognostic decisions on children with DLD. Each of these tools capture a

unique facet of a child's language skills at kindergarten, and in doing so, may be more or less

sensitive to language difficulty at this age as the language skills of children with typical language

undergo periods of plateau and growth. In particular, language sample measures, according to

the literature, may correspond to earlier-developing language abilities while an omnibus

standardized assessment, depending on the test, might both look at earlier- developing skills and

some skills that develop later. On the other hand, tasks that measure emergent literacy skills,

verbal short-term memory, and discourse tend to look at later-developing abilities (see

Scarborough, 2009). It is clear in the literature that there are still questions regarding the

apparent recovery of children previously diagnosed with DLD during kindergarten. If it is the

case that recovery around this age often proves to be illusory, it compels clinicians to find ways

to identify lingering or future challenges to ensure these children do not fall through the cracks.

After all, the underidentification of DLD in kindergarten remains a relevant issue to this day

(Norbury et al., 2016).

A Note on Case Series Approach

      The main goal of the present study was to look at the consistency among assessment tools in the degree to which they highlight language concerns in 5-year-old children with identified language delays in preschool. Because of the recruitment and data collection constraints presented due to the COVID-19 pandemic, this study follows a case series design focusing on four children who were identified with language delay as preschoolers and seen again for assessment at kindergarten age. As Macleod and colleagues (2020) noted: the aim of a case series approach is "to develop profiles of children, particularly the strengths, and weaknesses of their language, by drawing on diverse sources of information (p.4)." Given that the analysis was primarily descriptive, a case series design allowed us to look further into the variability of decisions when we considered all the available data from the assessments used in the study. The nuances that existed in the analyses of each participant afforded us the opportunity to observe different profiles of language difficulties and potential recovery, if any, in the different language assessments that were used. We would not have had the same opportunity to delve further in each participant trajectory had we collected data from a large sample.

      By looking at the data of four participants, we were able to simulate the decision-making process that clinicians encounter regarding the prognosis of children with DLD after kindergarten. As such, this study hopes to encourage a clinical discussion on what constitutes 'recovery' from DLD, if it is a real phenomenon, and explore how the potential for illusory recovery affects the decision-making during kindergarten based on the information that clinicians have.

Research Questions

To address the overarching issue of whether clinicians need to be concerned about illusory recovery, we considered the following questions in this study:

Q1. When typical assessment approaches are used, do we replicate the finding that a substantial proportion of children seem to recover from their previous language difficulties?

Hypothesis: When we consider the typical assessment tools that clinicians use (see Charest et al., 2019b) namely the Clinical Evaluation of Language Fundamentals - Preschool 2 (CELF-P2; Semel, Wiig, & Secord, 2003)—a standardized assessment of oral language— and LSA measures (i.e. MLU, syntactic complexity, lexical diversity, grammatical accuracy), we predict that some kindergarten children will likely show recovery evidenced by standard scores falling in the range of "**no** concern". This is supported by previous studies showing improvement in language scores over the preschool years in different language assessment measures (see Bishop & Edmundson, 1987; LaParo et al., 2004; Scarborough & Dobrich, 1990).

Q2. When considering a broader range of measures that look at kindergarten skills, do we see different results emerge?

Hypothesis: We expected that there would be some variation in outcomes across different measures when a wide range of assessment tools are used, given that different tools emphasize different skills and all test scores are only estimates of true ability. We examined how the inclusion of a broader range of measures affects conclusions about recovery versus ongoing need, namely:

- CCC-2 (Bishop, 2006), a measure of a child's overall communicative function according to parent report;

- ENNI (Schneider et al., 2005), an Edmonton-normed narrative task;

- Test of Early Language and Literacy (TELL; Phillips et al. 2016), a Canadian-normed measure of emergent literacy skills, and;

- Comprehensive Test of Phonological Processing - 2 (CTOPP-2; Wagner et al., 2013), a measure of phonological awareness skills and rapid automatized naming.

In general, we expected that there will be differences in the results among commonly-used assessments such as the CELF-P2 and traditional LSA measures and other assessments, in particular, measures that look at functional language use (i.e. ENNI, CCC-2), measures of emergent literacy skills, and sentence repetition.

Q3. Are there systematic patterns of similarity and difference in outcomes across measures?

Hypothesis: When standard scores are dichotomized to pass-fail decisions, we expected that there would both be a degree of concordance and discordance between measures. Some children potentially will score below cutoff (i.e. in the range of concern) on all measures, while others could score below cutoff on all measures. That being said, where there are differences between measures, we predicted that, similar to the Ascendancy hypothesis, participants will more likely be flagged for ongoing concern in assessments that look at later-developing or ascendant skills (i.e. narrative-based measures, emergent literacy, sentence repetition) than those that focus on earlier-developing abilities (Bishop & Edmundson, 1987; Scarborough, 2009; Stothard et al., 1998).

Method

Participants

This project was part of a larger cross-sectional study looking at kindergarten-age outcomes for children with previously identified language difficulties, when different assessment approaches are considered. Data were collected originally from five participants. However, out of the five, only data from four individuals (M_{age} = 5 years; 6 months; 2 girls) were reported. For the remaining child, prior records of language ability were ultimately not able to be located.

Table 1.0. Participant Characteristics at Intake

Participant	Age at time of testing	Language Diagnosis at Preschool Age	Speech Diagnosis at Preschool	NVIQ	Hearing Screen	First Language	Other Languages
Ava	5;10	Severe Expressive Language Delay (ELD); Mild Receptive Language Delay	Severe Speech Sound Delay (SSD)	Average	Pass	English	N/A
Carter	5;6	Severe ELD	Severe SSD	Average	Pass	English	N/A
Isla	5;6	Severe ELD	N/A	Average	Pass	English	N/A
Billy	5;3	Severe Delay involving Language	Severe SSD	Average	Pass	English	French

Note. Table 1.0 details the information gathered from the participants at intake. The children in this study are referred to as Billy, Isla, Carter, and Ava (not their real names). Billy was 5 years; 3 months old at the time of testing, Ava was 5 years; 10 months old, and Carter, and Isla were both 5 years; 6 months old. Prior to kindergarten, all four children received a diagnosis of severe expressive language delay according to their previous language assessment reports. Two of the children were also diagnosed with receptive language difficulties, indicated as mild for one child, and severity rating not given for the other. In addition, three out of the four children were also diagnosed with a speech sound delay. Given that recruitment was challenging for this specific population and that many children typically referred for language difficulties also have speech

sound difficulties (McGregor, 2020), we had included children whose speech sound difficulties did not severely impact their ability to participate in a language assessment.

At intake to the study, all children completed the nonverbal matrices section of the Kaufman Brief Intelligence Test - 2nd Edition (KBIT-2; Kaufman & Kaufman, 2004). All scores were in the average range, ranging from 86 to 112. All passed a hearing screen at 20 dB. By parent report, none of the children had other diagnoses affecting language development such as Autism Spectrum Disorder (ASD) or hearing loss. All four children were not previously exposed to another language besides English at home. One child started attending a French Immersion kindergarten program three months before their participation in the study began. Participants were recruited through their school district.

Measures

Clinical Evaluation of Language Fundamentals – Preschool II Edition (CELF-P2; Semel et al., 2003). The CELF-P2 is a standardized tool used to assess a child's language and communication skills (e.g. receptive and expressive language, language structure and content, literacy and phonological awareness). It was normed on 800 children aged 3 to 6 in the United States. The sum of three subtests (i.e. sentence structure, word structure, and expressive vocabulary) is called the Core Language Score (CLS), and this composite score is used to differentiate performance between TD children and those with language disorders. According to the manual, the sensitivity of the CELF-P2 CLS using a -1 *SD* (standard score = 85) cutoff is 0.85, while its specificity is 0.82. In this study, the standard scores from CLS as well as the Expressive Language (ELI) and Receptive Language (RLI) indices were used to compare among participants, and among assessments.

