Identifying Language Comprehension Impairment in Preschool Children

Elizabeth Skarakis-Doyle
Lynn Dempsey
Christopher Lee
University of Western Ontario, London, Ontario, Canada

Language impairment with associated comprehension deficits results in a more severely disabling condition with a poorer long-term prognosis than does expressive language impairment (Bishop & Edmundson, 1987a, 1987b; Silva, 1980; Silva, McGee, & Williams, 1983; Silva, Williams, & McGee, 1987). Children with both comprehension and expressive language impairment are at greatest risk for having continuing deficits in their teen and adult years (Johnson et al., 1999; Rutter & Mawhood, 1991; Silva, 1980; Silva et al., 1983; Silva et al., 1987; Stothard, Snowling, Bishop, Chipchase, & Kaplan, 1998; Tomblin, Freese, & Records, 1992), which impacts their social, academic, and vocational success, and hence, prohibits their full participation in society (Dockrell & Lindsay, 1998; Johnson et al., 1999). Fortunately, the prognostic outcome for such children can be improved with intervention (Bishop & Edmundson, 1987b), which makes early identification of comprehension deficits critical.

Accurate identification depends on the ability of assessment measures to discriminate children with language comprehension difficulties from those without such deficits. However, empirical investigation suggests that standardized measures of early comprehension generally do not have sufficient sensitivity and specificity to identify children with language comprehension impairment (Plante & Vance, 1994; Spaulding, Plante, & Farinella, 2006). Bishop (1997) suggested that assessment of discourse-level comprehension should be sensitive to comprehension impairment. Comprehension of discourse units is particularly vulnerable to impairment given its requisite integration of multiple sources of information (Bishop, 1997) and the slower processing rate of children with language impairment (Miller, Kail, Leonard, & Tomblin, 2001). Indeed, difficulties with narrative discourse comprehension have been shown to limit the acquisition of literacy skills (Dickinson & Smith, 1994). This highlights the need for early, accurate identification of language comprehension impairment, including discourse. This study sought to further empirically examine the validity of several measures of discourse comprehension, including two that are new, in children age 2 1/2 to 5 years. In building the evidence base for these measures, we were particularly interested in their discriminability.

Assessing Early Comprehension

Comprehension is a multifaceted construct that depends on a coalition of specific linguistic knowledge and general world knowledge, as well as on various cognitive processes. Early comprehension is commonly conceptualized as the processing of small linguistic units such as phonemes, words, and phrases (Fernald, 2002). Hence, assessment has focused on the young child’s understanding of those
units that they are expected to have mastered. Nevertheless, Thal (1991) cautioned that given the variability in young children’s performance, it is likely that multiple measures would be needed to accurately identify the comprehension component of language impairment in young children. Given its potential sensitivity to comprehension impairment, measures of discourse comprehension should be included in comprehension test batteries.

Unfortunately, the discourse level of comprehension is rarely assessed in young children. There is a paucity of valid tools that measure very young children’s oral language comprehension, particularly for discourse units such as stories. Possibly this is because little is known about the oral comprehension of larger discourse units of language before children learn to read. There is a renaissance of interest in reading comprehension and its “prehension” foundations in oral language (Pearson, 2005). Paris and Paris (2003) suggested that early narrative comprehension provides the foundation that children later use to map their understanding onto written texts. Thus, accurate assessment of discourse-level comprehension also has implications for identifying potential literacy problems.

Measuring receptive vocabulary and sentence comprehension may not be enough to assess discourse understanding. Although words and sentences are important elements that enable the comprehension of larger discourse units, it does not necessarily follow that they alone comprise a sufficient source from which to derive understanding of a story. For example, comprehension of words and sentences cannot account for one’s understanding of the connections between and across the elements of larger discourse units, such as stories. Evidence suggests that although children’s performances on comprehension measures of various levels (e.g., word and sentence) are typically related to each other (Skarakis-Doyle, Dempsey, & Lee, 2002), an individual child’s performance on one measure is not always strongly predictive of his or her performance on another measure. If, as Bishop (1997) suggested, discourse-level comprehension is particularly sensitive to impairment, and such an impairment threatens the acquisition of literacy skills, then measures of larger units of discourse such as stories are needed in addition to word- and sentence-level measures to accurately identify children with language comprehension impairment.

New Measures of Oral Discourse Comprehension

Skarakis-Doyle and colleagues developed the Joint Story Retell task (joint retell task) and the Expectancy Violation Detection task (detection task) to evaluate emerging comprehension of familiar stories in young children age 2-1/2 to 4 years (Dempsey, Jacques, Skarakis-Doyle, & Lee, 2002; Dempsey & Skarakis-Doyle, 2001; Dempsey, Skarakis-Doyle & Lockwood, 2000; Skarakis-Doyle, 2002; Skarakis-Doyle & Wootton, 1995). The joint retell task and detection task are ecologically valid measures of story comprehension that simulate the naturalistic activity of parent–child shared storybook reading. The joint retell task assesses a child’s understanding of story content within and across sentence boundaries. The detection task assesses several phases of comprehension monitoring. These procedures are based on the assumption that young children develop expectations for stories that they have heard frequently. It is further assumed that children may use their knowledge of familiar events as a scaffold for story comprehension, similar to other language learning situations (Nelson, 1996). Additionally, these procedures require little expressive language ability and measure comprehension online (i.e., during rather than after the story, as is typically done when using comprehension questions). As a result, the entire story representation need not be held in working memory, thereby reducing the memory demands placed on young children. For the joint retell task, this is accomplished using a cloze procedure whereby the examiner pauses during the story to permit the child to supply specifically chosen portions; for the detection task, this is done by altering specific components of the familiar story and then examining children’s nonverbal and verbal responses to those violations while the story is being read.

