Reliability and validity of the C-BiLLT: a new instrument to assess comprehension of spoken language in young children with cerebral palsy and complex communication needs.

Joke J. Geytenbeek, Lidwine B. Mokkink, Dirk L. Knol, R. Jeroen Vermeulen, Kim J. Oostrom

*Augmentative and Alternative Communication* 2014; 30(3): 252-266
ABSTRACT
In clinical practice, a variety of diagnostic tests are available to assess a child’s comprehension of spoken language. However, none of these tests have been designed specifically for use with children who have severe motor impairments and who experience severe difficulty with using speech to communicate. This article describes the process of investigating the reliability and validity of the Computer-Based instrument for Low motor Language Testing (C-BiLLT), which was specifically developed to assess spoken Dutch language comprehension in children with cerebral palsy (CP) and complex communication needs (CCN). The study included 806 children with typical development, and 87 non-speaking children with CP and CCN, and was designed to provide information on the psychometric qualities of the C-BiLLT. The potential utility of the C-BiLLT as a measure of spoken Dutch language comprehension abilities in children with CP and CCN is discussed.

Keywords: Cerebral palsy, complex communication needs, language test, language comprehension, validity, reliability
INTRODUCTION

Impairments in expressive and/or receptive communication skills are common in children with cerebral palsy (CP) and complex communication needs (CCN). It has been reported that the more severe the physical challenges associated with the CP, the more likely it is that impairments in the use of speech for communication will also be observed. In children with severe CP oro-motor impairments (i.e. severe dysarthria [unintelligible speech] or anarthria [absence of speech], are generally accompanied by severe limited mobility (WHO, 2001). Because of the challenges associated with communication for this population, there may be discrepancies between their expressive communication skills (production of speech, use of augmentative and alternative communication [AAC] systems) and their receptive communication skills (comprehension of spoken language).

The determination of a child’s comprehension of spoken language is important to the support of their participation and development. For example, insight into a child’s comprehension of spoken language has immediate consequences for how communication partners can most effectively interact with the child. Information on a child’s comprehension of spoken language also may influence choices regarding how augmentative and alternative communication systems can be best provided. Finally, the development of education programs to fit the particular needs of a child are also influenced by the child’s comprehension skills.

In clinical practice, various diagnostic tests are available to assess a child’s comprehension of spoken language. (for review: Geytenbeek et al.) However, none of these tests are designed specifically for children with severe motor impairment and anarthria or dysarthria. As a result, the assessment information that is gathered may not be a true reflection of the child’s ability, as it remains unclear whether test errors arise from the child’s difficulty in performing the needed motoric skills to provide a response (e.g., pointing to a response item) or from poor language comprehension.

For the assessment of children with CP and CCN, various adaptations to existing diagnostic tests have been made and alternative access methods have been used. Although these efforts can provide some useful information under some conditions, these ad hoc adaptations introduce new concerns. For children with severe limited mobility and for children not yet able to recognize numbers (i.e. young children and/or children with intellectual disabilities), standardized tests may require adaptations that are difficult to establish and/or may require additional time to administer. This may, in turn, have a negative influence on the motivation and/or performance of the child. Moreover, these adaptations may vary between individual testers, and the tests may not be administered as intended by the test developers. In these cases, the results have been obtained using methods that differ significantly from the standardized procedures provided by the test developers, thereby compromising the psychometric qualities of the test and limiting the interpretation of the obtained measurements. There is a need for an instrument that provides standardized procedures for assessment of language comprehension that can be delivered using the access methods.
(e.g., scanning, touch screens) used by children with CCN and physical disabilities. Just as importantly, this instrument should address the language and everyday experiences of children with physical disabilities, and should be developed and validated with appropriate samples. At present, there is no language comprehension assessment instrument available that specifically addresses the language and everyday experiences of Dutch children with severe physical disabilities and complex communication needs, who because of their disabilities may have more limited life experiences.5,16

In a series of studies, our research group has worked to address the challenge of reliably assessing comprehension of spoken Dutch language by children with severe physical and communication challenges in a series of studies.9,12 This exploratory work resulted in a pilot version of an assessment package called C-BiLLT (Computer-Based instrument for Low motor Language Testing).9 The C-BiLLT is designed to provide a test of a child’s comprehension of spoken Dutch language. It was developed so as to require very limited motoric responses from the child being assessed. Because the test is delivered using a computer, the child can provide responses using a variety of alternative access methods (e.g., scanning, touch-screen, eye-gazing or eye-gaze computer control).

Our goal has been to develop an assessment tool that would provide information on a child’s comprehension of Dutch spoken language relative to the performance of their peers without disabilities. For this reason, over time, we have collected response data from both Dutch children with severe CP and CCN, as well as from Dutch children (control sample) who were reported by their teachers and parents as being typically developing. Although the pilot version of the C-BiLLT showed promising construct, convergent, and face validity9, small but significant changes to the test were required to strengthen confidence in the reliability and validity of the tool as a measure of language comprehension. Therefore, for the investigation of the C-BiLLT reported in the present study, both the control and the experimental samples were enlarged and more complex test items were included. In addition, the pretest section of the C-BiLLT was extended with photographs and the earlier experimental section of sound recognition was deleted.

Administration of the C-BiLLT is meant to provide information on the child’s ability to understand Dutch spoken language, which is presented in both single words and spoken sentences of increasing difficulty in vocabulary and grammatical structure. As the test items are spoken aloud by the test examiner, a target (correct option) and a foil (incorrect option) are also simultaneously presented as photographs on the computer screen. The child is asked to select the photograph that best represents the spoken word, phrase, or sentence, and can provide a response to the computer-presented information using a variety of alternative access methods.

The C-BiLLT can be used with at least four alternative access methods; (a) a touch screen operated by any body part (e.g. with hand, foot, nose, etc); (b) input switch(es) that are adjustable to any body part, including switches integrated on a flexible and bendable shaft.
attached to the (wheel) chair of the child and operated by any body part; in test part I two switches are used (the left switch is used to select the left image and the right switch to select the right image on the screen), and in test part II one switch is used (for linear scanning); (c) the child’s own wheel-chair head support (used as an input switch); and (d) eye-gaze to the touch screen with observation of the selection by the test examiner who then inputs the choice with an external mouse or independent eye-gaze computer control with an infra-red camera or with an eye – control module.

The C-BiLL T uses vocabulary drawn both from two well-researched Dutch developmental receptive language lists - the "streeflijst woordenschat" and the "Lexilijst" as well as the Reynell Developmental Language Scales (RDLS, Dutch version). It was our belief that the vocabulary to be included in an instrument for language comprehension for children with severe disabilities and CCN should pertain to objects/situations that are relevant and identifiable to children with severe motor impairment, which may be different from the vocabulary familiar to children without disabilities, because of the more limited life experiences of children with severe disabilities. Therefore although the C-BiLL T does follow the linguistic hierarchy of the RDLS and includes items that are immediate derivatives of RDLS items, (e.g. “waarin kan je slapen?” [In what do you sleep?]), the majority of C-BiLLT items are new.

Aims of the Present Study
The present study had three major aims: (a) to investigate the clinical usefulness of the extended version of the C-BiLL T as a measure of comprehension of spoken Dutch language in children with severe CP and CCN; (b) to provide information on the performance of children with severe CP and CCN relative to their peers without disabilities, using test items and response activities that were specifically designed for children with severe physical disabilities; and (c) to provide information about the reliability, internal consistency (considered as an aspect of reliability), and validity of the C-BiLLT as a measure of spoken language comprehension.