Sentence Repetition

The recalling sentences subtest of the CELF-P2 looks at sentence repetition. Here, participants repeat phrases and sentences of varying length and complexity. The mean for the subtest is 10 with a *SD* of 3. A scaled score, and a percentile range can be derived from the raw score on this subtest. The scaled score also contributes to the ELI mentioned above. Although this subtest amounts to a scaled score, the CELF-P2 provides percentile ranges for each range to denote whether a child's sentence repetition skills are within typical range.

Children's Communication Checklist 2 (CCC-2; Bishop, 2006). The CCC-2 is a checklist of a child's skills and behaviours in the following domains of communication (referred to as scales in the checklist): structural language (i.e. speech, syntax, semantics, coherence), pragmatics/social communication (i.e. inappropriate initiation, stereotyped language, use of context, nonverbal communication), and other behaviours associated with Autism Spectrum Disorder (i.e. social relations, interests). This study used the US version of the checklist, normed on 950 children aged 4 to 16 years. The CCC-2 consists of 70 items, for which a parent or caregiver provides a rating of how often their child exhibits different communicative behaviours. Values range from 0 (i.e. less than once a week) to 3 (several times or more than twice a day or always). The total sum of the scores in the first eight scales is called the General Communication Composite (GCC), which is used to flag children who likely have clinically significant concerns. The best balance between sensitivity and specificity of the GCC in the CCC-2 is at -1*SD*, which is 0.70 and 0.85 respectively. At -2*SD*s, the sensitivity declines to 0.31, while the specificity increases to 0.96.

Conversational Language Sample Analysis Measures. Conversational language samples were collected following protocols set out in the Systematic Analysis of Language Transcripts (SALT; Miller et al., 2019) in which suggested topics are provided as well as recommendations for

asking questions that encourage more than yes/no responses. In the SALT program, language sample variables for a target child can be compared to a typically developing peer group that is matched to the child's age. Z-scores reporting comparison to the normative group are available for a range of language measures. In the present study, we focused on the following: (1) MLU in morphemes (MLUm), (2) mean number of verbs per utterance as a measure of sentence complexity (SC), (3) moving average NDW to measure lexical diversity (LD) (i.e. the average of the NDW for every 100-word window), and (4) percentage of utterances with errors for grammatical accuracy (GA).

Edmonton Narrative Norms Instrument (ENNI; Schneider et al., 2005). The ENNI is a tool to collect information regarding a child's language skills through storytelling. It was normed on 377 children aged four to nine years from Edmonton, Alberta, Canada. Here, children generated stories about from picture sequences to an examiner who could not see the pictures. The language sample collected can be scored to examine a child's knowledge of story grammar components in comparison to typically developing children or analyzed using SALT Software to look at traditional measures of language skills in discourse (e.g. sentence complexity, grammatical accuracy). In this study, scores from both macrostructure and microstructure analyses were included: story grammar, as well as MLU in morphemes, mean verbs per utterance, moving average NDW and percent grammatical errors. Additionally, only the Set A stories were used instead of both Set A and B. The story transcripts were compared according to the ENNI database in SALT to obtain *z*-scores.

Test of Early Language and Literacy (TELL; Phillips et al., 2016). The TELL is an assessment tool that measures both language and literacy skills of children aged 3-8 years. This assessment was normed on 1061 children from schools, daycares, and preschool programs

around Canada. The TELL includes subtests examining phonological awareness and written

language skills; however, for the purpose of this study, the following subtests were used: Print

Understanding, Letter Knowledge, Oral Vocabulary, Word Reading, Oral Reading and

Comprehension Subtests. These subtests look at emergent literacy and oral language skills,

including items that pertain to abstract or higher-level language concepts.

The TELL includes several options for how clinicians can use the scores derived from the

assessment: one is administering all the subtests to be able to calculate a composite score ($M =$

100, $SD = 15$) that indicates a holistic performance of a child in oral language and/or literacy.

The other option, which is relevant to this study, is to calculate the standard scores derived from

the scaled scores of each subtest and use them clinically to inform where a child is performing

for that specific skill. Although no sensitivity and specificity ratings were included, the authors

of the test have found notable consistencies with other language and literacy assessments such as

the CELF-P2 or CELF-4 (p.254). As noted by the authors of the test, this suggests that while

there are aspects that are similar between assessments, the TELL is able to identify specific

concerns that other assessments might not tap into (p.254). As well, when children with language

and/or literacy difficulties were age-matched with TD children, the TELL was able to

differentiate between the two groups except in the Phonological Awareness (PA) subtest.

Comprehensive Test of Phonological Processing – 2 (CTOPP-2; Wagner et al., 2013)

The CTOPP-2 is an assessment tool that measures phonological processing abilities of

individuals aged 4-24 years on four composites, namely: phonological awareness, phonological

memory, rapid symbolic naming, and alternate phonological awareness. This assessment was

normed on a representative sample of 1900 participants across the United States. Each composite

is composed of scaled scores from different subtests: (1) phonological awareness is comprised of

elision, blending words, and phoneme isolation; (2) phonological memory is a composite of

scores from memory for digits and nonword repetition; (3) rapid symbolic naming is a composite

from rapid digit and letter naming; (4) finally, the alternate phonological awareness composite is

derived from blending nonwords, and segmenting nonwords. For the purpose of this study, the

first three composite scores were used for analysis.

Procedure

After obtaining consent for a child to participate in the study, a certified S-LP or a

graduate trainee (the author) administered the assessments in individual sessions with each child

over four sessions. The order of the tasks was counterbalanced. Testing was completed at either

the University of Alberta or at the children's school sites. Breaks were given throughout each

session, and at the end of the last visit, children were given a small prize and certificate for their

participation. Parents also filled out the CCC-2 form and returned it for scoring. At the

conclusion, parents were provided the option to receive a summary of their child's results.

Reliability Analyses

To ensure consistency in the data being analyzed, several reliability measures were used.

In particular, we looked at reliability in scoring the assessment tools. Additionally, we also

considered agreement in how language samples were transcribed (i.e. similarity in word-for-

word transcription and morpheme segmentation, utterance segmentation through C-units as well

as identification of grammatical errors) both for a conversational language sample and a

narrative sample. In the case of reliability in scoring of test administration, the author attempted

to score the CELF-P2 from the videorecording of a S-LP administering the assessment on one of

our participants. This process was not completed, however, as the stimulus book was often

covered by the child, which made re-scoring of several items impossible. The author instead checked whether all scores were correctly recorded using the appropriate age norms.

In the case of transcription reliability, the analysis of two different transcripts from a conversational language sample of a participant showed that the rate of agreement for word-for-word transcription (including morpheme segmentation) was 96.3%, utterance segmentation was 93.2% and segmentation of grammatical errors were similar in 3 out of 4 opportunities. One disagreement was marking subject omission in one utterance. In this case, the author cross-checked other transcripts to ensure that missing the subject omissions was not a systematic transcription error. In spite of the differences in coding grammatical errors, database comparison outcomes for percent grammatical errors in the narrative, and conversational transcripts did not change the outcome (i.e. child still scored above cutoff in both samples transcribed by different coders).

Results

In this study we wanted to explore whether clinicians should be concerned about illusory recovery in kindergarten, and if so, which measures might be important to consider when assessing children previously diagnosed with language delay/disorder (i.e., consistent with DLD) at this age. To this end, we examined the rate at which children with previously identified severe expressive language delays, and half with receptive language delay as well, appeared to have recovered at kindergarten age using an omnibus language test focused at the word and sentence level, traditional language sample measures taken from a conversational language sample, as well as when using an expanded set of measures that included emergent literacy.

