University of Groningen

The Psychometric Evaluation of a Speech Production Test Battery for Children

van Haaften, Leenke; Diepeveen, Sanne; van den Engel-Hoek, Lenie; Jonker, Marianne; de Swart, Bert; Maassen, Ben

Published in:
Journal of Speech Language and Hearing Research

DOI:
[10.1044/2018_JSLHR-S-18-0274](https://doi.org/10.1044/2018_JSLHR-S-18-0274)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

van Haaften, L., Diepeveen, S., van den Engel-Hoek, L., Jonker, M., de Swart, B., & Maassen, B. (2019). The Psychometric Evaluation of a Speech Production Test Battery for Children: The Reliability and Validity of the Computer Articulation Instrument. *Journal of Speech Language and Hearing Research*, 62(7), 2141-2170. https://doi.org/10.1044/2018_JSLHR-S-18-0274

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Research Article

The Psychometric Evaluation of a Speech Production Test Battery for Children: The Reliability and Validity of the Computer Articulation Instrument

Leenke van Haaften,^a Sanne Diepeveen,^{a,b} Lenie van den Engel-Hoek,^a
Marianne Jonker,^c Bert de Swart,^{a,b} and Ben Maassen^d

Purpose: The aims of this study were to assess the reliability and validity of the Computer Articulation Instrument (CAI), a speech production test battery assessing phonological and speech motor skills in 4 tasks: (1) picture naming, (2) nonword imitation, (3) word and nonword repetition, and (4) maximum repetition rate (MRR).

Method: Normative data were collected in 1,524 typically developing Dutch-speaking children (aged between 2;0 and 7;0 [years;months]). Parameters were extracted on segmental and syllabic accuracy (Tasks 1 and 2), consistency (Task 3), and syllables per second (Task 4). Interrater reliability and test–retest reliability were analyzed using subgroups of the normative sample and studied by estimating intraclass correlation coefficients (ICCs). Construct validity was investigated by determining age-related changes of test results and factor analyses of the extracted speech measures.

Results: ICCs for interrater reliability ranged from sufficient to good, except for percentage of vowels correct of picture naming and nonword imitation and for the MRRs for bisyllabic and trisyllabic items. The ICCs for test–retest reliability were sufficient (picture naming, nonword imitation) to insufficient (word and nonword repetition, MRR) due to larger-than-expected normal development and learning effects. Continuous norms showed developmental patterns for all CAI parameters. The factor analyses revealed 5 meaningful factors: all picture-naming parameters, the segmental parameters of nonword imitation, the syllabic structure parameters of nonword imitation, (non)word repetition consistency, and all MRR parameters.

Conclusion: Its overall sufficient to good psychometric properties indicate that the CAI is a reliable and valid instrument for the assessment of typical and delayed speech development in Dutch children in the ages of 2–7 years.

A major task for speech-language therapists (SLTs) is to differentiate children with delayed or disordered speech development from typically developing peers and to determine eligibility for speech services.

For such an assessment, they should be able to rely on standardized tests and normative data. However, several reviews that evaluated the content and psychometric characteristics of speech assessments in English and other languages (Flipsen & Ogiela, 2015; Kirk & Vigeland, 2014; McCauley & Strand, 2008; McLeod & Verdon, 2014) concluded that, overall, the diagnostic tests reported on tend to lack fundamental psychometric properties, while sample sizes used for norm referencing were inadequate, and evidence of reliability and validity were poorly described.

Various speech assessments are available for the Dutch language. A survey by our research group (Diepeveen, Van Haaften, Terband, De Swart, & Maassen, submitted) revealed that the vast majority of SLTs (75.8%) in the Netherlands use “LOGO-Art Dutch Articulation Assessment” (Nederlands ArticulatieOnderzoek; Baarda, de Boer-Jongsma, & Jongsma, 2013), with 50.0% (also

^aDepartment of Rehabilitation, Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Center, Nijmegen, the Netherlands

^bHAN University of Applied Sciences, Nijmegen, the Netherlands

^cDepartment for Health Evidence, Radboud University Medical Center, Nijmegen, the Netherlands

^dCenter for Language and Cognition, Groningen University, the Netherlands

Correspondence to Leenke van Haaften:

Leenke.vanHaaften@radboudumc.nl

Editor-in-Chief: Julie Liss

Editor: J. Scott Yaruss

Received July 9, 2018

Revision received November 1, 2018

Accepted December 10, 2018

https://doi.org/10.1044/2018_JSLHR-S-18-0274

Disclosure: The authors have declared that no competing interests existed at the time of publication.