Reliability of the C-BiLL T
The reliability of an instrument refers to its ability to measure the target construct (in this case comprehension of language) consistently across time, individuals, and situations despite measurement errors. emphasized that reliability parameters are required for instruments that are used for discriminative purposes and agreement parameters for instruments that are used for evaluative purposes. In the literature, agreement and reliability parameters are often used interchangeably despite their differences. Agreement and reliability parameters focus on two different questions. Information on agreement parameters addresses the question of how good is the agreement between repeated measurements (i., e., calculation of the measurement error and the closeness of scores for repeated measures). Information
on reliability addresses the questions of how well the participants can be distinguished from each other, despite measurement errors.\textsuperscript{25}

The C-BiLL T was developed to evaluate spoken language comprehension in children with CP and CCN. However, information on both agreement and reliability parameters are relevant for the investigation of the reliability of the C-BiLL T. To address these questions, we provide information on the intraobserver and interobserver reliability, standard error of measurement (SEM), and the intraclass correlation coefficient (ICC).

Intraobserver reliability, also referred to as test–retest reliability, refers to the same rater, using the same scale/instrument, assessing the same participants at different times. Interobserver reliability refers to different raters, using the same scale/instrument, assessing the same participants.\textsuperscript{25} For a test–retest reliability study, there are no standard rules on the appropriate time-interval between the measurements.\textsuperscript{26} Interference of test performance, can occur, due to changes over time (progression in comprehension) or learning effects (e.g. remembering previous items). The C-BiLL T contains a long list of items limiting the participant’s recall of previously tested items. However, to avoid interference of learning effects and to avoid changes over time in language comprehension, we followed a two weeks’ time interval that is most commonly used in test-retest reliability studies.\textsuperscript{26}

As guidelines recommend, our aim was to collect a subgroup of at least 50 participants for the investigation of each reliability measure.\textsuperscript{26,27}

We also provide information on the ICC as a measure of reliability. The ICC relates the measurement error variance to the variability between persons.\textsuperscript{26} There are several ICC formulas\textsuperscript{26} however, since we are mainly interested in absolute agreement (i.e. the test examiners draw the same conclusions about the test results) we will use ICC\textsubscript{agreement}. ICC values of 0.80 are considered acceptable for research and values between 0.90 and 0.95 for diagnostic instruments.\textsuperscript{28}

For the reliability of the C-BiLL T we considered as appropriate:

- ICC’s of > 0.80 for both intra and interobserver reliability of children with TD.
- ICC’s of > 0.80 for both intra-and interobserver reliability of children with CP and CCN.

The SEM is the square root of the measurement error variance and expresses the closeness of the scores of repeated measurements. For an evaluative instrument the variability between persons in the population sample does not matter at all; only the measurement error is important.\textsuperscript{25} If measurement error is small in comparison to variability between participants, the reliability parameter approaches 1. This means that discrimination between participants is scarcely affected by measurement error.\textsuperscript{26} The measurement error (represented by the (SEM) is a suitable parameter of agreement\textsuperscript{25} and is expressed in the unit of the scale. Reliability and SEM of the C-BiLL T were investigated both in the TD group, and in the CP and CCN group (Table 4.1).
Internal consistency of the C-BiLLT

Internal consistency relates to single test administration and the accuracy of the measurement. The most often used measure of internal consistency is Cronbach’s alpha coefficient. However, alternatives such as Guttman’s lambda provide important information and are often considered superior. Nevertheless, because the use of alpha continues to be very important and continues to be the subject of misinterpretation, the advice of Sijtsma was followed to investigate the internal consistency of the C-BiLLT and we report both Cronbach’s alpha coefficient and Lambda 2.

For the investigation of internal consistency of the C-BiLLT a Cronbach’s alpha ≥ 0.80 and Lambda2 of ≥ 0.80 was considered as appropriate.

Validity of the C-BiLLT

Validity is defined by the COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments) panel as ‘the degree to which an instrument truly measures the construct(s) it purports to measure’, in this case comprehension of spoken Dutch language. Evidence of the validity of an instrument for a particular task can be assessed by examining the relationship of test scores with similar (convergent validity) or dissimilar (discriminant validity) instruments, and by the use of exploratory item factor analysis (EFA) to provide information on construct (structural) validity. The RDLS and the Peabody Picture Vocabulary Test (PPVT-III) were chosen for the investigation of convergent validity and the Coloured Progressive Matrices (CPM), for discriminant validity. For the investigation of validity measures, the same rule for sample sizes was followed as for reliability measures.

To investigate the validity of the C-BiLLT we hypothesized:

1) a high correlation between the C-BiLLT and RDLS scores in children with TD because the items of the C-BiLLT were based on the linguistic hierarchy of the RDLS;
2) a slightly lower (difference of 0.1-0.2) correlation between the C-BiLLT and PPVT than between the C-BiLLT and RDLS in children with TD and children with CP and CCN because the PPVT tests receptive vocabulary instead of sentence comprehension;
3) a moderate correlation between the C-BiLLT and CPM in children with TD because the CPM involves non-verbal reasoning that may be associated with verbal comprehension but actually pertains to a different cognitive function; and
4) that the EFA analysis would yield a unidimensional solution with the factor accounting for ≥ 70% of the variance.

In addition, construct validity as defined by the COSMIN panel also includes the degree to which the scores of a measurement instrument are consistent with hypotheses, not only with regard to relationships with scores of other instruments but also with regard to differences between relevant groups. Therefore, expected differences between children with CP and TD children and age effects were also investigated.
Table 4.1
Demographic Information and Clinimetric Analyses of Study Sample (N = 893)

<table>
<thead>
<tr>
<th>Children with</th>
<th>Number of children</th>
<th>M/F</th>
<th>Mean age (SD)</th>
<th>Age range</th>
<th>GMFCS IV/V</th>
<th>Additional test measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>TD</td>
<td>806</td>
<td>424/382</td>
<td>4;4 (1;5)</td>
<td>1;6 – 7;6</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>87</td>
<td>46/41</td>
<td>6;10 (2;11)</td>
<td>1;9 – 12;0</td>
<td>34/53</td>
<td></td>
</tr>
</tbody>
</table>

Factor analysis and internal consistency

| TD and CP     | 893                | 470/423 | 4;7 (1;9)   | 1;6 – 12;0|            |                        |

Convergent validity

| TD            | 423                | 223/200 | 3;11 (1;5)  | 1;6 – 6;3  | NA         | RDLS                  |
| TD            | 117                | 60/57   | 4;9 (1;6)   | 2;3 – 7;0  | NA         | PPVT-III              |
| CP            | 33                 | 19/14   | 7;6 (2;8)   | 3;2 – 12;0 | 14/19      | PPVT-III              |

Discriminant validity

| TD            | 103                | 56/47   | 5;0 (0;7)   | 4;6 – 6;6  | NA         | CPM                   |

Intra-observer reliability and SEM

| TD            | 137                | 71/66   | 4;3 (1;6)   | 1;6 – 6;10 | NA         |                        |
| CP            | 32                 | 16/16   | 7;0 (2;8)   | 1;11 – 12;0| 17/15      |                        |

Inter-observer reliability and SEM

| TD            | 37                 | 20/17   | 4;10 (1;3)  | 1;8 – 7;0  | NA         |                        |
| CP            | 35                 | 14/21   | 5;11 (2;9)  | 1;7 – 11;4 | 15/20      |                        |

Note. TD = typical development; CP = cerebral palsy; GMFCS = Gross Motor Function Classification System; RDLS = Reynell Developmental Language Scales; PPVT III = Peabody Picture Vocabulary Test – III; CPM = Raven’s Coloured Progressive Matrices; SEM = standard error of measurement; NA = not applicable.
Hypotheses were:

5) a higher correlation (≥ 0.2) between the C-BiLLT and age in children with TD than in children with CP and CCN, because the latter children show a wider variation in their language comprehension skills (i.e., a greater range of performance at different ages) than children with TD;

6) a significant difference between the mean scores on the C-BiLLT of children with TD and children with CP and CCN, with children with CP and CCN receiving, on average, lower scores. Finally, normed scores of the C-BiLLT (to establish a norm-referenced test), are derived from raw test scores. Therefore raw test scores should reflect the linguistic hierarchy of language comprehension for different ages. In order to investigate if raw C-BiLLT scores discriminate between age intervals in TD children we hypothesized:

7) a significant difference between the mean scores on the C-BiLLT of TD children within age groups in years.