In answering the research questions, we will discuss the individual results for each child, and then turn to consider the results at a group level. Table 2.0 and 2.1 summarizes the standard

scores ($M = 100$; $SD = 15$) or the z-scores as appropriate of each child for each of the language

assessment measures that we looked at. Where available, beside their scores are confidence

intervals (CI) set at 95%, percentile ranks or ranges. Note that the scores for the Recalling

Sentences subtest from the CELF-P2 are scaled scores ($M = 10$, $SD = 3$). In Table 2.1, notice that

Isla did not have scores for the ENNI measures as she was not able to complete the task— she

had difficulty generating a story from pictures in the training story and in the first story (i.e. A1)

even when given the appropriate prompting. In addition, some CTOPP-2 scores (i.e. rapid

symbolic naming) on the table were not calculated as some of the participants' raw scores were

unusable. According to the manual, these scores were not valid for interpretation in this case as

some participants made 4 or more errors in the rapid naming subtests.

In Tables 2.0, 2.1, 3.0 and 4.0, measures are organized with word- and sentence- level

and conversation-based assessments at the top, and then narrative and then literacy measures that

may tap ascendant skills. Table 2.0 and 2.1 present the standard scores of each child for each

measure, and Tables 3.0 and 4.0 present the test results for each child for each measure,

dichotomized into a "below cutoff" or "above cutoff" decision at two respective cutoff points: -1

SD (Table 3.0) and -2 SD (Table 4.0). These two cutoff points were analyzed for two reasons: (1)

clinicians typically use -1 SD to flag language concerns (Charest et al., 2019b) and; (2) this

cutoff point is also the best balance between sensitivity and specificity in the case of CELF-P2.

Two SDs below the mean, on the other hand, is used in many jurisdictions as a funding cutoff for

severe language concerns (Spaulding et al. 2006). In the following section, differences in how

the language measures flag concern depending on the cutoff point will be discussed.

Table 2.0. Summary of Standard Scores among Language Assessment Measures I

Assessment Measure	Ava			Carter		
	Standard Score	z-score	95% CI	Standard Score	z-score	95% CI
Earlier Developing Abilities						
LSA MLU in Morphemes		-2.14**			-2.11**	
LSA SC Mean Verbs per Utterance		-2.18**			-1.39*	
LSA LD Moving Average NDW		-1.47*			0.26	
LSA %Utterances with Errors		-1.15			-1.25	
Mid-range Developing Abilities						
CELF-P2 CLS	77	-1.53*	69 - 85	84	-1.10*	76 - 92
CELF-P2 ELI	65	-2.33**	58 - 72	77	-1.50*	70 - 84
CELF-P2 RLI	87	-0.87	79 - 95	75	-1.67*	67 - 83
Functional/Ongoing Difficulties						
CCC-2	61	-2.60**	58 - 70	99	-0.07	93 - 105
ENNI Story Grammar A1		1.19			-0.97	
ENNI Story Grammar A3		0.15			-0.86	
ENNI MLU in Morphemes		-0.74			-2.39**	
ENNI SC Mean Verbs per Utterance		-0.82			-2.30**	
ENNI LD Moving Average NDW		-2.09**			-0.40	
ENNI % Utterance with Errors		0.24			-0.90	
Ascendant Skills/Predictors						
TELL Print Understanding	63	<-2**		79	-1.4*	
TELL Letter Knowledge	78	-1.47*		85	-1*	
TELL Oral Vocabulary	65	<-2**		96	-0.27	
TELL Word Reading	83	-1.13*		83	-1.13*	
TELL OR&C Running Record	70	-2**		100	0	
TELL OR&C Comprehension	80	-1.3*		80	-1.33*	
CTOPP PA Composite	71	-1.93*		86	-0.93	
CTOPP Phonological Memory	58	<-2**				
CTOPP Rapid Symbolic Naming				104	0.27	
CTOPP Rapid Non-Symbolic Naming	70	-2**		85	-1*	
CELF-P2 Recalling Sentences*	4**			5**		

Note. Score in Recalling Sentences is a scaled score based on *M* = 10, *SD* = 3.**Z scores 1 *SD* below the mean.
** *Z*-scores 2*SD*s below the mean.

Table 2.1. Summary of Standard Scores among Language Assessment Measures II

Assessment Measure	Isla			Billy		
	Standard Score	z-score	95% CI	Standard Score	z-score	95% CI
Earlier Developing Abilities						
LSA MLU in Morphemes		-1.39*			-1.10*	
LSA SC Mean Verbs per Utterance		-0.61			-1.61*	
LSA LD Moving Average NDW		-0.17			1.41	
LSA %Utterances with Errors		-2.07**			-0.81	
Mid-range Developing Abilities						
CELF-P2 CLS	98	-0.13	90 - 107	104	0.27	96 - 112
CELF-P2 ELI	92	-0.53	85-99	96	-0.27	89 - 103
CELF-P2 RLI	85	-1.00*	77 - 93	105	0.33	97 - 113
Functional/Ongoing Difficulties						
CCC-2	93	-0.47	87 - 100	78	-1.47*	73 - 85
ENNI Story Grammar A1					0.47	
ENNI Story Grammar A3					-0.05	
ENNI MLU in Morphemes					0.56	
ENNI SC Mean Verbs per Utterance					-0.23	
ENNI LD Moving Average NDW					0.65	
ENNI % Utterance with Errors					-0.71	
Ascendant Skills/Predictors						
TELL Print Understanding	71	-1.93*		89	-0.73	
TELL Letter Knowledge	78	-1.47*		80	-1.33*	
TELL Oral Vocabulary	102	0.13		80	-1.33*	
TELL Word Reading	75	-1.67*		100	0	
TELL OR&C Running Record	<70	<-2**		102	0.13	
TELL OR&C Comprehension	<57	<-2**		74	-1.73*	
CTOPP PA Composite	60	<-2**		92	-0.53	
CTOPP Phonological Memory	82	-1.2*		85	-1*	
CTOPP Rapid Symbolic Naming						
CTOPP Rapid Non-Symbolic Naming	92	-0.53		95	-0.3	
CELF-P2 Recalling Sentences	5**			9		

Note. Score in Recalling Sentences is a scaled score based on *M* = 10, *SD* = 3. *Z scores 1 *SD* below the mean. **Z-scores 2*SD*s below the mean.

Table 3.0 Above/Below Cutoff Table of Assessments at **1 *SD*** Below the Mean

	Language Assessment Measures	Ava	Carter	Isla	Billy	Below Cutoff Count	% Below Cutoff
	Earlier-Developing Abilities						
1	LSA MLU in Morphemes	-	-	-	-	4/4	100%
2	LSA SC Mean Verbs per Utterance	-	-	+	-	3/4	75%
3	LSA LD Moving Average NDW	-	+	+	+	1/4	25%
4	LSA %Utterances with Errors	+	+	+	+	0/4	0%
	Mid-range Developing Abilities						
5	CELF-P2 CLS	-	-	+	+	2/4	50%
6	CELF-P2 ELI	-	-	+	+	2/4	50%
7	CELF-P2 RLI	+	-	-	+	2/4	50%
	Functional/Ongoing Difficulties						
8	CCC-2	-	+	+	-	2/4	50%
9	ENNI Story Grammar	+	+		+	0/3	0%
10	ENNI MLU in Morphemes	+	-		+	1/3	33%
11	ENNI Syntactic Complexity (Mean Verbs per Utterance)	+	-		+	1/3	33%
12	ENNI LD Moving Average NDW	-	+		+	1/3	33%
13	ENNI % Utterance with Errors	+	+		+	0/3	0%
	Ascendant Skills						
14	TELL Print Understanding	-	-	-	+	3/4	75%
15	TELL Letter Knowledge	-	+	-	-	3/4	75%
16	TELL Oral Vocabulary	-	+	+	-	2/4	50%
17	TELL Word Reading	-	-	-	+	3/4	75%
18	TELL OR&C Running Record	-	+	-	+	2/4	50%
19	TELL OR&C Comprehension	-	-	-	-	4/4	100%
20	CTOPP PA Composite	-	+	-	+	2/4	50%
21	CTOPP Phonological Memory	-	-	-	+	3/4	75%
22	CTOPP Rapid Non-Symbolic Naming	-	+	+	+	1/4	25%
23	CELF-P2 Recalling Sentences	-	-	-	+	3/4	75%
	Below Cutoff Total	17	12	10	6		
	Total Tests	23	23	18	23		
	Below Cutoff percentage	74%	52%	55%	26%		
	Above Cutoff percentage	26%	48%	45%	74%		

Note. "+" denotes scores that were *above* the given cutoff point (i.e. -1*SD*); "-" denotes scores that were *below* the cutoff. Below Cutoff counts details the number of scores below cutoff across participants. Below Cutoff totals, on the other hand, are the number of scores below cutoff across the different language assessment measures.