As presented elsewhere (Dempsey & Skarakis-Doyle, 2001; Milosky & Skarakis-Doyle, 2007; Skarakis-Doyle, 2002), the joint retell task and detection task have been shown to be reliable and valid measures of oral story comprehension that meet psychometric standards (American Psychological Association, 1985). In a preliminary investigation, Skarakis-Doyle and Wootton (1995) compared the performance of typically developing children and children with language impairment on the joint retell task and detection task, with results demonstrating a significantly poorer performance on both measures by the latter group. Skarakis-Doyle and Dempsey (2007) replicated these findings for the detection task in a larger study that used different stimuli. Further, they found that the children with language impairment not only performed more poorly than an age-matched control group, but also performed more poorly than a group of younger children matched for receptive language vocabulary. These group differences support Bishop’s contention that discourse-level comprehension may indeed be vulnerable to impairment. However, because validity is an evolving concept (Messick, 1995), additional evidence is required for our experimental measures in order that the nature of these preliminary group differences be understood. Of particular import is the evidence of the ability to accurately identify children with language impairment.

Classification and Measurement Considerations

Validity is a multidimensional phenomenon. Not only must a test score be shown to be a valid representation of the domain it assesses, but additionally, any action taken based on the interpretation of those scores must also be validated (Messick, 1995). Plante and Vance (1994) suggested that when a test is used for the purpose of identifying language impairment, then classification accuracy is one of several types of evidence that are required before one can conclude that a measure is valid for that clinical purpose. In their review of preschool language tests, Plante and Vance examined the sensitivity and specificity of four standardized language tests—two tests of expressive language ability, one test of receptive language ability, and one general language test. These authors proposed that language tests should meet standards of classification accuracy that reflect the social value of accurate identification and the social consequence of misidentification of language impairment. That is, in order to allocate scarce resources effectively, standards should be set sufficiently high as to minimize misidentification and unnecessary treatment yet maximize identification of those children who truly require service. Accordingly, they considered 90% and higher accuracy to reflect good classification accuracy and 80%–90% accuracy to reflect fair classification accuracy. Tests with less than 80% accuracy were considered.
unacceptable (Plante & Vance, 1994). Only one expressive test of the four examined was found to meet the standards of adequate sensitivity. More recently, Spaulding, Plante, and Farinella (2006) found that out of 43 language tests examined, only five met their criterion of 80% identification accuracy and only two of these five tests were appropriate for preschool children.

Because eligibility for treatment criteria often require that multiple tests be administered, Plante and Vance (1994) also examined combinations of standardized tests to determine whether sensitivity improved. Contrary to what might be expected, accuracy was reduced relative to that attained with single measures. Plante and Vance concluded that multiple measures do not necessarily converge to improve identification, making efforts to understand how multiple tests work together an important endeavor. These efforts may be particularly critical when considering young children’s comprehension. As noted previously, the multifaceted nature of comprehension requires assessment at multiple levels; additionally, young children’s notoriously variable performance necessitates the use of multiple measures even within levels of comprehension (Thal, 1991). Thus, additional effort to determine the sensitivity and specificity of a combination of comprehension measures is warranted when considering young children.

Given the need for discourse measures that are appropriate for young children and their potential utility as an indicator of impairment, this study had two objectives. First, we sought to further evaluate two newly developed tests of oral story comprehension—the detection task, which evaluates comprehension monitoring, and the joint retell task, a measure of primary discourse comprehension. In addition, we examined the traditional procedure used for measuring primary discourse comprehension—comprehension questions (comp questions). We did so to seek additional evidence for the validity of our experimental measures as representative of different aspects of discourse comprehension and hence, the inferences that could be legitimately made from them. Specifically, we posed the following question: “With what level of accuracy do individual measures of discourse comprehension identify group membership based on language status?” If the individual measures could be shown to validly identify the language comprehension status of young children, we could then seek further evidence for their clinical use as part of a multimeasure combination to identify comprehension impairment. Thus, our second objective was, for clinical purposes, to examine how these discourse comprehension measures functioned together. Examination of a measurement combination permits us to account for the concerns raised by Thal (1991) regarding the variability in young children’s comprehension performance, and to address the clinical reality that multiple measures may need to be employed with this age group. Thus, we investigated whether a combination consisting of several language comprehension measures could identify children with language comprehension impairment with good clinical accuracy and compared how individual children performed on the different measures of discourse comprehension.

METHOD

Participants

Children with language impairment (LI group) and children with typically developing language (TDL group) were recruited in approximate proportion to the prevalence of language impairments in the preschool population (National Institute on Deafness and Other Communication Disorders [NIDCD], 1995). Our goal in using proportionate allocation to language groups was to achieve an ecologically valid sample in which to evaluate classification accuracy. Because prevalence reports vary across studies with the region and age range of children included (Beitchman, Nair, Clegg, & Patel, 1986; Tomblin et al., 1997), we approximated the 22% rate reported in two studies of preschool children that were completed in our local area, which is Middlesex County, Ontario (Flood, 1994; Warr-Leeper, Smith, & Dunn, 2001). This prevalence rate reflects the inclusion of phonological impairment in the category of language impairment. The rate is higher than that reported in most large-scale kindergarten prevalence studies (e.g., Tomblin et al., 1997); however, the local data reflect prevalence rates for an age group that is more consistent with our cohort.

Participants were 49 children between the ages of 30 and 61 months ($M = 45.5, SD = 9.8$). Of these, 12 children (10 boys and 2 girls) had language impairment and 37 children (15 boys and 22 girls) had typically developing language. The mean age of the LI group was 45.0 months ($SD = 11.1$). The mean age of the TDL group was 46.2 months ($SD = 9.3$). Table 1 shows the age range for both cohorts of children in 6-month intervals. All children, both LI and TDL, were from middle-class families in which English was the first language of the home. All children had normally developing motor skills and no uncorrected visual or hearing impairment. According to parent report, story reading (an important element of all of the discourse comprehension procedures) was a part of every child’s regular routine. The TDL group demonstrated age-appropriate receptive vocabulary as measured by the Peabody Picture Vocabulary Test—Third Edition (PPVT-III; Dunn & Dunn, 1997) and scored above the 20th percentile for their age on the total score for the MacArthur-Bates Communicative Development Inventory—Level III (MCDI—III; Fenson et al., 2007), an extension of the well-known parent report measure for children over 30 months of age.