METHODS

Participants

The TD group consisted of 806 children recruited from mainstream primary schools and day-care centers throughout the Netherlands. There were five major exclusion criteria: (a) documented history of speech/language delay; (b) auditory or visual problems, developmental delay or learning disability, and/or any neurological or otherwise chronic disease(s) as reported by parents or nursery school teacher (in the Netherlands, children receive regular developmental screenings for physical, motor, and mental development by qualified medical professionals); (c) younger than 1;6 (years; months), or older than 7;6 years; (d) performance more than 1 sd below the mean on administered (dis) similar instruments (RDLS, PPVT-III, CPM); and (e) no Dutch speaking parent(s)/caregiver(s).

The CP group consisted of 87 children with CP and CCN recruited from rehabilitation centers, special schools for children with physical disabilities, and special day-care centers throughout the Netherlands. There were six major inclusion criteria: (a) medical diagnosis of CP (i.e. spastic, dyskinetic, or ataxic); (b) severe motor impairment corresponding to Gross Motor Function Classification System (GMFCS) levels IV and V; (c) productive spoken vocabulary of ≤ 5 words; (d) no medically diagnosed or otherwise documented history of auditory or severe visual perceptive problems; (e) older than 1;6 years and younger than 12;0 years; (f) Dutch speaking parent(s) or caregiver(s).

Table 4.1 presents the demographic characteristics of the two groups.

The Medical Ethics Committee of the VU University Medical Centre Amsterdam approved the study. All parents provided written informed consent.

The C-BiLLT

The C-BiLLT consists of three sections: a pre-test, a learning module, and a computer-based test. Each is described below.
The pre-test is used to explore whether a child is able to communicate a choice between two concrete objects (or digital photos of the objects) when provided with the spoken name of the objects. Eight pre-selected familiar objects (as identified by parents) from the child’s home environment (i.e., the child’s milk-bottle or cup, ball, spoon, coat, trousers, toy car or doll, favourite book, and favourite digital video disc cover) are presented in pairs to the child. While holding a target and foil object in separate hands, the examiner verbally asks the child to select the target object (e.g., ‘Where is the ball?’) by reaching, pointing, eye-gazing to his/her object of choice. This is repeated in fixed order until all eight pairs of objects have been presented. Next, the child is presented with the same eight objects but now presented as generic photographs of the item (photographic print size A5). Again in fixed order of eight pairs, the child is asked verbally to select the target item. Although the investigator was allowed to encourage the child to provide a response in any way possible, the actual question could only be repeated once. The response of the children was considered correct if the child (based on Heim, 2001) fixated his/her eyes on the targeted object for at least two seconds; or reached for the targeted object with his/her arm, foot, elbow, or hand; or pointed to the targeted object with his/her hand, arm, foot, or head; or turned his/her head to the targeted object with an accompanying vocal sound. If the child correctly identified at least 5 objects and/or 5 photographs, he/she was considered to be able to communicate a selection in response to a spoken response opportunity and advanced to the learning module.

The learning module is incorporated in the assessment package for two reasons. First, to make the child aware of the association between the access methods and the visual representations (response selection) on the 19-inch touch screen. Second, to empirically find the best response mode for the individual child (e.g. eye gazing, touch-screen, input switches, or the child’s own wheelchair head support) when there are concerns regarding access method.

The computer test consists of 75 items presented in two parts. Part I consists of three primary sections, with each section containing 10 items referring to nouns, verbs, animals, objects and persons. For each item, the child is shown two digital photographs (a target and a foil) arranged horizontally on the computer screen. To control for effects of chance for each of these three sections, a parallel section exists, presenting the same test items in a different order and with a different foil for each target item (see Figure 4.1). The examiner always completes the primary form. Only when a child delivers one or more incorrect responses on the primary form, the examiner assesses the entire parallel form.

Part II consists of 45 test items pertaining to spoken sentences with increasing complexity of grammatical structures and is organized into 8 sections. The response options per test item are now represented by four digital photographs in a 2 x 2 matrix on the computer screen (see Figure 4.2), providing one target and three foils. For both the first and second part of the computer test, visual feedback of the child’s response is shown by the appearance of a red square around the chosen photograph. Information on all sections is described in Appendix I.
**Score calculation.** The child can attain a pretest score of either 1 point (for identification of at least 5 real objects or 5 photographs) or 2 points (for identification of at least 5 of both real objects and photographs). The learning module activities do not contribute to the final score. For the computer test Part I (Sections 1, 2 and 3) one point is given for each correctly identified item in both original and parallel version with a total of 30 points. For instance if a child answered the question “Where is the telephone” incorrect in the primary form and correct in the parallel form, no point is scored for the item “telephone”. If the child answered this question correctly both in the primary as well as in the parallel form, one point is scored for the item “telephone”. For the computer test Part II (Sections 4 to 11), one point is scored for each correct response with a total of 45 points. The maximum achievable C-BiLLT score is 77, i.e. the sum of the pretest and the scores of the two parts in the computer test.

**Administration Procedures**
Administration of the C-BiLLT took place in a distraction-free room at the educational environment of the child. An experienced speech and language pathologist (JG) performed the first C-BiLLT administration of all 87 children with CP and CCN. As a first step, the most appropriate access method for the child was determined in consultation with the child’s personal speech language therapist (SLT) and/or occupational therapist (OT) or, in some cases, with the child himself/herself. If an access method that was similar to the child’s current access strategy was possible, this access method was chosen. With regard to eye-gaze computer control (i.e., infra-red camera, eye-control module), this technique was only used with children who were already familiar with the calibration techniques and the use of this technology (n=4).
If additional support of the child’s personal SLT or OT was necessary to find the best response mode for the child, the actual session of the C-BiLLT may involve a second session (depending on concentration and fatigue of the child). Moreover, if a child showed symptoms of fatigue, anxiety, or inattentive behaviour at the start or during the assessment, the testing activities were ended and not included in the dataset.

In children with TD, administration of the C-BiLLT was carried out by 1 speech language pathologist (SLP), 1 SLT, 12 Master’s students in speech pathology and 3 neuropsychologists. Both for children with TD, and for children with CP and CCN, administration of the C-BiLLT was performed in the presence of an observer (i.e. one of the 17 test administrators). The access methods used by TD children were either the touch screen or switch activators (operated by finger pointing) or eye-gazing at the screen (which was observed and recorded by the test administrator). Because (future) normed scores of the C-BiLLT will be derived from children with TD, for these children (and especially the younger ones) we applied the rule that if a child did not participate (perhaps because of anxiety, shyness, or inattentive behavior), the test scores were excluded from the dataset.