Table 4.0 Above/Below Cutoff Table of Assessments at **2 *SD*s** Below the Mean

	Language Assessment Measures	Ava	Carter	Isla	Billy	Below Cutoff Count	% Below Cutoff
	Earlier-Developing Abilities						
1	LSA MLU in Morphemes	-	-	+	+	2/4	50%
2	LSA SC Mean Verbs per Utterance	-	+	+	+	1/4	25%
3	LSA LD Moving Average NDW	+	+	+	+	0	0%
4	LSA %Utterances with Errors	+	+	+	+	0/4	0%
	Mid-range Developing Abilities						
5	CELF-P2 CLS	+	+	+	+	0	0%
6	CELF-P2 ELI	-	+	+	+	1/4	25%
7	CELF- P2 RLI	+	+	+	+	0	0%
	Functional/Ongoing Difficulties						
8	CCC-2	-	+	+	+	1/4	25%
9	ENNI Story Grammar	+	+		+	0	0%
10	ENNI MLU in Morphemes	+	-		+	1/3	33%
11	ENNI SC Mean Verbs per Utterance	+	-		+	1/3	33%
12	ENNI LD Moving Average NDW	-	+		+	1/3	33%
13	ENNI % Utterance with Errors	+	+		+	0	0%
	Ascendant Skills						
14	TELL Print Understanding	-	+	+	+	1/4	25%
15	TELL Letter Knowledge	+	+	+	+	0	0%
16	TELL Oral Vocabulary	-	+	+	+	1/4	25%
17	TELL Word Reading	+	+	+	+	0	0%
18	TELL OR&C Running Record	+	+	-	+	1/4	25%
19	TELL OR&C Comprehension	+	+	-	+	1/4	25%
20	CTOPP PA Composite	+	+	-	+	1/4	25%
21	CTOPP Phonological Memory	+	+	+	+	0	0%
22	CTOPP Rapid Non-Symbolic Naming	+	+	+	+	0	0%
23	CELF-P2 Recalling Sentences	-	-	-	+	3/4	75%
	Below Cutoff Total	7	4	4	0		
	Total Tests	23	23	18	23		
	Below Cutoff percentage	30%	17%	22%	0%		

| Above Cutoff percentage | 70% | 83% | 78% | 100% |

Note. "+" denotes scores that were *above* the given cutoff point (i.e. -2*SD*); "-" denotes scores that were *below* the cutoff. Below Cutoff counts detailed the number of scores below cutoff across participants. Below Cutoff totals, on the other hand, are the number of scores below cutoff across the different language assessment measures.

Individual Comparisons

In Tables 3.0 and 4.0, red shading indicates scores that were below cutoff at -1 *SD* (Table 2) or -2 *SD* (Table 3.0); green shading indicates scores that were above cutoff.

Ava

Omnibus Standardized Assessment and Traditional LSA Measures. If we take - 1*SD* as a cutoff suggesting ongoing language concerns, Ava's scores on the CELF-P2, the Core Language as well as the ELI were in the range of concern. At -2*SD*, the ELI suggested concern, as well as three LSA measures with the exception of grammatical accuracy. The CLS painted a different picture in that it did not flag Ava in the range for concern.

Functional Language Assessments. In both cutoff points (i.e. 1 *SD* and 2 *SD*s below the mean), the CCC-2 flagged ongoing concern. This result concurred with her scores in the common language assessments mentioned above.

Ascendant Skills Assessments. At 1 *SD* below the mean, Ava consistently scored within the range of concern when looking at the TELL and the CTOPP-2—measures that look at the ascendant skills. This was also the case when we consider her score in a predictor task like the sentence repetition measure in the CELF-P2. On the other hand, at -2*SD*, Ava was flagged for concern in 2 subtests of the TELL out of 6, and in the CELF-P2 sentence repetition task, but not on the CTOPP measures.

Overall Results. Overall, as can be seen in Tables 3.0 and 4.0, there is a high degree of consistency among the expanded set of measures in Ava's case such that in both cutoff points,

Ava would be likely to get referred for ongoing concerns in the common language assessments (i.e. CELF-P2, LSA measures), in the CCC-2 as well as the known predictors of ongoing concern (i.e. ascendant skills, and sentence repetition).

Carter

Omnibus Standardized Assessment and Traditional LSA Measures. As can be seen in Table 3.0, at 1 *SD* below the mean Carter was flagged for concern in 5 out of 7 of the typical assessments used in kindergarten (i.e. LSA measures and CELF-P2). However, at -2*SD*, only the MLU flagged Carter for ongoing concern; he scored above -2*SD*s in the other traditional language assessment measures.

Functional Language Assessments. At 1*SD*, Carter scored within the expected range of communicative behaviours for his age on the CCC-2. In addition, he was also in the average range in 50% of the micro- and macrostructure analysis of his ENNI samples. At -2*SD*, Carter was only flagged for ongoing concern in half of the ENNI microstructure measures (i.e. MLU and sentence complexity), and was not flagged for a low score in story grammar.

Ascendant Skills Assessments. As can be seen in Table 3.0, at 1 *SD* below the mean, Carter scored in the range of concern in half of the TELL subtests and in 1 out of 3 CTOPP-2 subtests. At 2*SD*s below the mean, he was not flagged for concern in any of the aforementioned subtests. Interestingly, his sentence repetition score also indicated ongoing concern both at 1 and 2 *SD*s below the mean.

Overall Results. When we consider a broad range of assessments, Carter would likely be flagged for ongoing needs by roughly half of the measures at 1*SD* below the mean. This likelihood sharply decreases to 17% at the -2*SD* cutoff. Another way to look at it is depending on the assessments that his clinician used for assessment, the probability of Carter being referred for

further evaluation is near to chance at the -1*SD* cutoff. Altogether, in Carter's scenario, the

concordance rate among the assessments was not as consistent as what was seen in Ava's results.

Isla

Omnibus Standardized Assessment and Traditional LSA Measures. At 1 *SD* below the

mean, the CELF-P2 RLI scored Isla in the range of concern. In addition, MLU also reported

ongoing concern. Isla's standard score in the RLI was an 85 (see Table 2.1), and while this was

shown in the table as "red," this score indicated borderline concern as per the CELF-P2

interpretation guidelines. At 2*SD*s below the mean, she was not flagged for concern in any of

these assessments.

Functional Language Assessments. As can be seen in Tables 3.0 and 4.0, Isla's CCC-2

standard score showed no concern at either cutoff point. ENNI scores were not calculated as Isla

was not able to complete the task even with the appropriate prompting. When asked to say what

was happening in both the training story and first story, Isla said, "orange" and named one or two

objects in the pictures. After a few trials, she refused to continue with the task, and would ask to

play with toys in the room instead. She also circled back to describing what playing outside was

like and needed to be redirected back to the task.

Ascendant Skills Assessments. Looking at measures tapping into ascendant skills, at -1*SD*, Isla

scored below cutoff in all but one TELL subtest and in two out of three CTOPP-2 subtests.

Moreover, she also scored in the range of concern for the sentence repetition task. At -2*SD*, Isla

scored below cutoff in the Oral Reading and Comprehension subtests of the TELL. Isla scored

below cutoff in the PA composite in the CTOPP-2. As well, she was still in the range of concern

for sentence repetition at 2 *SD*s below the mean.