Assignment to the LI group was made based on the children’s status in speech and language treatment, not on the standardized test scores used by referring clinicians. All children were receiving intervention for both expressive and receptive impairment. Their speech was intelligible 80% of the time to an unfamiliar listener. Using treatment status as the basis for assigning children to the LI group presupposes that the initial identification carried out by the referring clinicians was accurate.

As argued by Dollaghan and Campbell (1998), the definition of language impairment continues to be controversial with no gold standard.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age 30–35</th>
<th>Age 36–41</th>
<th>Age 42–47</th>
<th>Age 48–53</th>
<th>Age 54–61</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDL</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>LI</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Note. The data are displayed in 6-month intervals except for the final group, which involves an 8-month interval to encompass the two oldest children.
standard for identification available. Enrollment in intervention is presumably the result of multifaceted judgments that reflect not only standardized test scores, but also parental and professional concern for a child’s ability to function in important communication contexts. Although errors in initial identification cannot be ruled out, a multifaceted clinical judgment is a credible basis for establishing language status (Dollaghan & Campbell, 1998). Furthermore, the selection of children on the basis of treatment status is more likely to reflect the nature of a typical caseload than is selection based on standardized test scores alone.

Although standardized tests were not used to establish assignment to the LI group, performance on several language tests and a nonverbal cognitive measure are provided in Table 2 as additional descriptive information. Table 2 shows that, as expected, given their status in language therapy, children in the LI group performed significantly worse than children in the TDL group on all of the standardized tests of language comprehension that were administered before the story comprehension measures. To verify that children were language impaired and did not demonstrate a general developmental disability, a measure of nonverbal cognitive ability, the Perceptual-Performance scale of the McCarthy Scales of Children’s Abilities (McCarthy, 1972), was administered as well. Similar to the mean nonverbal performance score of the LI group in the Dollaghan and Campbell (1998) study, the mean nonverbal performance score of the LI group in this study was within normal limits but was significantly lower than that of the TDL group.

In order to further verify that the LI group indeed demonstrated a comprehension deficit (i.e., that the original classification of these children as LI was correct), their performance on three of these measures was examined more closely. The three measures included the PPVT–III, the Grammatical Morpheme subtest of the Test of Auditory Comprehension of Language—III (TACL–III; Carrow-Woolfolk, 1999), and the Language Use subtest of the MCDI–III, which largely assesses understanding of linguistic concepts. Half of the LI group had either a standard score on the PPVT–III or TACL–III Grammatical Morpheme subtest that was greater than 1 SD from the test’s mean or performed below the 20th percentile for their age on the MCDI–III Language Use subtest. The other half of these children demonstrated this level of performance on two or more of these measures.

### Test Materials

A patterned children’s story, “Splish Splash” (Skarakis-Doyle & Wootton, 1995), was written specifically to be the story stimulus for the story comprehension measures. The theme of this story is the familiar routine event of bath time, which is a goal-based, caretaking routine that is typical of events for which young children are likely to have scripts (Nelson & Gruendel, 1986). The story is of a girl who tries to avoid taking a bath and who is ultimately successful. There are five episodes organized around two conflicting goals: (a) the child’s goal of avoiding the bath and (b) the mother’s goal of bathing her daughter. The story itself is told in a patterned fashion. Three of the episodes are parallel (i.e., child requesting something for her bath and mother searching for it), in which syntactic frames, some vocabulary, and a rhythmic song-like refrain are systematically repeated. The story incorporates core elements associated with bath time (e.g., soap and water) as well as optional elements to support the core (e.g., testing the temperature of the water, a rubber duckie). Use of a script-based theme and patterned format provide a memory scaffold for the young participants. The book contains a total of 298 words and is eight pages long. Each page of text is accompanied by a corresponding color illustration. A sample from the story is provided in the Appendix. Two additional modified versions of the story were used in the test phase of the study, one for the joint retell task and one for the detection task.

For the joint retell task, 10 story elements from across five element categories (three actions, three objects, two locations, one agent, and one goal) were omitted from the original text and substituted with pauses. The goal item taps children’s understanding of the mother’s explicitly stated objective of giving her daughter a bath. A cloze item was also created to tap children’s comprehension of the overall gist or theme of the story (i.e., the little girl does not want to take a bath), which is not made explicit in the text, resulting in a total of 11 cloze items. Most cloze items can be successfully completed with one or two words (e.g., soap). In the case of the goal and gist items, children can convey the meaning of the items in just a few words (e.g., don’t like bath). This modified version was 238 words in length, and as with the detection task version, resembled the original story in all other aspects. An excerpt from the story with cloze items is included in the Appendix.

### Table 2. Mean and standard deviation values for study participants on background standardized language and cognitive tests.