In order to gather information on test-retest reliability, the C-BiLLT was administered twice under the same conditions, with a time interval of 2 weeks, in a subset of children (see: Reliability and Standard Error of Measurement). To investigate construct validity, additional tests were administered (see Construct validity) immediately prior to administration of the C-BiLLT by the same test administrator. For reasons of efficiency, an exception was made for the children with CP and CCN (n=68) in whom the additional test (i.e. the PPVT-III) was carried out by the child’s personal SLT; considering the workload of the child’s own SLT and to avoid changes over time in language comprehension skills of the child, these test data were incorporated in our study if test administration was done within 4 weeks after the administration of the C-BiLLT.

Psychometric Evaluation and Statistical analysis
Information was collected on the administration of the C-BiLLT for both groups of children.

Access methods. We collected information on the access methods used for completion of the C-BiLLT by both groups of children; this information is presented in the Results section.

Reliability and SEM (standard error of measurement).
Children with TD. For evaluation of intra-observer reliability and SEM, C-BiLLT administration was repeated in 137 TD children by 1 SLT, 1 SLP and 3 neuropsychologists (i.e., 5 of the in total 17 test administrators) after a time interval of 2 weeks after the initial assessment. The ICC was calculated by dividing the variance of children nested in raters, and raters by the total variance37 (see Appendix A for formulas).

For evaluation of inter-observer reliability and SEM of the C-BiLLT in the TD group, C-BiLLT
administration was repeated in 37 children by 2 raters (see Limitations section). Responses of children with TD were rated twice, independently, by 2 of the 5 raters. Hence, the design was incomplete, and ICC was calculated by dividing the variance of the child by the total variance (see Appendix A for formulas). Variance components were assessed using SPSS VARCOMP.

**Children with CP and complex communication needs.** For evaluation of intra-observer reliability and SEM, 32 children were retested. Because intra-observer reliability of the children with CP and CCN was investigated by 1 rater, a one-way design was used to calculate reliability. For evaluation of inter-observer reliability and SEM, C-BiLLT administration was repeated in 35 children with CP and CCN, by 8 experienced and trained SLT’s, after a time interval of 2 weeks after the initial assessment. To investigate SEM, the same variance components as described above (see Appendix A) were used.

**Internal consistency.** To investigate to which degree the items of the C-BiLLT measure the same construct, the internal consistency of the C-BiLLT was examined with both Cronbach’s alpha coefficient and Guttman’s Lambda 2.

**Validity**

**Structural validity.** To investigate structural validity, i.e. the degree to which the scores of an instrument are an adequate reflection of the outcome of the dimensionality of the construct to be measured, in this case comprehension of spoken language, an exploratory item factor analysis (EFA) was performed on all data \( n=893 \). EFA was performed because we studied the factor structure of a newly developed scale. Moreover, instead of using linear factor analysis we performed item factor analysis, which is a more appropriate method for handling the skewed answering patterns and the ordered categorical item responses of the C-BiLLT. Item factor analysis uses the matrix of polychoric correlations rather than Pearson’s correlations. The model parameters were estimated by the method of maximum likelihood with robust standard errors that can handle missing data. The chosen number of factors was based on the percentage explained variance and the factor loadings (i.e. essentially correlation coefficients between the construct to be measured). Tabachnick and Fidell cite .32 as a minimum loading of an item which equates to approximately 10% overlapping variance with the other items in that factor. Therefore, factor loadings of > 0.40 were considered to be satisfactory. Items that showed high residual correlations (> 0.2) with several other items were considered for removal, because this can also be viewed as a violation of unidimensionality.

**Construct validity.** Convergent validity examines the degree to which the C-BiLLT is similar to other tests that are designed to measure similar constructs. Discriminant validity examines the degree to which the C-BiLLT is, different from tests that do not measure comprehension of spoken language. Therefore, the test results for the C-BiLLT were compared with the
children’s performance on three tests: 1) for convergent validity the RDLS (Dutch version, standardized for children aged 1;6 years to 6;3 years) was administered only in children with TD because of its motor demands; 2) for convergent validity the Peabody Picture Vocabulary Test-III (PPVT-III) (receptive vocabulary test, standardized for children and adults from 2;3 to 90;0 years) was administered to both groups of children; and 3) for discriminant validity the Raven’s Coloured Progressive Matrices (CPM) (non-verbal logic reasoning, standardized for ages 4-11 years) administered in children with TD aged 4-6 years. An overview of the different clinimetric analyses are given in Table 4.1.

Descriptive statistics were calculated for the raw score(s) of the C-BiLLT in both groups of children (mean, range and SD) (see Table 4.2).

With regard to the investigation of the validity of the C-BiLLT, hypotheses 1-3 and 5 were tested by calculating Pearson’s correlations. A correlation of 0-0.3 was considered poor, 0.3-0.6 as moderate, and above 0.6 as high. Hypothesis 4 was investigated with an EFA, hypothesis 6 was tested with an independent t-test and hypothesis 7 with an univariate analysis of variance. A summary of the hypotheses is presented in Table 4.3.

All statistical analyses (except for the factor analysis) were performed with SPSS-16 for Windows. The EFA was performed using Geomin rotation in Mplus Version 6.1.

RESULTS

C-BiLLT administration and Response Methods

The test was administered to 831 children with TD and 90 children with CP and CCN. For both children who were TD (n = 25) and children with CP (n = 3), 3% of the children did not participate because of anxiety, shyness, or inattentive behaviour.

Of the children with TD aged < 2 years, 74% used the touch screen, 13% eye-gazing (n = 6, < 1;8 years, [observed by the test administrator]) and 13% used the switch activators to indicate their responses while children aged ≥ 2 years were more likely to use the switch activators (73%). The remaining 27% of children with TD ≥ 2 years used the touch screen to indicate their response. All children with TD completed the assessment within 10-20 min.

Of the children with CP and CCN, 76% (n=66) indicated their responses in a way similar to their current access strategy. Thirty percent (n=27) of the children used the touch screen function either with their hand, arm, foot, forehead or nose to indicate their responses. Eighteen percent (n=16) of the children used their own wheelchair head support to make a selection using scanning. Selection with the switch activators with their hands or head was used by 6% (n=5) of the children. Eye-gazing to the screen (as observed by the test administrator) was used by 14% (n=12) of the children and 5% (n=4) of the children used eye-gaze computer control. Different combinations of access methods to indicate their responses (e.g. one switch activator with their left hand to indicate left oriented images and eye-gazing to right oriented images, or the switch activators with their head and with their better hand) was used by 12% (n=10) of the children. The remaining children (15% of
<table>
<thead>
<tr>
<th>Age a</th>
<th>N</th>
<th>Mean b (sd)</th>
<th>95% CI</th>
<th>Range</th>
<th>Age a</th>
<th>N</th>
<th>Mean b (sd)</th>
<th>95% CI</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1;11</td>
<td>45</td>
<td>23.84 (11.19)</td>
<td>20.84 – 27.21</td>
<td>2 - 46</td>
<td>3</td>
<td>20.33 (17.16)</td>
<td>0 – 62.95</td>
<td>2 - 36</td>
<td></td>
</tr>
<tr>
<td>2-2;11</td>
<td>129</td>
<td>46.24 (9.04)</td>
<td>44.68 – 47.83</td>
<td>11 - 61</td>
<td>5</td>
<td>28.40 (12.16)</td>
<td>13.30 – 43.50</td>
<td>11 - 42</td>
<td></td>
</tr>
<tr>
<td>3-3;11</td>
<td>137</td>
<td>57.91 (5.88)</td>
<td>56.92 – 58.90</td>
<td>39 - 71</td>
<td>11</td>
<td>28.82 (18.86)</td>
<td>16.15 – 41.49</td>
<td>0 - 48</td>
<td></td>
</tr>
<tr>
<td>4-4;11</td>
<td>191</td>
<td>64.35 (4.70)</td>
<td>63.68 – 65.02</td>
<td>50 - 74</td>
<td>9</td>
<td>28.89 (27.86)</td>
<td>7.47 – 50.30</td>
<td>0 - 66</td>
<td></td>
</tr>
<tr>
<td>5-5;11</td>
<td>202</td>
<td>67.55 (4.27)</td>
<td>66.96 – 68.14</td>
<td>55 - 76</td>
<td>5</td>
<td>40.40 (26.27)</td>
<td>7.78 – 73.02</td>
<td>1 - 65</td>
<td></td>
</tr>
<tr>
<td>6-6;11</td>
<td>91</td>
<td>70.64 (3.34)</td>
<td>69.94 – 71.33</td>
<td>57 - 77</td>
<td>19</td>
<td>29.05 (26.0)</td>
<td>16.52 – 41.58</td>
<td>0 - 76</td>
<td></td>
</tr>
<tr>
<td>7-7;5</td>
<td>11</td>
<td>72.18 (1.60)</td>
<td>71.11 – 73.26</td>
<td>71 - 77</td>
<td>6</td>
<td>37.67 (29.84)</td>
<td>6.35 – 68.99</td>
<td>1 - 74</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>51.00 (10.79)</td>
<td>37.60 – 64.40</td>
<td>40 - 65</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>41.17 (25.75)</td>
<td>14.15 – 68.19</td>
<td>2 - 62</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>48.75 (25.81)</td>
<td>27.18 – 70.32</td>
<td>8 - 73</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>51.30 (28.55)</td>
<td>30.88 – 71.72</td>
<td>0 - 74</td>
<td></td>
</tr>
</tbody>
</table>