Overall Results. Depending on which measure(s) a clinician has available or if one were to only have access to a limited range of scores, the likelihood of being flagged at -1*SD* would be just above chance (i.e. 55%) for Isla. This likelihood drops to 22% at -2*SD*. In this scenario, Isla would likely not be eligible for further support if only the traditional assessments were considered. A different picture emerges when the ascendant skill measures are considered; they reveal that Isla still has ongoing concerns that point to the likelihood of struggling with academic language and literacy in the school years. Altogether, Isla's results highlight the possibility that there is a degree of discordance that can exist with an expanded set of language assessment measures. Even though we were not able to obtain scores from the ENNI, it was evident from both from the Oral Reading tasks of the TELL and the training story of the ENNI that Isla had difficulty with narratives.

Billy

Omnibus Standardized Assessment and Traditional LSA Measures. At 1*SD* below the mean, only half of the LSA measures (i.e. MLU, sentence complexity) reported ongoing concern for Billy. He scored within the typical range expected for his age in all the CELF-P2 indices. At 2*SD*s below the mean, none of the 7 commonly used language assessment measures scored Billy below cutoff.

Functional Language Assessments. As can be seen in Table 3.0, at -1*SD* the CCC-2 reported ongoing concern for Billy. However, he was not flagged in any of the ENNI measures. Billy scored within the typical range expected for his age at -2*SD* in all the functional language assessment measures (see Table 3.0).

Ascendant Skills Assessments. In measures that looked at ascendant skills, at the -1*SD* cutoff, Billy scored below cutoff in half of the TELL subtests, and in none the CTOPP-2 composites. At

the -2*SD* cutoff, Billy scored above cutoff in all ascendant skill measures and in sentence

repetition.

Overall Results. When considering the broad range of assessments, at 1*SD* below the mean,

Billy would likely be reported for ongoing concern only 26% of the time, suggesting that he will

not be as likely to be flagged for further monitoring as the other three participants. This

probability declines to 0% at 2*SD*s below the mean.

Group-Wise Examination of Different Measures

        In this section, we looked at the assessments in two ways: (1) the rate at which each

measure flagged a child for concern at kindergarten age, and (2) how the areas of concern

differed between the two cut-off points.

Omnibus Standardized Assessment and Traditional LSA Measures

        As seen in Table 3.0, the CELF-P2 ELI and CLS flagged for concern 50% of the time

among four participants at 1 *SD* below the mean. These percentages decreased to 25% and 0%

respectively when looking at 2 *SD*s below the mean as can be seen in Table 3.0. When looking at

LSA measures at the -1*SD* cutoff, MLU revealed ongoing concern in all four participants.

Sentence complexity revealed concern 75% of the time. Lexical diversity given by the moving

average NDW, on the other hand, had a rate of 25%. Grammatical accuracy did not report any

ongoing concern in any participants (0%). At the -2*SD* cutoff, MLU was likely to report ongoing

concern in half the participants while sentence complexity only flagged concern in one

participant.

        On the whole, these results indicate that at 1 *SD* below the mean, both the CELF-P2, and

some LSA measures, particularly MLU and sentence complexity, were more likely to report

ongoing concern among the four participants in this category of commonly used assessments in kindergarten (see Charest et al., 2019b).

Assessments of Functional Language Use

The results from the parent report were mixed: only half of the children were reported for concern in the CCC-2 at 1SD below the mean. The likelihood of being in the area of concern among these individuals dropped to 1 of 4 children at 2 SDs below the mean. Interestingly, the results from the ENNI show that none of the three children scored in the range of concern at either cutoff point in the case of the macrostructure analysis (i.e. story grammar). In regard to the microstructure analyses, MLU, sentence complexity and lexical diversity indicated concern only 33% of the time at both cutoff points. None of the three children with whom we were able to complete the ENNI were flagged for concern in the grammatical accuracy analysis of the ENNI. It should be emphasized, however, that the missing ENNI data is due to one child being unable to complete the task at all.

Measures of Ascendant Skills and Sentence Repetition

The results from Table 3.0 indicate that the TELL and CTOPP subtests, measures that look at emergent literacy skills, reported ongoing concerns more consistently. In particular, most of the TELL subtests, and the CTOPP-2 Phonological Memory subtest flagged children in the area of concern 75% of the time at the -1SD cutoff. These probabilities dropped to 25% of the time in 3/6 TELL subtests and in 1 out of 3 CTOPP-2 subtests at the -2 SD cutoff.

In the case of sentence repetition, as can be seen in Tables 3.0 and 4.0, the Recalling Sentences subtests flagged concern 75% of the time in both cutoff points.

Severity of Language Disorder

As can be seen in Tables 2.0, 2.1 and 3.0, the rate of being placed in the area of concern in every child decreased as the cutoff moved from one to two *SD*s below the mean. Each participant was more likely to be flagged for concern at the -1*SD* cutoff. It is worth noting that these results differ from the severity of all four children before intake. In particular, they all received a "severe language delay" diagnosis prior to kindergarten.

Criteria-Based Decision-Making

Given the results, an important question from a clinical perspective was to consider quantitatively which results would warrant further evaluation and which ones would potentially be deemed as having "recovered". As an exercise, we adapted and modified the criteria for outcomes as described by Bishop and Edmundson (1987). Bishop and Edmundson (1987) considered a child as having "good" outcome if most of their scores on the different language assessment measures fell within the satisfactory range (i.e. above 10th percentile): that is, the child would not have any score in the "impaired range" (i.e. 3rd percentile or below), and would have "no more than one score below the satisfactory range (p.161)" (i.e. between 3rd to 10th percentile). Here, instead of using the same ranges, we looked at how the participants can be differentiated according to cutoff points adopted in this study. To have a "recovered" outcome, the children must have no scores below -2*SD*s, and only have one score below -1*SD* across language assessments. Rather than counting each of the 25 different scores separately, we established criteria for each of the 7 groups of assessments. For the conversational LSA measures, a child's performance was considered in the range of concern if at least 2 LSA measures fell in the range of concern. That is, if the child had at least 2 LSA scores below -1 *SD,* they were scored as having an "overall" LSA in the concern range at -1 *SD*. To be considered

"below" -2 *SD*, a child would need to have at least 2/4 of the LSA scores more than 2 *SD*s below the mean. For the CELF-P2, and the CCC-2, we considered each composite score separately. For the ENNI, we counted performance as in the range of concern for -1 or -2 *SD* if 2/5 micro- and macrostructure analyses measures were below cutoff. For the TELL, to be indicative of concern, a child must score below cutoff at either cutoff points in at least 2 out of the 6 subtests. For the CTOPP-2, a child would need to score below cutoff in at least 1 out of 3 subtests at either cutoff points. Finally, the sentence repetition subtest was also considered on its own. By this definition, when we look at both the typical assessment measures (i.e. CELF-P2, LSA measures) and the expanded set of language assessment measures, none of the participants would be deemed as having "recovered" from earlier language difficulties— including Billy, whose scores were all above -2*SD*s, but who had more than one score below -1*SD*.

Discussion

In the current study, we looked at the consistency among the assessment materials that clinicians use in kindergarten. The main research question that guided our investigation was whether or not recovery would be seen in kindergarten for children previously diagnosed with language difficulties consistent with or suggesting DLD, and the extent to which systematic differences in outcomes are seen among different measures. We followed a case series approach, looking at four children who were diagnosed with "severe expressive language delay" prior to kindergarten, two of which had identified receptive language delays as well. Six assessments targeting different dimensions of language were administered on each child, and the results were analyzed in two ways: (1) how likely each participant will be flagged for further evaluation given their assessment results, and (2) how frequently each measure indicated concern across the

participants. Several findings of interest emerged from the analysis and they will be discussed

accordingly.