<table>
<thead>
<tr>
<th>Test</th>
<th>TDL</th>
<th>LI</th>
<th>Group comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>PPVT–III*</td>
<td>112.41</td>
<td>13.99</td>
<td>82.75</td>
</tr>
<tr>
<td></td>
<td>*(n = 37)</td>
<td></td>
<td>*(n = 12)</td>
</tr>
<tr>
<td>TACL–III – GM*</td>
<td>13.25</td>
<td>2.06</td>
<td>8.56</td>
</tr>
<tr>
<td></td>
<td>*(n = 32)</td>
<td></td>
<td>*(n = 9)</td>
</tr>
<tr>
<td>MCDI–III b</td>
<td>106.61</td>
<td>15.60</td>
<td>34.83</td>
</tr>
<tr>
<td></td>
<td>*(n = 36)</td>
<td></td>
<td>*(n = 12)</td>
</tr>
<tr>
<td>McCarthy*</td>
<td>58.77</td>
<td>9.77</td>
<td>39.83</td>
</tr>
<tr>
<td></td>
<td>*(n = 31)</td>
<td></td>
<td>*(n = 12)</td>
</tr>
</tbody>
</table>


---

*a* standard score, *b* total raw score.
For the detection task, eight core and optional elements were altered or violated. Following Skarakis-Doyle and Milosky (1999), three types of discourse-level violations were made to the story elements: two goal disruptions, five within-story substitutions, and one prosodic alteration, for a total of eight violations. A goal disruption is created by inserting a short, plausible event that is consistent with the story goal but that disrupts the forward progression toward that goal. These actions were not present in the original story. For example, the phrase jumped in the tub was inserted in the story in place of the original phrase ran outside to make more mud-pies. Getting into the bathtub would thwart the child’s goal of avoiding the bath. A within-story substitution is created by substituting one of the story’s actors, objects, or actions with another from elsewhere in the story. For example, Sarah, the main character, was replaced with Mother in the testing bath water sequence. The third type of violation is created by altering the content of the rhythmic refrain that occurs throughout the story. The new phrase is related to the story but contrary to the goal structure and content of the original story. The altered refrain becomes Ooey, gooey let’s make a mess…, which is contrary to the goal of taking a bath. This version of the story was 267 words in length but resembled the original story in all other aspects. Examples of an episode from the original version of the story, with examples of experimental alteration, are provided in the Appendix.

A set of 12 traditional wh-comprehension questions was also employed as a measure of story comprehension. This additional measure was included in order to evaluate the strengths and weaknesses of the experimental measures relative to the accepted standard for measuring discourse comprehension. Ten wh-questions queried the same content as the cloze version used in the joint retell task. An additional two questions were used to evaluate participants’ understanding of the gist or underlying theme of the story. Examples are provided in the Appendix.

Procedure

Children completed the experimental tasks during one session of approximately 1-1/2 hr in length or over two sessions in the event that they became fatigued. During the session(s), they completed the PPVT–III and the TAACL–III Grammatical Morphemes subtest, and a parent completed the MCDI–III (Fenson et al., 2007). These measures helped verify that children in the TDL group had age-expectated language abilities and that children in the LI group indeed had comprehension problems. The nonverbal cognitive measure was also administered at this time.

Following completion of the standardized tests, the experimenter read the unaltered stimulus story while the children followed along in the picture book. After listening to the story once, children completed the three story comprehension tests. Children always completed the joint retell or comp questions first, in counterbalanced order, and the detection task last. This prevented joint retell task and comp question performance from being hampered by exposure to the story violations presented during the detection task. Children were videotaped during these story comprehension tasks for later analysis of responses.

For the joint retell task, children were told, “Now we’re going to tell the story again. This time you can help me. This is how we’ll do it.” Then four practice cloze sentences were presented. Practice sentences were taken from the story but were not the same as the items in the joint retell task. The practice sentences were administered to ensure that all participants understood the requirements of the oral cloze task. To continue in the study, participants were required to demonstrate comprehension of the task by responding verbally to at least one of the practice sentences. All participants in this study did so. Following the practice items, the investigator introduced the experimental task by saying, “Now we’re going to tell the whole story. You can keep helping me, just like that.” Then the examiner read the joint retell altered story, with the child following along in the picture book. At each pre-established cloze point, the examiner paused and waited expectantly for the child’s response. Once a response was made, the examiner continued with the story. If the child did not respond within 5 s, the investigator repeated the words just before the cloze item and again waited for 5 s. If the child still did not respond, the investigator supplied the missing word(s) and continued with the story.

For the comp questions, children were asked the set of 12 questions following presentation of the original story (or joint retell depending on the counterbalanced order). The examiner asked the question, allowing 5 s for a response. If there was no response after 5 s, the examiner repeated the question and again waited 5 s. If there was still no response or the child indicated that he or she did not know the answer, the next question was presented until the entire set of questions had been completed. The story illustrations were not available to children during the question-asking period.

As was the case for the joint retell task, the detection task began with children completing practice items using a familiar book that the child’s parent had supplied. The examiner introduced the task as follows: “I’m going to read your story to you. But, I’m going to say some things that sound silly or out of place and I want you to catch me.” The examiner then read a portion of a familiar storybook provided by the child’s parent. Familiar elements such as the main actors and their actions were altered. If the child failed to respond either verbally or nonverbally to the alteration, the examiner encouraged the child to “catch” the next mistake. When a child detected a violation during the practice session, verbal reinforcement was given (e.g., “Good you caught me!”). Up to five practice trials were presented, but as soon as the child responded to one, the remaining practice items were discontinued. All participants in this study demonstrated understanding of the detection task by responding to a practice violation item. Upon completion of the practice items, the violated version of the story was read to the child while he or she followed along in the book. Children demonstrated that they had detected violations by responding within 5 s of their occurrence through nonverbal behaviors (e.g., change in eye gaze, facial expression, or body movement) and/or a verbal response. If the child produced a verbal protest (e.g., “No!”) in response to a violation, the examiner provided an open-ended probe to elicit further correction of the violation (e.g., “No? No what?”). The examiner did not overtly acknowledge nonverbal responses made following violations. If no response was made to a violation, no prompts were given, and the examiner continued to read the story until completed.