CP = cerebral palsy; CCN = complex communication needs; a = chronological age range in years; b = mean C-BiLLT raw scores; CI = confidence interval; range = range in score
Table 4.3
Hypotheses and outcomes on Construct Validity

<table>
<thead>
<tr>
<th>Hypotheses (H):</th>
<th>Result</th>
<th>Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1: C-BiLL T versus RDLS in TD: high</td>
<td>$r = .93$</td>
<td>Yes</td>
</tr>
<tr>
<td>H2: C-BiLL T versus PPVT in TD/ CP: 0.1 lower than H1</td>
<td>$r = .88/ .87$</td>
<td>No</td>
</tr>
<tr>
<td>H3: C-BiLL T versus CPM in TD: moderate</td>
<td>$r = .43$</td>
<td>Yes</td>
</tr>
<tr>
<td>H4: EFA yield a unidimensional solution with the factor accounting for ≥ 70% of the variance.</td>
<td>Factor 76%</td>
<td>Yes</td>
</tr>
<tr>
<td>H5: C-BiLL T versus age in TD/ CP: 0.2 higher in TD</td>
<td>$r = .82 / .36$</td>
<td>Yes</td>
</tr>
<tr>
<td>H6: C-BiLL T in TD versus CP: significant difference</td>
<td>Sign ($p &lt; 0.0001$)</td>
<td>Yes</td>
</tr>
<tr>
<td>H7: C-BiLL T raw scores in TD children yields significant differences between age –intervals</td>
<td>Sign ($p &lt; 0.0001$)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

TD, children with typical development; CP, children with severe Cerebral Palsy; RDLS, Reynell Developmental Language Scales; PPVT-III, Peabody Picture Vocabulary Test; CPM, Raven’s Progressive Coloured Matrices; sign, significant; $r$, Pearson’s correlation coefficient; $a$, independent t-test; $b$, univariate analysis of variance with Bonferroni corrections.

Table 4.4
Access Methods of Children with TD and Children with CP and CCN [n (%)]

<table>
<thead>
<tr>
<th>Access method</th>
<th>TD&lt; 2 years of age</th>
<th>TD≥ 2 years of age</th>
<th>CP and CCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Touch screen</td>
<td>33 (74)</td>
<td>204 (27)</td>
<td>27 (30)</td>
</tr>
<tr>
<td>Eye-gazing (observed)</td>
<td>6 (13)</td>
<td>-</td>
<td>12 (14)</td>
</tr>
<tr>
<td>Switch activators</td>
<td>6 (13)</td>
<td>547 (73)</td>
<td>5 (6)</td>
</tr>
<tr>
<td>Combination of the above</td>
<td>-</td>
<td>-</td>
<td>10 (12)</td>
</tr>
<tr>
<td>Wheelchair head support</td>
<td>N.A.</td>
<td>N.A.</td>
<td>16 (18)</td>
</tr>
<tr>
<td>Eye-gazing (computer controlled)</td>
<td>N.A.</td>
<td>N.A.</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Drop out due to failure in pre-test</td>
<td>-</td>
<td>-</td>
<td>13 (15)</td>
</tr>
<tr>
<td>Drop out due to anxiety, shyness or inattentiveness</td>
<td>22$^a$ (3)</td>
<td>3$^a$ (0.004)</td>
<td>3$^a$ (3)</td>
</tr>
</tbody>
</table>

Note: TD = typically developing children; CP = cerebral palsy; CCN = complex communication needs; N.A. = not applicable; $^a$ = number of children excluded from de dataset.
the total, \(n = 13\) did not provide correct responses for at least 5 objects or photographs of the pre-test and consequently did not advance to the computer test. See Table 4.4 for an overview of the different access methods for TD and CP children.

The test administrators reported that for those children who completed the C-BiLLT assessment activities (97%), none of the assessments of children was hampered because of limited mobility function. In the CP and CCN group, the assessment time ranged from 15-60 min.

**Reliability and SEM**

Both intra-observer \((n=137)\) and inter-observer reliability \((n=37)\) were excellent for children with TD, with ICCs of 0.97 (95% confidence interval [CI] 0.94-0.99) and 0.96 (95% CI 0.94-0.97), respectively. SEM for intra- and inter-observer reliability for children with TD was 3.11 and 2.40, respectively.

Intra-observer \((n=32)\) and inter-observer reliability \((n=35)\) of the CP group was excellent with ICCs of 0.97 (95% CI 0.95-0.99) and 0.97 (95% CI 0.95-0.98), respectively. SEM for intra- and inter-observer reliability of the CP group was 3.40 and 3.00, respectively.

**Internal consistency:** Cronbach’s alpha and Guttman’s lambda2 calculated with 75 items (see Structural validity) in the TD group were 0.88 and 0.93, respectively, compared with 0.91 and 0.94, respectively, in the CP group.

**Validity**

An overview of the results of the hypotheses on validity measures of the C-BiLLT are presented in Table 4.3.

**Structural validity.** Residual correlations of the one-factor solution were high (> 0.2) for item 54 with 24 of the 76 computer items; therefore, item 54 was deleted. The exploratory item factor analysis based on the remaining 75 items resulted in a unidimensional factor solution. The factor had an explained variance of 76%. Factor loadings on the factor ranged from 0.43 for item 70 to 0.99 for item 28 (Table 4.5). EFA was performed on all data, including 87 children with CP. Results of the analysis of data from the children with TD only (data not shown) did not change our conclusions.