Recovery in Kindergarten

Previous studies have suggested that recovery from earlier language difficulties is

possible (Bishop and Edmundson, 1987; Stothard et al., 1998). However, whether this recovery

is true or illusory still remains a question for many researchers, and a concern for clinicians in

the field (Charest et al., 2019a; Charest et al., 2019b; Scarborough & Dobrich, 1990). In this

study, we found some patterns from our participants that may point to recovery from language

difficulties in kindergarten.

From the perspective of the commonly used assessment tools in practice (i.e. CELF-P2,

traditional LSA measures), two of the four participants—Ava and Carter— obtained scores that

continue to suggest the presence of a language disorder, albeit of a milder severity than indicated

in their earlier assessments. The other two participants—Isla and Billy— scored in the area of no

concern. While we only had four participants, it is interesting that "recovery" in this sense

occurred in half of the children in this study. This is similar to the recovery rate (i.e. 44% out of

the 68) that Bishop and Edmundson (1987) observed in their study at age 5 ½.

The two latter cases point to two instances of possible "recovery." Billy scored in the

typical range for their age across the majority of measures— suggesting that he might fall within

the category of a possible "true" recovery. The only way to confirm this outcome would be to

collect data when he is older and see whether or not Billy will have residual language

difficulties. On the other hand, Isla was consistently flagged for concern in the emergent skills

assessments, thereby suggesting that her scenario could present a case of "illusory" recovery,

depending on the structure of the language assessment, as she scored within the average range on the majority (i.e. 5 out of 7) of the common measures used in kindergarten.

Looking at recovery according to severity of language disorder, it is striking that all four participants were less likely to be in a range of concern at 2*SD*s below the mean, which is considered 'severe' in many assessments and jurisdictions (Semel et al., 2003; Spaulding et al., 2006). With the exception of one child who scored in the average range in 68% of the total measures used, all participants scored within the average range in most of the assessments at -2*SD*s. This finding is consistent with several studies which found that language difficulties at kindergarten tend to be less severe and more domain-specific in children with DLD (Aram & Nation, 1975; Bishop & Edmundson, 1987; Scarborough & Dobrich, 1990). Similar to Scarborough and Dobrich (1990), it is entirely possible that intervention services helped remediate language difficulties as all four participants received intervention services in their respective preschool programs, and one still continued receiving S-LP services at the time of testing. However, there is also merit in considering the fact that this change could be explained by the nature of the language skills being measured— primarily because of the difference we see in how these measures differed in flagging concern according to the skills that they look at.

Consistency among Assessment Tools

In our study of the four children, we found that degrees of concordance and discordance occurred among an expanded set of language assessment tools. In particular, there was high agreement for both Ava and Billy wherein the tools scored them below and above -1*SD*, respectively, 74% of the time. On the other hand, the agreement in scoring below cutoff at -1*SD* in Carter's and Isla's case were both just above chance. These findings highlight that when most assessments are concordant, it may be easier to decide whether a child needs support or not.

However, in a case where they do not agree, it may be more challenging to identify which tools should be given more weight when making prognostic decisions.

Different from the results by Scarborough and Dobrich (1990), our study showed that at age 5, not all of our participants scored within the range expected for their age in language sample measures. This indicates that some of the language measures we used were able to capture ongoing difficulties that these children had at kindergarten age. In particular, some LSA measures in addition to known predictors of ongoing concern in the literature (i.e. emergent literacy skills, sentence repetition) were more likely to flag a child for concern than other language assessment tools. On the contrary, the omnibus standardized assessment measures along with the functional language tools (i.e. parent report, narratives) reported concern in only half of the children. These results underscore how it is important to use more than one assessment when making decisions regarding a child with DLD, especially in those cases in which children are more likely to score within the average range in the common assessments used in kindergarten (Bishop & Macdonald, 2009).

Language Sample Measures

The microstructure analyses of the conversational and narrative samples revealed differences in scoring children in an area of concern. Particularly, we found that some LSA measures in conversation flagged children more often than in narratives.

According to extant research, morphology and syntax are considered hallmark areas of challenge in language disorders (Domsch et al., 2012; Leonard et al., 2014). For conversational language sample measures, sentence complexity and MLU in morphemes were more likely to show ongoing difficulties in conversation in most of the participants. This result is not surprising given the existing evidence suggesting that MLU is linked to the development of complex

sentences (Tyack & Gottsleben, 1986), despite previous findings that reported plateau effects after preschool age (Bishop & Edmundson, 1987; Scarborough & Dobrich, 1990). Complex sentences are known to develop during the preschool years and continue well into school age in TD children; however, it is said to occur on a later time course in children with DLD (Arndt & Schuele, 2013; Charest et al., 2019a; Guo et al., 2020). For the children in this study, both length and complexity of their sentence showed important differences from their TD peers.

A surprising result was the fact that both macro- and microstructure analyses in the narrative tasks were less likely to indicate concern among our participants. Several studies have pointed to narratives as a reliable predictor of language concerns (Bishop & Edmundson, 1987; Schneider et al., 2006; Stothard et al., 1998). Indeed, the findings of Guo and Schneider (2016) as well as Guo et al. (2020) specifically show measures of clausal density and grammatical accuracy on the ENNI to differ between children with DLD and their TD peers. Perhaps an explanation for why we did not find a difference is the testing age for our participants; Guo and Schneider (2016) found strong diagnostic accuracy for percent grammatical errors at 6 years of age in the narrative samples of children with DLD. The participants in this study were all 5 years old. Perhaps the challenges might be more apparent at a later age.

It is important to note that the results of the ENNI should be taken with caution. We only had data from three out of the four participants. Isla, even with the appropriate prompting, was not able to produce a narrative with the training story as well as the first story. Isla also had difficulties in the oral reading task in the TELL. It is very possible that narrative production is something that Isla also struggles with, and this might also change how likely the ENNI reports concern if all children were included in the analysis.

Emergent Literacy Skills and Sentence Repetition

Our results are consistent with existing research which indicate that both emergent literacy skills and sentence repetition are reliable indicators of ongoing concern (Archibald & Joanisse, 2009; Conti-Ramsden et al., 2001; Everitt et al., 2013; Milburn et al., 2017; Scarborough, 2009).

In the case of sentence repetition, it has been shown that the Recalling Sentences subtest of the CELF-P2 was more likely to predict persistent expressive language delays (Everitt et al., 2013). Because 3 out of 4 children scored below cutoff in this assessment tool, it is possible that there are still underlying difficulties in verbal short-term memory that these children are still struggling with, despite some of them scoring either above cutoff at -1*SD* or -2*SD*s in the CELF-P2 CLS and ELI. It is interesting to note that while this task is not functional in nature, it may be an important marker to alert that functional challenges may show up as in the case of Isla who scored below cutoff in this task but not in the other indices of the CELF-P2. On the other hand, it could also provide consistency when a child is showing some form of recovery as in Billy who scored above cutoff in the task as well as in most language assessment measures.

In regard to emergent literacy skills, it has been cited in the literature that phonological awareness and letter knowledge are predictive of later reading ability (Milburn et al., 2017; Scarborough, 2009). If it is the case that some children already exhibit difficulties in these areas, it would be helpful to identify them for services earlier to remediate current concerns and possibly prevent later difficulties in complex language tasks (e.g. reading, writing) required in a school setting.

Clinical Decision-Making

From the evaluation of six different language assessment tools used in kindergarten, three points became clear: (1) recovery from language difficulties can potentially occur in kindergarten; (2) a child with an "illusory" recovery profile should not be considered to have 'recovered'; (3) lastly (and perhaps the most relevant), the type of assessment tools and cutoff points that we choose *matter* and have an impact on which children are likely to receive support.

To that end, an important question to ask is if there is a criterion that we can adopt when making decisions based on discrepant results. Does this mean that if a child has at least one low score, it would warrant further assessment? Should we require at least 2 — or more — low scores? And from what measures? Additionally, what are the implications for clinicians who have increasingly large caseloads and limited time to administer an expanded set of language assessment measures?