Scoring

The joint retell task was scored by calculating the total number of items (out of a maximum of 11) that were responded to correctly. Correct responses are those that are verbatim from the story or those that are variations of the words in the story that convey the same meaning. The comp questions were scored similarly, but with a
maximum of 12 items. Performance on the detection task was scored from the videotape using criteria established by Skarakis-Doyle (2002). Both verbal and nonverbal responses that occurred within 5 s of a violation were considered responses. Verbal responses include behaviors such as single-word protests (e.g., “No!”), repetitions with emphatic stress or rising inflection that indicated a challenge to or query of the violation (e.g., “Mom!” Or “Mom?”), and spontaneous or prompted corrections (e.g., “Not Mom, Sarah!”). As these sample responses indicate, little expressive ability is needed for the child to verbally indicate detection of a violation. Nonverbal responses included changes in eye gaze (e.g., from book to reader or vice versa), changes in facial expression (e.g., smile), and changes in body movement (e.g., head turns). For a nonverbal response to be considered a detection of a story violation, it must have occurred within 5 s of the violation and no later than the end of the phrase immediately following occurrence of the violation. Skarakis-Doyle (2002) demonstrated that these nonverbal behaviors were reliable and robust indicators of detection.

**Agreement**

A research assistant trained in scoring the joint retell task, comp questions, and detection task independently reanalyzed 30% of the data from the entire cohort for each measure. Protocols were randomly selected from each group of participants. Agreement was calculated by dividing the number of agreed-on items by the number of agreements + disagreements. One hundred percent agreement was achieved for the comp questions, 95% for the joint retell task, and 93% for the overall presence of responses to violations on the detection task. In addition, agreement was determined separately for the solely nonverbal and verbal responses, resulting in 87% for the former and 100% for the latter.

**RESULTS**

Several issues were of interest in this investigation, including (a) the extent to which the experimental discourse comprehension measures, the detection task and the joint retell task, could accurately classify participants into language status groups for purposes of examining additional evidence for their validity; and (b) the accuracy with which a coalition of comprehension measures functioned together to meet a clinical standard for identification of comprehension impairment. As an initial analysis, independent samples t tests were conducted to determine whether, on average, test scores on each of the experimental measures were significantly different in the two language status groups. The effect size of significant differences are reported using Cohen’s d (Cohen, 1988). Mean scores and standard deviations for these measures are displayed in Table 3. In effect, a measure that does not yield significant average differences between groups is unlikely to accurately classify individuals according to language status group. Average differences, however, do not guarantee an accurate classification of individuals into these groups; consequently, discriminant analyses were also performed to appraise the accuracy with which the detection task and joint retell task could each classify the language status of individual children.

On the joint retell task, the LI group performed poorer than the TDL group, t(29) = 6.26, p < .001, d = 1.28. The discriminant analysis also revealed that the joint retell task differentiated the two groups. The overall Wilk’s lambda was significant, L = .66, \( \chi^2(1, N = 47) = 18.06, p < .00, \) and 33% of the variance in joint retell task scores was accounted for by group language status. The joint retell task correctly identified the language status of 89% of participants, with stronger specificity (94%) than sensitivity (73%), as shown in Table 4. Thus, although the joint retell task demonstrated strong group differences, its ability to identify children with LI was weaker than might have been expected given those differences.

<table>
<thead>
<tr>
<th>Measure</th>
<th>TDL M</th>
<th>SD</th>
<th>LI M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint retell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 36)</td>
<td>6.36</td>
<td>2.77</td>
<td>2.18</td>
<td>1.60</td>
</tr>
<tr>
<td>Detection task</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 36)</td>
<td>5.03</td>
<td>2.32</td>
<td>.55</td>
<td>.69</td>
</tr>
<tr>
<td>Comp questions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 37)</td>
<td>6.08</td>
<td>2.97</td>
<td>.64</td>
<td>.92</td>
</tr>
</tbody>
</table>

Table 3. Mean and standard deviation values for study participants on discourse comprehension measures.

On the joint retell task, the LI group performed poorer than the TDL group, t(29) = 6.26, p < .001, d = 1.28. The discriminant analysis also revealed that the joint retell task differentiated the two groups. The overall Wilk’s lambda was significant, L = .66, \( \chi^2(1, N = 47) = 18.06, p < .00, \) and 33% of the variance in joint retell task scores was accounted for by group language status. The joint retell task correctly identified the language status of 89% of participants, with stronger specificity (94%) than sensitivity (73%), as shown in Table 4. Thus, although the joint retell task demonstrated strong group differences, its ability to identify children with LI was weaker than might have been expected given those differences.

In terms of the detection task, the LI group performed poorer than the TDL group, t(29) = 10.20, p < .001, d = 1.54. The discriminant analysis also revealed that the detection task differentiated the two groups. The overall Wilk’s lambda was significant, L = .53, \( \chi^2(1, N = 47) = 27.95, p < .00, \) and 43% of the variance in detection task scores was accounted for by group language status. Further, detection task scores as shown in Table 4 correctly identified the language status of 92% of participants overall. The detection task also achieved both a high degree of specificity and a high degree of sensitivity, accurately classifying 92% of the TDL group and 91% of the LI group.

Finally, a third traditional discourse comprehension measure, comp questions, was evaluated as a benchmark against which to evaluate the relative performance of the experimental discourse measures. As with the experimental measures, a significant group difference was found for the comp questions, t(45) = 10.20, p < .001, d = 1.51. The discriminant analysis also revealed that the comp questions differentiated the two groups. The overall Wilk’s lambda was significant, L = .56, \( \chi^2(1, N = 48) = 26.05, p < .001, \) and 93% of the variance in detection task scores was accounted for by group language status. The comp questions correctly identified the language status of 92% of participants overall. The detection task also achieved both a high degree of specificity and a high degree of sensitivity, accurately classifying 92% of the TDL group and 91% of the LI group.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint retell</td>
<td>89%</td>
<td>94%</td>
<td>73%</td>
</tr>
<tr>
<td>Detection task</td>
<td>92%</td>
<td>92%</td>
<td>91%</td>
</tr>
<tr>
<td>Comp questions</td>
<td>92%</td>
<td>92%</td>
<td>91%</td>
</tr>
<tr>
<td>Joint retell + detection task</td>
<td>96%</td>
<td>94%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 4. Percentage of accurate classification for the individual experimental discourse comprehension measures and their combinations.
The detection task was examined in combination with both the traditional comp questions and the joint retell task. Both primary comprehension measures were included in the combination because incorporating measures with differing task demands into the combination creates the possibility of assessing a broader range of children with LI, allowing in particular those with very limited expressive language abilities to be assessed. The discriminant analysis revealed that the combination of predictor measures differentiated the two groups. The overall Wilk’s lambda was significant, $\Lambda = .47$, $\chi^2(3, N = 46) = 32.02, p < .00$, with 53% of the variance in composite created by the combined scores accounted for by group language status. The classification result for this combination of measures is displayed in Table 4. When the measures were combined, overall group identification accuracy was 96%, with 94% specificity and 100% accurate identification of children with language comprehension impairment.