**Construct validity.** Two of the three hypotheses on relationships with other instruments were confirmed, with the strongest correlation between the C-BiLLT scores and the RDLS scores in children with TD \((r=0.93, [n=423])\). A somewhat less lower than expected (but still high) correlation was found between the C-BiLLT scores and the PPVT-III scores in children with TD \((r=0.88 [n=117])\) and in children with CP \((r=0.87 [n=33])\). As predicted, the least strong correlation was found between C-BiLLT scores and CPM results \((r=0.43)\) in children with TD \([n=103]\). All hypotheses concerning differences between sub-groups of participants were confirmed (Table 4.3). Correlation between age and C-BiLLT performance for children with TD was high \((r=0.82)\) and was moderate for children with CP \((r=0.36)\).
Table 4.5
Exploratory Factor analysis (EFA)

Factor loadings of the items of the C-BiLLT (Computer-Based instrument for Low motor Language Testing). (n = 893; explained variance 76%)

<table>
<thead>
<tr>
<th>Part</th>
<th>Section</th>
<th>Number of items</th>
<th>Focus (Identification of ...)</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
<td>nouns</td>
<td>0.907-0.968</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>10</td>
<td>verbs</td>
<td>0.933-0.967</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10</td>
<td>animals, objects and persons</td>
<td>0.960-0.984</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>More difficult nouns</td>
<td>0.927 - 0.958</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>Simple sentences with objects in combination functions and prepositions</td>
<td>0.895-0.955</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>“who”questions of persons performing activities</td>
<td>0.746 - 0.944</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>4</td>
<td>Passive and active sentences with objects and prepositions combined with events</td>
<td>0.431*-0.754</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>9</td>
<td>sentences with more complexity in grammar and semantic</td>
<td>0.562 - 0.966</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>6</td>
<td>simple active sentences referring to non-observable situations of four persons</td>
<td>0.754 – 0.881</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>9</td>
<td>complex sentences with comprehension of two and more concepts referring to food products.</td>
<td>0.422*-0.941</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>4</td>
<td>compound sentences referring to the lunch table</td>
<td>0.632 – 0.887</td>
</tr>
</tbody>
</table>

* Only item 45 of section 7 and item 70 of section 10 had a factor loading of < 0.500

Table 4.6
Multiple Comparison of Age Groups for C-BiLLT scores between Children with Typical Development

<table>
<thead>
<tr>
<th>Age</th>
<th>Mean difference</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 - 1</td>
<td>22.41*</td>
<td>19.31 – 25.51</td>
</tr>
<tr>
<td>3 - 2</td>
<td>11.66*</td>
<td>9.46 – 13.85</td>
</tr>
<tr>
<td>4 - 3</td>
<td>6.43*</td>
<td>4.43 – 8.44</td>
</tr>
<tr>
<td>5 - 4</td>
<td>3.20*</td>
<td>1.40 – 5.01</td>
</tr>
<tr>
<td>6 - 5</td>
<td>3.25*</td>
<td>1.08 – 5.43</td>
</tr>
<tr>
<td>7 - 6</td>
<td>1.54</td>
<td>4.37 – 7.46</td>
</tr>
</tbody>
</table>

* = comparison of age intervals with Bonferonni correction; b = mean difference in scores between age groups in years
* p < 0.001
Mean C-BiLLT scores showed significant differences between children with TD (mean 59.72, range 0-77, SD 13.19, CI 58.81- 60.63) and children with CP and CCN (mean 36.38, range 0-76, SD 25.08, CI 31.03- 41.72 [p < 0.0001]).

In TD children mean C-BiLLT score differed within groups of children at different ages and increased for children as they became older (e.g. a mean score of 23.84 for children younger than 2 years, 57.91 for children between 3;0 to 3;11 years up to 72.18 for children older than 7 years) and the range in scores became smaller as chronological age increased.

In children with CP and CCN, mean C-BiLLT scores varied considerably across the different ages and range in scores showed a large variety within age groups, although the maximum score showed a steady increase up to the age of 6 years (from 36 at the age of 1 year up to 76 at the age of 6 years) (See Table 4.2).

For both groups of TD and CP children ceiling effects were observed in individual cases (see discussion).

C-BiLLT mean scores in children with TD showed significant differences between age intervals (p < 0.001) with the exception of the age interval between 6 and 7 years (see Table 4.6).

**DISCUSSION**

To our knowledge, the C-BiLLT is the first instrument specifically designed to assess comprehension of spoken language in young children with CP and CCN, and to address issues of access for children with severe physical disabilities. The present study elaborates on our earlier work by providing an in-depth clinimetric evaluation of the C-BiLLT. Items of the extended and present version of the C-BiLLT were tested on a substantial cohort of children with TD, and children with CP and CCN, to investigate the reliability and validity of the C-BiLLT as a measure of spoken language comprehension.

**C-BiLLT administration and Response Methods**

The finding that 97% of the children with severe CP were attentive during the assessment of the C-BiLLT may be due to the fact that we included items and “real life” photographs that were drawn from the life experiences of children with severely limited mobility and CCN. Moreover, the various access methods incorporated in the administration of the C-BiLLT enabled children to take advantage of the different options to indicate their response by any body part. However, the proposed assessment may have limited utility in children with the most severe motor problems. Despite a clear reduction of motor demands, the assessment does not entirely eliminate the interference of motor impairment. Moreover, we cannot exclude the possibility that children possess language comprehension abilities despite absence of a clearly intelligible response, even when assessed with the accommodations supported by C-BiLLT. Furthermore, the development of any assessment tool is a continuous process of review and analysis for its usefulness for a particular purpose. The development and appliance of the C-BiLLT is no exception on this matter.
Reliability of the C-BiLLT
Excellent values of reliability were found with CIs that are appropriate for use of diagnostic measures. The small confidence intervals for both intra as well as inter-observer reliability indicate that the ICC’s true value are greater than 0.90. In other words, the repeated measurements with the same or different raters is scarcely affected by measurement errors and therefore the reliability is high. The fact that the C-BiLLT applies an observable visual feedback for both the child (test taker) and the test examiner may improve the reliability of the scoring of answers on the test items, as well as the test-retest reliability of the C-BiLLT. In addition, with regard to eye-gazing, the 19 –inch screen of the C-BiLLT portrays a clear distinction between the left and right image(s) enabling clear observation of on which image the child focusses. Moreover, the SEMs (expressed in the unit of the measurement) were small for children with TD as well as for those with CP and CCN. The smaller the SEM, the smaller the changes that can be detected beyond measurement error. In other words, for clinical use of the C-BiLLT it is likely that the measurement error is smaller than the improvements or deterioration one wants to detect with the C-BiLLT. For instance if a child achieved a C-BiLLT score of 30 then its true score lies between 25-35 (i.e. observed score ± 1.96×SEM). This range of the true score seems acceptable for a test that consists of a total of 77 items.

Validity of the C-BiLLT
The factor analyses revealed a clear unidimensional structure, best characterized as comprehension of spoken language. Factor loadings were uniformly high, i.e. ≥ 0.50. Only two items had a factor loading of ≤ 0.50 but ≥ 0.40, which is higher than the indicated minimum loading of an item. Loadings for the first 39 items were particularly high, indicating a high correlation between the items. An explanation for this correlation is that we deliberately developed relatively more undemanding items (referring only to linguistic mapping) for the first sections of the C-BiLLT to guarantee small increments of increasing difficulty. Although some redundancy of items in the first three sections is possible, this has not yet been investigated. On the other hand, since we plan to develop a computer adaptive test version of the C-BiLLT, some redundancy of items is essential. Computer adaptive testing might be of considerable advantage in children with CP and CCN because it adapts to the test taker’s ability level and reduces the length of the assessment.
In addition, with regard to the hypotheses on convergent and discriminant validity as measured by statistical relationships with other instruments, we observed high correlations with existing language and receptive vocabulary tests (RDLS and PPVT-III, respectively) and moderate correlation with a test for non-verbal logic reasoning. We had expected this finding, as both the RDLS and PPVT-III measure similar constructs. However, the correlation between the PPVT-III and C-BiLLT was slightly higher than expected. Nevertheless, these findings provide evidence of the construct validity of the C-BiLLT with highly investigated
measures of language comprehension, and underline the high correlation between receptive vocabulary and sentence comprehension in children with TD as well as in children with CP and CCN. Moreover, notably and in line with earlier findings, PPVT-III assessments were ended in 19 children (22%) with CP and CCN because the necessary alternative administration of the PPVT could not be done reliably. In addition, PPVT assessments were not completed and/or were reported as unreliable in 40% (n=35) of the children with CP and CCN. These findings underline the importance of the development of the C-BiLLT, providing an appropriate test administration in children with severe CP and CCN.