Depending on how these questions are answered, the outcomes for each child will be different. A clear-cut distinction can be hard to make in this instance, but it is still possible to make judgements according to the degree of concern that language assessment tools present depending on what is available to a clinician. In our study, we found that MLU in morphemes, sentence complexity, emergent literacy skills, and sentence repetition were tools that were more likely to show ongoing concern in our participants. For children like Isla, we must consider using other assessments to identify ongoing language concerns in areas like emergent literacy skills and verbal short-term memory to complement standardized assessments and language samples. On the other hand, with students like Billy, who appear to have recovered in most language tasks in kindergarten, we must look at considering a long-term criterion of what constitutes "success in language/ literacy skills" over a period of time beyond the kindergarten years.

Another consideration is to be mindful of recovery in the severity of language difficulties in kindergarten age. In our study, we found that all children, except for Ava who might be flagged for concern at 2 *SD*s below the mean, scored mostly between the -1 *SD* and -2 *SD* boundaries. That being said, if it is the case that many jurisdictions only provide funding supports when children score in the severe range (i.e. 2 SDs below the mean; Spaulding et al., 2006), then potentially only 1 out of the 4 children in our study would continue to receive S-LP intervention for their language concerns or maybe even none depending on the number of low scores required in a specific jurisdiction. Despite their 'mild' difficulties, it is important to think about what the potential academic outcomes might be for children who have struggled with language previously, but whose current scores are not quite low enough to qualify them for services.

Children with DLD are an underserved population, and as a field, we need to do more to help children with DLD become more visible (McGregor, 2020). One step to realizing this is to share the information that, essentially, passing a test does not always indicate that a child should be released from services. When looking at the short-term trajectories from identification, it is evident in our study that there are some children who continue to struggle in different aspects of language even as their early intervention funding for S-LP services wrap up in kindergarten in a province like Alberta. As well, there is still uncertainty whether *recovery* for some is long-term or if it is only specific to certain areas of language. What we do know from existing studies is that there are long-term consequences for children who have [had] DLD (Bishop et al., 2017; Dockrell & Hurry, 2018; Eadie et al. 2018; Stothard et al. 1990) such as increased risk of reading and academic difficulties in the school years. This study as well as existing literature emphasize the need for extreme caution before concluding that ongoing support is *not* needed.

Limitations

There are several important limitations of the present study that should be borne in mind. The first is the small sample size. Given that we only had four children, our sample size was not large enough to complete group-based, statistical comparisons group differences among assessments. However, we were able to collect a wealth of information on how to arrive at a decision for further support given a set of language assessment tools. The second is the fact that, without longitudinal data, interpretation of normal-range scores as 'true' or 'illusory' can only be speculative. As well, even if recovery was apparent, there is no way of knowing whether this would be maintained in the absence of intervention— which has also been expressed as a concern by clinicians (Charest et al., 2019b).

Additionally, while we have seen a snapshot of what happens in kindergarten for children with previously identified language difficulties, there are still questions that remain to be addressed. One is the likelihood of children with DLD getting identified during school entry. Most of the children in our study had an accompanying speech sound disorder at the time of first identification and continued to show speech sound errors at the time of assessment in this study. It is possible that speech may have affected our assessments. Our attempts to recruit children with language difficulties gave us children with both speech and language difficulties. This is in line with reports that these profiles of children are more likely to be on caseloads than children who only have specific language needs (McGregor, 2020). That being said, it is important to look at what happens in the identification process, and find ways to ensure that other educational professionals (e.g. teachers, educational assistants, learning coaches) also have an idea of what to look for in children who might have DLD—not only asking questions about speech intelligibility but also including questions on how children in their caseload use language. Children with

language difficulties who do not have speech issues need to be identified given that language predicts reading outcomes (Milburn et al., 2017; Scarborough & Dobrich, 1990).

Conclusion

The goal of this study was to shed light on the possibility of recovery from DLD in kindergarten. Using a case series design, we evaluated six different language assessments that were administered to four five-year-old children who were identified with DLD prior to starting kindergarten. The use of a case series approach here has emphasized several points that would not have potentially emerged in an aggregate analysis of data: in particular, that there are profiles of children that may present recovery at kindergarten age in children diagnosed with earlier language difficulties, and that severity of language difficulties decreases at kindergarten age. We were able to further explore the different scenarios of decisions that likely can happen once their S-LP services wrap up in kindergarten. By looking at these different profiles of children with DLD coming from four different classrooms, we hope that clinicians, and researchers start to re-evaluate the current assessment tools and practices that they use in kindergarten.

In closing, much of the work that needs to be done falls on the ways in which researchers, and clinicians, can better identify, support children with DLD throughout the lifespan, and advocate for their services. McGregor (2020) is right: "too many children with DLD are not identified; too many are likely to be missed (p.10)." As a field we must do better; we must collaborate and share information with different organizations, education professionals and with families that we serve (McGregor, 2020). This way, we can collectively work towards an equitable path where no child gets truly left behind.

References

Anderson, S.A.S, Hawes, D.J., & Snow, P. C. (2016). Language impairments among youth offenders: A systematic review, *Children and Youth Services Review*, *65*, 195-203. https://doi.org/10.1016/j.childyouth.2016.04.004.

Archibald, L. M. D., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research, 52,* 899–914. doi:10.1044/1092-4388(2009/08-0099)

Bishop, D. V. M. (2006), *CCC-2: Children's communication checklist-2, United States edition, manual.* San Antonio, TX: Pearson.

Bishop, D. V. M., & Adams, C. V. (1990). A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. *Journal of Child Psychology & Psychiatry*, *31*(7), 1027–1050.

Bishop, D. V. M., & Edmundson, A. (1987). Language-impaired 4-year-olds: distinguishing transient from persistent impairment. *Journal of Speech & Hearing Disorders*, *52*, 156–173.

Bishop, D. V. M., & McDonald, D. (2009). Identifying language impairment in children: combining language test scores with parental report. *International Journal of Language & Communication Disorders*, *5*, 600.

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

Bishop D. V. M., Snowling, M. J., Thompson P. A., Greenhalgh, T., & CATALISE consortium.

(2016). CATALISE: A multinational and multidisciplinary delphi consensus study

Identifying language impairments in children. *PLoS ONE, 11,* 1–26.

doi:10.1371/journal.pone.0158753

Boudreau, D. (2005). Use of a parent questionnaire in emergent and early literacy assessment of

preschool children. *Language, Speech & Hearing Services in Schools*, *36*(1), 33–47.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future

reading difficulties in kindergarten children: A research-based model and its clinical

implementation.*Language, Speech, and Hearing Services in Schools, 32,* 38–50.

doi:10.1044/0161-1461(2001/004)

Catts, H. W., Nielsen, D. C., Bridges, M. S., & Liu, Y.-S. (2016). Early identification of reading

comprehension difficulties. *Journal of Learning Disabilities*, *49*(5), 451.

Charest, M., Borger, P., Chan, C., Sanders, K., Yip, B., McFarlane, L.A., & Schneider, P.

(2019a). Assessment, diagnosis, and recovery from language disorder at kindergarten

age: A survey of clinicians. *Canadian Journal of Speech Language Pathology and

Audiology, 43(*1), 1-18.

Charest, M., Borger, P., Chan, C., Sanders, K., Yip, B., McFarlane, L.A., & Schneider, P.

(2019b). 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), 1-13.

Charest, M, & Skoczylas, M. (2019) Lexical diversity vs. 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.

DOI: 10.1044/2019_AJSLP-18-0143

Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of lexical diversity in the

    narratives of children with typical language development and developmental language

    disorder. *American Journal of Speech-Language Pathology*, *29*(4), 1866 -1882.

    https://doi.org/10.1044/2020_AJSLP-19-00176

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific

    language impairment (SLI). *The Journal of Child Psychology and Psychiatry, 42,* 741-

    748. doi:10.1111/1469-7610.00770

Dale, P. S. (1996). Parent report assessment of language and communication. In K. W. Cole, P.