Our findings demonstrate that the great majority of the participants in our cohort were correctly classified. In order to better understand how these comprehension measures worked together, we further examined the children both with and without language impairment who were misidentified. Table 5 illustrates how different discourse comprehension measures either converged or diverged in the misidentification of these participants. Four children with LI were misclassified as having TDL. As shown in Table 5, only 1 of these children was misidentified by two tests, joint retell and detection together. The other 3 were misidentified by only one test, in two cases, the joint retell and, in the other, the comp questions. Six children with TDL, who were 3 years of age or younger, were misclassified as having LI. This constitutes half of all TDL children in the study who were under the age of 3. One of the misidentified children was classified as LI by all three discourse measures; the remainder were misclassified by only one test. For all of the children who were misidentified as LI by only one test, this test was either the comp questions or the detection task. None of the LI group who were 36 months of age or younger ($n = 4$) was incorrectly identified.

### DISCUSSION

The importance of identifying language comprehension impairment early and its associated challenges led us to investigate the classification accuracy of multiple measures of discourse comprehension, both primary comprehension and comprehension monitoring. Doing so permitted evidence-based examination of their validity. Of particular interest to us was the performance of the two experimental discourse measures, one of which (detection task) was intended to measure a level of comprehension (early comprehension monitoring) that typically is not assessed in young children, and the other of which (joint retell task) was intended to minimize the typical enabling factors that could interfere in measuring discourse comprehension relative to traditional comprehension questions. Further, given the potential clinical utility of multiple comprehension measures, we examined the accuracy of the combination of three measures in identifying language comprehension impairment in children with and without LI.

In our cohort of children between 30 and 61 months of age, those with and without language comprehension impairment differed significantly in their ability to perform our discourse comprehension measures. The effect sizes obtained were large by Cohen’s standards, indicating that there was little overlap between the two groups. Additional evidence was sought by further examining the discriminability of these measures, and although they varied by 10% in the amount of variance they accounted for in group status and in the percentage of children with LI identified, these findings offer support for the validity of interpreting the experimental measures as representing discourse comprehension.

### Table 5. Comparison of misidentified participants on three discourse comprehension measures.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age</th>
<th>Detection task</th>
<th>Joint retell</th>
<th>Comp questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1LI</td>
<td>45</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2LI</td>
<td>43</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3LI</td>
<td>56</td>
<td>1</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>4LI</td>
<td>61</td>
<td>1</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>1TDL</td>
<td>34</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>2TDL</td>
<td>32</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>3TDL</td>
<td>33</td>
<td>1</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>4TDL</td>
<td>37</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>5TDL</td>
<td>34</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>6TDL</td>
<td>35</td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

**Note.** * in cell indicates that the participant was misidentified by the measure.
and monitoring. We first discuss the results for the individual experimental measures to provide a clearer picture of how they performed across the broad age group of young children participating in this study and to suggest what that means for the measures’ overall validity as tests of discourse comprehension. We then proceed to discuss the clinical implications of the level of identification accuracy demonstrated by the combination of discourse measures.

The Expectancy Violation Detection Task

Our detection task assessed several phases of comprehension monitoring using an online format. This level of language comprehension usually is not evaluated in young children because it has been thought that they are not developmentally ready to engage in such tasks. However, Skarakis-Doyle (2002) demonstrated that the stimuli and format of the detection task could reveal emerging comprehension monitoring in young children, and evidence for its validity has been accumulating. Our findings offer additional support for this detection task’s validity as a measure of early comprehension monitoring by demonstrating not only group differences, but also that it accurately identified language comprehension impairment in our LI group.

The detection task is essentially a novelty detection task. Novelty detection is an early emerging process that informs executive functioning abilities such as self-monitoring and regulation (Metcalf, 1994; Skarakis-Doyle, 2002). As within other cognitive domains (e.g., infant intelligence), novelty detection procedures may be particularly good at identifying young children who are at risk for developmental impairment (Fagan, Singer, Montie, & Shepard, 1986; Thompson, Fagan, & Fuller, 1991). The online format of the task requires children to sustain attention throughout the story and to detect disruptions to their expectations. There are no pauses or specific queries to cue children as to when to respond or to support what the structure of that response should be. Further, the fact that the task requires the coordination of a coalition of information sources (like other discourse measures) and requires vigilance may contribute to its identification effectiveness. Thus, the detection task provides unique assessment information about the emerging comprehension monitoring abilities of young children with and without language impairment when used as part of a clinical assessment package for the comprehension of preschool children who are 3 years of age and older. Although the detection task was very accurate in identifying the children in our LI group under 3 years of age, it tended to somewhat overidentify children with TDL in that same age range. Although this tendency may in part reflect properties of the task itself, the variability known to be present in the early stages of language development may also have affected the results (Fenson et al., 2000). That is, the rate at which young children acquire language varies considerably, and the range of normal is wide. Some of the young children in the TDL group may not have been as advanced in their comprehension monitoring ability as their peers, obtaining lower enough scores on the detection task that they may have been misclassified as having LI. Given that the detection task seemed to overidentify children under the age of 3, a qualification on its validity with this age range is suggested. Additional research to examine the variability in performance on the detection task for children less than 3 years of age is ongoing. However, for clinical purposes, it appears that when used as a component of a package of discourse comprehension measures, this qualification may be somewhat mitigated. The clinical implications of the influence of age on detection task performance will be discussed further below.