With regard to the hypotheses concerning expected differences between the sub-groups of participants, application of the C-BiLLT confirmed increasing comprehension of spoken language with age in children with TD and to a lesser account in children with CP and CCN. With the exception of the age interval between 6 and 7 years, the C-BiLLT differentiated between age groups of children with TD. Older children achieved higher scores. In addition, the standard deviation (SD) in children with TD became lower as children became older. The SD shows how much variation or dispersion from the average exists. A low standard deviation indicates that the data points tend to be very close to the mean which is to expected in children older than 5 years, because receptive grammar development is generally completed by the age of 5-6 years; a high SD (indicating that the data points are spread out over a large range of values) was found in the youngest children who indeed show more variability in language development.

As expected, in children with CP and CCN, C-BiLLT scores varied considerably and (overall) were significantly lower than in children with TD. Moreover, high SD were found for all age groups indicating that C-BiLLT scores showed a large range of values and thus a large variability in scores. At the same time, 11% (n =10) of the children with CP and CCN (aged 6;5 to 11;5 years) achieved the highest or nearly highest possible score on the C-BiLLT (data not shown). For these children, language comprehension skills exceeded clinicians’ expectations, or at least objectified clinical or parental impressions. Age equivalent scores were derived from the distribution of raw scores of the TD children and this highest or nearly highest possible score was comparable with an age equivalence of 6;6 years. The findings suggest that the C-BiLLT can distinguish between a group of children with TD and a group of children diagnosed with CP and CCN. Results indicate that raw test scores of TD children sufficiently discriminate and provide adequate data to calculate standardized scores at least up to the age of 6;6 years.

To summarize, reliability and validity measures of the TD children on the C-BiLLT justifies the calculation of normed data to establish a norm-referenced test for the assessment of spoken language comprehension in children with severe CP and CCN. With these normed data, the C-BiLLT can provide information on (Dutch) spoken language comprehension performance of children with CP and CCN relative to their peers without disabilities, the principal goal of this investigation.
Limitations of the Present Study

Although the C-BiLLT promises to be a valuable instrument, some considerations and limitations need to be mentioned.

For example, a subgroup of about 50 participants is preferable for reliability and validity studies. Although we aimed to collect as much data as possible, the number of children available for investigation of reliability measures (i.e. additional time for repeated measures) depended on the available time of the participating schools and on the availability of children within a certain age group of the different grades of the participating schools and rehabilitation centers. For inter-observer reliability (in children with TD, and with CP and CCN) less than 50 children participated. Also for evaluation of inter-observer reliability in children with CP and CCN a smaller subgroup than the preferred number of 50 participants was available. However, it should be noted that children with CP and CCN represent a small population that is often difficult to recruit. In addition, because children with CP and CCN are occupied with a relatively full schedule of rehabilitation/education, their available time is limited. Nevertheless, because these subgroups of children were representative for the whole age groups of children with TD and CP, and because ICCs showed excellent results with small CIs, these subgroups were found to be sufficient.

From the distribution of raw scores of the TD group a set of scaled scores can be derived. According to the “classical” norming procedure, tests intended for evaluative purposes and less important decisions (e.g. evaluation of language performance), norm groups smaller than 200 are insufficient. However, unlike the rules that apply to “classical norming” procedures, continuous –norming produces the same accuracy using smaller individual norm groups than with classical norming. Bechger, Hemker & Maris showed that a norm group size of about 70 in the continuous approach (for eight groups) produced the same accuracy as a norm group of 200 in the classical approach. We intend to use a continuous-norming procedure for the C-BiLLT. However, in the outer groups (in the present study the groups who were younger than 2, or between the ages of 6;6 and 7;6 years) accuracy is less than in the middle groups. Children not included in the dataset (because of inattentive behaviour, shyness etc.) were predominantly the younger ones. As a result the youngest age group of TD children was smaller than 70.

Furthermore, items of the C-BiLLT were developed for investigation of Dutch spoken language comprehension up to the chronological age of 6;6 years. This age limit was chosen as receptive grammar development is generally considered to be complete by the age of 5-6 years. As a consequence fewer children older than 7 years were included in the present study explaining the small age group of 7 years. However, recent studies on language comprehension have shown that particular aspects of complex syntax continues to develop until early adolescence. Specifically for the children with CP and CCN older than 6;6 years who achieved maximum C-BiLLT scores, ceiling effects may have occurred. Moreover, items of the C-BiLLT did not discriminate between the age groups of 6 and 7 years.
in TD children and ceiling effects were observed in individual cases from the age of 6 years. Therefore, items of the C-BiLLT should be expanded to guarantee a more comprehensive representation of the development of spoken language comprehension. To establish this, and to establish larger outer groups, additional data collection for the youngest as well as the oldest group is currently taking place in our department.

We anticipated that the C-BiLLT might not be administrable to all children with CP. Despite adjustments to optimize test circumstances to the individual child’s conditions, children with severe (cerebral) visual impairment, or in whom there was uncertainty about their vision, were excluded. Children with severe visual impairment may experience difficulties because all answer alternatives are visually presented. In addition, our decision to use a wide 19 inch monitor display may not solve this issue entirely for children with mild visual impairments. Continued work is needed to develop access and assessment techniques to investigate the spoken language comprehension of children with severe and multiple disabilities.

CONCLUSION
The present study had three major aims. First to investigate the clinical usefulness of the extended version of the C-BiLLT as a measure of comprehension of spoken Dutch language in children with severe CP and CCN. Second, to provide information on their performance relative to their peers without disabilities by using test items and response activities that were specifically designed for children with severe physical disabilities. And third, to provide information about the reliability, internal consistency and validity of the C-BiLLT.

Few markers of early delay in spoken language comprehension are available, particularly for children with CCN and physical disabilities. However, early insight in a child’s language comprehension skills is important because it provides parents and other communication partners with guidelines as to which level of spoken language input is appropriate for the child. In addition, it can provide valuable information on how augmentative and alternative communication systems can be introduced in order to best support communication growth and development. Emerging evidence suggests that if young children at risk for CCN receive appropriate augmented language experience early in the course of their development, the communication skills of the child may have the chance to follow a more ‘normalized’ developmental route.

The current study provides evidence of good reliability (as measured by ICC and SEM), and good validity as measure of spoken Dutch language comprehension (as measured by factor analysis and correlation coefficients) of the C-BiLLT. The testing activities of the C-BiLLT were also accessible for children who are often unable to participate in standardized assessment (97% completed C-BiLLT assessment).