    S. Dale, & D. J. Thal (Eds.), *Assessment of communication and language* (pp. 161–182).

    Baltimore: Paul H. Brookes.

Dockrell, J. E., & Hurry, J. (2018). The identification of speech and language problems in

    elementary school: Diagnosis and co-occurring needs. *Research in Developmental

    Disabilities*, *81*, 52–64.

Domsch, C., Richels, C., Saldana, M., Coleman, C., Wimberly, C., & Maxwell, L. (2012).

    Narrative skill and syntactic complexity in school-age children with and without late

    language emergence. *International Journal of Language & Communication Disorders

    (Print)*, *47*(2), 197–207.

Eadie, P., Conway, L., Hallenstein, B., Mensah, F., McKean, C., & Reilly, S. (2018). Quality of

    life in children with developmental language disorder. *International Journal of Language

    & Communication Disorders*, *53*(4), 799–810.

Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language

    impairment based on grammaticality. *Language, Speech, and Hearing Services in

    Schools*, *44*(1), 20–31.

Everitt, A., Hannaford, P., & Conti-Ramsden, G. (2013). Markers for persistent specific

    expressive language delay in 3–4-year-olds. *International Journal of Language &*

    Communication Disorders, 48(5), 534-553.

Fazio, B. B., Naremore, R. C., & Connell, P. J. (1996). Tracking children from poverty at risk for

    specific language impairments: A 3-year longitudinal study. *Journal of Speech and*

    Hearing Research, *39*, 611-624.

Goffman, L., & Leonard, J. (2000). Growth of language skills in preschool children with specific

    language impairment: Implications for assessment and intervention. *American Journal of*

    Speech-Language Pathology, 9, 151–161. doi:10.1044/1058-0360.0902.151

Griffin, T. M., Hemphill, L., Camp, L., & Wolf, D. P. (2005). Oral discourse in the preschool

    years and later literacy skills. *First Language, 24,* 123-147.

Guo, L. Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances

    between 4 and 9 years of age for the Edmonton narrative norms nstrument: Reference

    data and psychometric properties. *American Journal of Speech-Language*

    Pathology, *28*(4), 1448–1462.

Guo, L. Y., & Schneider, P. (2016). Differentiating school-aged children with and without

    language impairment using tense and grammaticality measures from a narrative task.

    Journal of Speech, Language, and Hearing Research, 59, 317-329.

    doi:10.1044/2015_JSLHR-L-15-0066

Guo, L. Y., Schneider, P. & Harrison, W. (2020). Clausal density between ages 4 and 9 for the

    Edmonton narrative norms instrument: Reference data and psychometric properties.

    Language, Speech, and Hearing Services in Schools, 51(4)*,* 1-15.

    https://doi.org/10.1044/2020_LSHSS-20-00043

Hall, N. E., & Segarra, V. R. (2007). Predicting academic performance in children with language impairment: The role of parent report. *Journal of Communication Disorders*, *40*, 82–95.

Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools, 41*, 84–95. doi:10.1044/0161-1461(2009/08-0075).

Joaquin, N. (1988). *Culture and history.* Pasig City, Philippines: Anvil Publishing.

Kaufman A.S, & Kaufman N.L. (2004). *Kaufman brief intelligence test second edition*. San Antonio, TX: The Psychological Corporation.

Klee T, Stokes SF, Wong AM, Fletcher P, & Gavin WJ. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language & Hearing Research*, *47*(6), 1396–1410.

LaParo, K. M., Justice, L., Skibbe, L. E., & Pianta, R. C. (2004). Relations among maternal, child, and demographic factors and the persistence of preschool language impairment. *American Journal of Speech-Language Pathology*, *13*, 291-303. https://doi.org/10.1044/1058-0360(2004/030)

Leonard, L. B. (2014). *Children with specific language impairment* (2nd ed.). Cambridge, MA: MIT Press.

Macleod, A., Meziane, R. S., & Pesco, D. (2020) Language abilities of children with refugee backgrounds: Insights from case studies. *Applied Psycholinguistics*, 1-21. doi:10.1017/S0142716420000405

Manhardt, J. and Rescorla, L. (2002) Oral narrative skills of late talkers at ages 8 and 9. *Applied Psycholinguistics, 23,* 1–21.

Milburn, T. F., Lonigan, C. J., Allan, D. M., & Phillips, B. M. (2017). Agreement among

traditional and RTI-based definitions of reading-related learning disability with preschool

children. *Learning & Individual Differences*, *55*, 120–129.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2019). *Assessing language production using SALT*

software: A clinician's guide to language sample analysis (3rd Edition). Madison, WI:

SALT Software, LLC.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E. & Pickles, A. (2016).

The impact of nonverbal ability on prevalence and clinical presentation of language

disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry*,

57(11), 1247–1257.

Paul, R., Norbury, C., & Gosse, C. (2018). *Language disorders from infancy through*

adolescence: Listening, speaking, reading, writing, and communicating (5th ed.).

Maryland Heights, MO: Elsevier/Mosby.

Phillips, L. M., Hayward, D. V., & Norris, S. P. (2016). *Test of early language and literacy*.

Scarborough, Ontario: Nelson Publishing.

Rescorla, L. (2009) Age 17 language and reading outcomes in late-talking toddlers: support for a

dimensional perspective on language delay. *Journal of Speech, Language, and Hearing*

Research, 52, 16–30.

Scarborough, H. S. (2009). Connecting early language and literacy to later reading (dis)abilities:

Evidence, theory, and practice. In F. Fletcher-Campbell, J. Soler, & G. Reid (Eds.),

Approaching difficulties in literacy development: Assessment, pedagogy and programmes

(pp. 23-38). Thousand Oaks, CA: Sage.

Scarborough, H.S. (1998). Early identification of children at risk for reading disabilities:

Phonological awareness and some other promising predictors. In B.K. Shapiro, P.J.

Accardo, & A.J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–119). Timonium, MD: York Press.

Scarborough H. S., & Dobrich, W. (1990). Development of children with early language delay. *Journal of Speech and Hearing Research*, *33*(1), 70-83.

Schneider, P., Dubé, R. V., & Hayward, D. (2005). The Edmonton narrative norms instrument. Retrieved from University of Alberta's Faculty of Rehabilitation Medicine website: http://www.rehabresearch.ualberta.ca/enni

Schneider, P., Hayward, D., & Dubé, R. V. (2006). Storytelling from pictures using the Edmonton Narrative Norms Instrument. *Journal of Speech-Language Pathology and Audiology, 30*, 224-238.

Semel, E., Wiig, E. H., & Secord, W. A. (2004). *Clinical Evaluation of Language Fundamentals - Preschool* (2nd ed.). San Antonio, TX: Pearson.

Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences*, *21*, 1–16.

Spaulding, T. J., Hosmer, S., & Schechtman, C. (2013). Investigating the interchangeability and diagnostic utility of the PPVT-III and PPVT-IV for children with and without SLI. *International Journal of Speech-Language Pathology*, *15*(5), 453–462.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37,* 61-72. doi:10.1044/0161-1461(2006/007)

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, 407–418. doi:10.1044/ jslhr.4102.407.

Tyack, D. L., & Gottsleben, R. H. (1986). Acquisition of complex sentences. *Language, Speech, and Hearing Services in Schools, 17,* 160–174. doi:10.1044/0161- 1461.1703.160

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing–Second edition*. Austin, TX: PRO-ED.

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals* (5th ed.). Bloomington, MN: Pearson.

Winstanley, M., Webb, R.T., & Conti-Ramsden, G. (2018), More or less likely to offend? Young adults with a history of identified developmental language disorders. *International Journal of Language & Communication Disorders, 53*, 256-270. https://doi.org/10.1111/1460-6984.12339

Wong A., Klee T, Stokes S.F., Fletcher P., & Leonard L.B. (2010). Differentiating Cantonese-speaking preschool children with and without SLI using MLU and lexical diversity (D). *Journal of Speech, Language & Hearing Research*, *53*(3), 794–799.