The Joint Story Retell Task

Our second experimental measure, the joint retell task, did not identify children with LI as well as the other measures we evaluated. This measure was originally developed to assess primary discourse comprehension, as do traditional comprehension questions. Both the joint retell task and comp questions measure comprehension of story content and scaffold responses by providing a prompt of some sort for the response. However, the joint retell task was additionally intended to minimize the enabling factors of memory and expressive language abilities that can confound the interpretation of comp questions. Recall that the joint retell task is presented in an online format such that the child responds while the story is read, which should reduce memory demands. This is in contrast to traditional comp questions, which are not presented until after the entire story has been presented, and hence require that the child hold the story in memory in order to respond. Additionally, we presume that the verbal ability required to formulate a correct response is not as great with the joint retell task as it is with the comp questions. In most cases, children completing the joint retell task need supply only a single word. There was a significant group difference found by the joint retell task: It identified approximately 75% of the children in the LI group and it accounted for approximately one third of the variance distinguishing language group status. The purpose of developing the joint retell task was to control expressive language and memory demands as enabling factors in a primary discourse comprehension task that may introduce error into the interpretation of results. It appears that these controls were beneficial to very young children (those under 3-1/2 years) regardless of language status. Only 2 of the 16 children in the entire cohort who were under 3-1/2 years of age were misidentified by the joint retell task. Thus, the task does permit young children to demonstrate what they know about the story and to be identified accurately. However, it may not have sufficiently stressed the systems of at least some of the older children with LI, resulting in some missed identifications. As shown in Table 5, all of the children (n = 3) who were misidentified by the joint retell task were 43 months of age or older. Better understanding of the influence of age and stage of expressive language development on joint retell performance is required to clarify their impact on empirical tests of the measure’s validity. Although the limitations of the joint retell task that we found with our particular sample must be explored further, they must also be balanced against the potential clinical utility of the task in the alternative format that it provides. The fact that even with a relatively small sample size of children with LI, approximately 75% were identified correctly suggests that this measure is better at identifying such children than many standardized tests that are available for use with preschoolers (Spaulding et al., 2006). This justifies further work on its development and, as will be discussed shortly, an examination of its utility when combined with other measures of discourse comprehension.

Clinical Identification of Comprehension Impairments

Messick (1995) stated that not only the interpretation of test performance must be validated, but also the actions that follow from
that interpretation. Thus, recognizing that multiple measures are typically preferred when evaluating young children due minimally to their notoriously variable comprehension performance, we believed it important to examine a combination of our discourse measures to validate this clinical practice. Further, we have asserted that the multilevel nature of comprehension itself also requires multiple measures in order to measure it adequately. Thus, in order to obtain evidence of the clinical utility of our discourse comprehension measures, we evaluated the discriminative accuracy of the combination of all three. Unlike the tests evaluated by Plante and Vance (1994), the particular combination of discourse comprehension measures that we evaluated was a highly accurate indicator of language comprehension status for our cohort of children.

All of the children with LI and well over 90% of the children with TDL across our age group were accurately classified by this combination. Given that this combination met the socially determined clinical standard set by Plante and Vance, we explored how the three measures worked together to achieve this accuracy.

More often than not, the three discourse comprehension measures converged in their classification of the participants; hence, the high level of overall identification. However, each of the measures misidentified between 4 and 5 children; the detection task and comp questions more often mistakenly identified a problem in the TDL group (overidentification), and the joint retell task more often missed a problem in the LI group (underidentification). Several possibilities exist that may account for the misidentifications. First, as we discussed when considering the validity of the individual experimental measures, the misidentification may be the result of psychometric deficits in the measures themselves.

Second, there may be little or no psychometric weakness with our measures; rather, the misclassifications may reflect the well-known variability in early language development that any valid measure would necessarily reflect (Fenson et al., 2000). Previously, we discussed different ways that age might impact the classification accuracy of the detection task and the joint retell task. Given that participants in our study spanned an age period in which language undergoes arguably its greatest developmental change, considering the interplay of age and performance on these tasks is important. The more variable language abilities of children under 3 years of age may indeed be reflected in our results and supports the preference for a multimeasure approach with very young children (Olswang, Rodriguez, & Timler, 1998; Thal & Katich, 1996).

Finally, at this time, we cannot rule out a third alternative: It is possible that some or all of the misidentified children in the TDL group demonstrated a primary discourse or monitoring problem that went undetected by the other discourse comprehension measures, or that some of the LI group might have had stronger discourse abilities than could have been predicted by the original comprehension tests used to verify their initial classification, none of which was at the discourse level. Interestingly, as shown in Table 5, 1 TDL child (participant 3TDL, 33 months) was misidentified by both of the primary discourse measures and the comprehension monitoring measure. Although the detection task and comp questions tended to overidentify children in this age range, the joint retell task tended to benefit the youngest children, potentially balancing their impact. Yet in this child, all three measures suggested that the child had a discourse comprehension performance similar to that of a child with LI. It was beyond the scope of this study to follow up on our participants over a longer period, so we do not know if this child would have demonstrated discourse comprehension difficulties at a later age. The convergence of all three measures in this case illustrates the potential advantage of having multiple measures when making a clinical decision about very young children who may be quite variable in their comprehension performance. The convergence of multiple measures that have been shown to represent the same domain may enhance the confidence of interpreting a delay.