When normed data are further established, performance on the C-BiLLT may provide well-founded information of the language comprehension abilities of an individual child with severe CP and CCN relative to the development of peers without disabilities (a norm-
referenced test). C-BiLLT results may also provide information on the child’s comprehension of types of language structures that are typical of that stage of development. As such, it may provide parents and professionals with important information on a child’s language comprehension skills. With its standardized assessment procedure and different access methods, the C-BiLLT shows promise as a reliable measurement of spoken Dutch language comprehension in children with CP and CCN. Adequate and early insight into a child’s language comprehension skills can provide valuable information to optimize the child’s language environment in daily activities and participation.

In order to develop a fuller understanding of the performance of children without disabilities and thus a well-founded basis to interpreted deviant performances on the C-BiLLT assessment activities, the collection and calculation of norm data for the C-BiLLT is justified. The construction of normed data is currently underway in our department.

Future Research
In the near future, item-response theory will be performed to establish an adaptive C-BiLLT test. Moreover, continued work in establishing norm data will mean that administration of the C-BiLLT can be used to follow the developmental course of language comprehension more consistently and contribute to the understanding of language development in children with CP and CCN. In addition, to address information on performance across time the responsiveness of the C-BiLLT will be investigated with follow-up research of comprehension of spoken language in children with severe CP and CCN. In addition, it is our intention to investigate the introduction of C-BiLLT assessments in children with progressive neurological conditions to monitor final stages of language deterioration.

The development of a reliable measure of spoken language comprehension for individuals with severe disabilities is a large and challenging process, but is an important need in many language communities. First contacts have been made for international collaboration and translation of the C-BiLLT into English and German. We are also continuing work on ensuring that the C-BiLLT can be administered using procedures that address both the need for standardization of the testing activities (to support interpretation of the results), as well as the need for appropriate individualization (e.g., alternative access techniques).

Acknowledgments
The authors express special thanks to all children and their parents who participated in this study and contributed to the development of the C-BiLLT. We are grateful for the cooperation of rehabilitation centres, schools for special education and day-care centres throughout the Netherlands (Heliomare, Reade, Trappenberg-Merem, de Hoogstraat, Mtylschool de Regenboog, Mtylschool Amsterdam, Mtylschool de Thermiek, Mtylschool Arianne de Ranitz, Revant [locatie Zeeland], Leypark, Rijndam, Rijnland, Revalidatiecentrum Gouda,
OKDC Nifterlake, de Kokkel, de Blauwe Vogel). We are grateful for the cooperation of kindergartens and primary schools throughout the Netherlands.

The authors thank the following persons for assisting with the development of the C-BiLLT, and the collection and management of data: Valentine van den Ancker, Rieneke Baaijens, Lieke Berkhout, Kelly Blackstone, Erica Boone, Nicole Buijze, Vanessa Goos, Aisha de Graaff, Bernadette Haquebard, Laurike Harlaar, Nancy Kersch, Maaike de Kleijn, Jonne Kronemeijer, Kristin Maria, Marieke van der Meijden, Suzanne Mijnans, Katja Schmitt, Karen Schol, Marloes Stam, and Jorieke Valkenburg.

References


(38) Euser AM, le CS, Finken MJ, Wit JM, Dekker FW. Reliability studies can be designed more efficiently by using variance components estimates from different sources. *J Clin Epidemiol* 2007;60:1010-4.


Appendix A. ICC and SEM

Intra-observer design

The linear random effects model for the intra-observer design is
\[ Y_{or} = m + a_{or} + b_o + e_{or}, \]
with
\[ Y_{or} \] the response of subject \( s \) by observer \( o \) at the \( r \)-th replication (\( r = 1, 2 \)),
\( m \) a constant (fixed parameter),
\( a_{or} \) the random subject effect, nested within observer,
\( b_o \) the random observer effect, and
\( e_{or} \) the measurement error.

It is assumed that \( a_{or}, b_o \) and \( e_{or} \) are independent normally distributed each with mean zero and variances \( s_{so}, s_o, s_e \), respectively. The intra-observer reliability is defined as the correlation between any pair of measurements by the same observer on the same subject:
\[ ICC_{intra} = Corr(Y_{or}, Y_{or'}) = \frac{s_{so} + s_o}{s_{so} + s_o + s_e}. \]
(Vangeneugden et al, 2005).

The variance components can be estimated by restricted maximum likelihood (REML), which also yields standard errors of the estimated variance components. To obtain a more closely normally distributed statistic, \( ICC_{intra} \) is transformed using the Fisher-z transformation
\[ z = \frac{1}{2} \log \frac{1 + ICC_{intra}}{1 - ICC_{intra}} = \frac{1}{2} \log \left( \frac{s_{so} + 2s_o + s_e}{s_{so} + s_o + s_e} \right). \]

As in Euser et al (2007), let the vector of variance components be \( \mathbf{t} = (s_{so}, s_o, s_e)' \) where the prime denotes transposition. The derivative of \( z \) with respect to \( \mathbf{t} \) is
\[ \frac{\partial z}{\partial \mathbf{t}} = \frac{1}{2s_{so} + 2s_o + s_e} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -s_o/(s_{so} + s_o + s_e) \\ 0 & -s_o/(s_{so} + s_o + s_e) & s_e/(s_{so} + s_o + s_e) \end{pmatrix}. \]

Applying the first-order delta method yields
\[ \text{vár}(\hat{z}) = \frac{\partial z}{\partial \mathbf{t}} \text{cov}(\hat{\mathbf{t}}) \frac{\partial z}{\partial \mathbf{t}}^T, \]
where \( \text{cov}(\hat{\mathbf{t}}) \) is the estimated covariance matrix of \( \hat{\mathbf{t}} \) and \( \partial z / \partial \mathbf{t} \) is evaluated at \( \mathbf{t} = \hat{\mathbf{t}} \). A 95% confidence interval (CI) for \( z \) is
\[ (\hat{z} - 1.96\sqrt{\text{vár}(\hat{z})}, \hat{z} + 1.96\sqrt{\text{vár}(\hat{z})}) \]
and back-transformation yields a 95% CI for \( ICC_{intra} \):
\[ \left( \exp(2\hat{L}) - 1, \exp(2\hat{U}) - 1 \right) / \left( \exp(2\hat{L}) + 1, \exp(2\hat{U}) + 1 \right). \]

For any ICC, the standard error of measurement (SEM) is defined as
\[ SEM = s_{Y} \sqrt{1 - ICC}, \]
where \( s_{Y} = \sqrt{s_{so} + s_o + s_e} \) is the standard deviation of \( Y \). Hence, the intra-observer SEM equals \( SEM_{intra} = \sqrt{s_e} \).
Inter-observer design:
The model for the inter-observer data is
\[ Y_{so} = m + a_s + b_o + e_o \]
with
- \( Y_{so} \): the response of subject \( s \) by observer \( o \),
- \( m \): a constant (fixed parameter),
- \( a_s \): the random subject effect,
- \( b_o \): the random observer effect, and
- \( e_o \): the measurement error.

It is assumed that \( a_s \), \( b_o \), and \( e_o \) are independent normally distributed each with mean zero and variances \( s^2_s \), \( s^2_o \), and \( s^2_e \), respectively. The inter-observer reliability is defined as the correlation between any pair of measurements by different observers on the same subject:
\[ ICC_{\text{inter}} = \text{Corr}(Y_{so}, Y_{sw}) = \frac{s^2_s}{s^2_s + s^2_o + s^2_e}. \]

A CI for \( ICC_{\text{inter}} \) can be obtained by the same reasoning as above, and coincides exactly with the design and method in Euser et al (2007).

Finally, the inter-observer standard error of measurement (\( SEM_{\text{inter}} \)) is defined as:
\[ SEM_{\text{inter}} = \sqrt{s^2_o + s^2_e}. \]