**CLINICAL IMPLICATIONS**

The evidence obtained in this study adds to the confidence with which we can interpret the joint retell task and detection task as measures of discourse comprehension and monitoring with young children. Further, the evidence accrued regarding our combination of discourse measures comprises a first step in validating the clinical action of applying them to identify comprehension impairment in children with language impairment. Additional study with a larger and more diverse sample of children with language impairment (e.g., those with expressive impairment only) is needed, as is investigation of the manner in which the individual experimental procedures measure discourse comprehension in children under the age of 3. The current study suggests that the combination examined may be both accurate and efficient. All three measures require a total of approximately 20 min to administer. Given the limited time available to clinicians to conduct assessments, identification may be only one of several possible objectives of the evaluation process (Lahey, 1988). Information about potential areas requiring treatment and the relative strengths and weaknesses of input and output modalities may also be sought during an initial assessment (Miller & Paul, 1995). Clinicians seek to accomplish as many goals as possible in as time efficient a manner as possible. Knowing that good identification accuracy could be achieved with the combination of measures studied, a clinician might then compare performance on the two primary discourse measures (which vary in expressive language demand and the extent to which memory is scaffolded) in order to inform decisions about input and output modalities for treatment. With further research into the potential of differential performance at the primary versus monitoring levels of discourse comprehension, clinical decision making regarding specific intervention targets may be informed.

Finally, the results of this study add to the accumulated evidence that discourse comprehension can be validly measured via orally presented stories in children as young as 30 months (Dempsey & Skarakis-Doyle, 2001; Milosky & Skarakis-Doyle, 2006; Skarakis-Doyle, 2002). In summary, we have offered preliminary empirical evidence of the valid use of a combination of both primary discourse comprehension and comprehension monitoring measures to identify young children with comprehension impairment at the discourse level. However, as noted previously, further investigations are necessary with children 3 years and younger, as well as with a greater diversity of language impairment. The initial success of our measures suggest that oral language comprehension problems that may interfere with later reading comprehension may be identified before a child has learned to decode written language. Our combined measures may then add to those proposed by Paris and Paris (2003) in the evaluation of this aspect of early literacy. Further, these measures may be useful as part of an even larger battery of tests used in the early identification of potential language comprehension impairment in hopes of ameliorating their severely disabling impact.
ACKNOWLEDGMENTS

This work was supported by grants from the Toronto Hospital for Sick Kids Foundation Grant XG01-081 and the Ontario Ministry of Health and Long Term Care—Children with Special Needs Branch. We are indebted to the late Siobhan Wooten for her contributions, without which this work could not have been conducted. We express our appreciation to Josephyne Jacques for her multiple contributions as project manager and to Melanie Beaudin, Sarah Pifer, and Brooke Thornton for their assistance with data collection and analysis. We are also grateful to Linda Milosky and Weronah Campbell for their helpful comments on earlier versions of the manuscript. Finally, we extend gratitude to the children, parents, and clinicians who participated or otherwise assisted in this study. Portions of this manuscript were presented at the Symposium on Research in Children Language Disorders (SRCLD) in June 2004, Madison, Wisconsin.

REFERENCES


Dockrell, J., & Lindsay, G. (1998). The ways in which speech and language difficulties impact on children’s access to curriculum. Child Language and Teaching Therapy, 14, 117–123.


Olswang, L., Rodríguez, B., & Timler, G. (1998). Recommending intervention for toddlers with specific language learning difficulties: We may not have all the answers, be we know a lot. American Journal of Speech-Language Pathology, 7, 23–32.


Received October 3, 2006
Revision received February 22, 2007
Accepted May 11, 2007
DOI: 10.1044/0161-1461(2008/006)

Contact author: Elizabeth Skarakis-Doyle, School of Communication Sciences and Disorders, Elborn College, University of Western Ontario, London, Ontario, N6G 1H1. E-mail: eskaraki@uwo.ca.

Lynn Dempsey is now at Brock University, St. Catharines, Ontario Canada.
APPENDIX. EXAMPLES FROM THE EXPERIMENTAL STIMULI

Splish Splash Story Stimuli (Skarakis-Doyle & Wootton, 1995)
One day a little girl named Sarah made very messy mud-pies in the backyard. Sarah’s mom took one look at her and said, “Splish, splash, Sarah needs a bath. Mommy says you’re dirty and she can’t have that.”
So Mom carried Sarah upstairs to the bathroom. Then she filled the bathtub with water and helped Sarah take off her dirty clothes. But Sarah said, “Mommy, I just can’t have a bath. First, I must test the water.”

Joint Story Retell Task Stimuli
One day a little girl named Sarah made very messy mud-pies in the backyard. Sarah’s mother took one look at her, and said, “________________.”
(Splish splash Sarah needs a bath, Mommy says you’re dirty and she can’t have that)
So mom carried Sarah upstairs to the bathroom and filled the bathtub with water.
"I must test the water”, said __________. (Sarah)
So she __________ (put) her big toe into the bathtub and said, “Oh Mommy, the water’s too cold. The water must be nice and warm.”

Expectancy Violation Detection Task Stimuli
One day a little frog* named Sarah made very messy mud-pies in the backyard. Sarah’s mom took one look at her and said, “Splish, splash, Sarah needs a bath. Mommy says you’re dirty and she can’t have that.”
So mom carried Sarah upstairs to the bathroom. Then she filled the bathtub with water and helped Sarah put on all her dirty clothes. But Sarah said, “Oh, Mommy I just can’t have a bath. First I must test the water.”
* practice item

Content and Gist Comprehension Questions
1. What did Sarah’s mother say when she saw Sarah in the backyard? (Splish Splash, Sarah needs a bath …)
2. Who wanted to check the bath water? (the little girl / Sarah)
3. In the story, why did Sarah keep starting to leave the bathroom? (she didn’t want a bath / she wanted to play)

Note. Violations and sample acceptable responses are in bold. In items where the “Splish Splash” refrain was used, brief portions were acceptable as a correct response.