using the Dutch version of the Metaphon Screening Assessment (Leijdekker-Brinkman, 2002). The Dutch version of the Hodson Assessment of Phonological Patterns is used by 31.1% (Van de Wijer-Muris & Draaisma, 2000), whereas another 30.3% evaluate a spontaneous speech sample; 22.7% administer the “Dyspraxia Program” similar to the Nuffield Dyspraxia Program (Eurlings-van Deurse, Freriks, Goudt-Bakker, Van der Meulen, & Vries, 1993), 13.6% use oral motor assessments, whereas 9.80% use a qualitative observation based on the Motor Speech Hierarchy framework used for PROMPT therapy: Verbal Motor Production Assessment for Children (Hayden, 2004), and 8.30% use the Articulation subtest (Klankarticulatie subtest) of the TAK (Taaltoets Alle Kinderen), a Dutch Language Proficiency Test for All Children (Verhoeven & Vermeer, 2001), with 4.50% employing their own (custom-made) speech assessments. None of these assessments is norm based or provides information about reliability and validity except for the TAK (Verhoeven & Vermeer, 2001). For the TAK, normative data are provided based on a representative normative group of 807 children with an age range of 4;7–8;3 (years; months), and the manual states that the test’s reliability and validity were sufficient to good (Verhoeven & Vermeer, 2006).

Moreover, all tests measure only one aspect of speech production. The production of speech sounds is a complex process that comprises both linguistic (or phonological) and speech motor aspects. Psycholinguistic models of speech production describe speaking as a series of sequential and parallel processes, where the first is the conceptualization of a preverbal message, either from memory or from perception, as occurs in picture naming. The next process is formulating a word or sentence, which is driven by two steps of lexicalization: the selection of a *lemma*, containing meaning and grammatical information, and the corresponding *lexeme* or word form. The *lexeme* constitutes the input for the next stage, phonological encoding, during which the sequence of speech sounds is specified together with the syllabic and prosodic structures. Syllables are the basic units of the next process: articulomotor planning and programming. The final process is execution, where the articulatory movements are performed, resulting in an acoustic speech signal (Maassen & Terband, 2015). Children with speech production deficits or speech sound disorders (SSDs) can experience problems at the level of lexeme retrieval, phonological encoding, articulomotor planning and programming, and/or execution. Speech assessment should evaluate these different aspects of SSDs to be able to obtain a complete speech profile. The LOGO-Art Dutch Articulation Assessment (Baarda et al., 2013) analyzes speech in terms of substitution errors in initial, medial, and word-final positions (three-position test). The Articulation subtest of the TAK (Verhoeven & Vermeer, 2001) comprises a word imitation test, with each of its 45 items being dichotomously scored (correct or incorrect) without any further analysis of speech errors. The Dutch version of the Metaphon Screening Assessment (Leijdekker-Brinkman, 2002) and Hodson Assessment of Phonological Patterns (Van de Wijer-Muris & Draaisma, 2000) are scored based

on phoneme inventories and the analysis of phonological processes. The Dyspraxia program (Eurlings-van Deurse et al., 1993) and the Verbal Motor Production Assessment for Children (Hayden, 2004) assess speech motor abilities such as sequencing. To date, there is no Dutch test that systematically assesses speech performance using a broad set of tasks while providing norm data that allow a speech profile to be compiled. Such a comprehensive speech profile is the first step toward a process-oriented diagnosis in which underlying deficits are identified. Because the available diagnostic tools merely yield a description at the symptom level without assessing other aspects of speech production or providing norm-referenced scores, we developed the Computer Articulation Instrument (CAI).

The CAI is a computer-based speech production test battery consisting of four tasks that we based on a series of studies in children with developmental and acquired SSDs (Nijland, 2003; Thoonen, 1998). It has a modular structure and requires interactive administration. Gauging both phonological and speech motor skills of children in the ages between 2 and 7 years, the tasks comprise (a) picture naming, (b) nonword imitation, (c) word and nonword repetition, and (d) maximum repetition rate (MRR). As demonstrated in Figure 1, picture naming taps into the whole chain of speech processes, from preverbal visual-conceptual processing to lemma access, word form selection, phonological encoding, motor planning, and articulation (motor execution; Maassen & Terband, 2015). During nonword imitation, a child is asked to reproduce nonwords (or nonsense words). In contrast to picture naming, a child cannot revert to its lexicon during this task and thus either analyzes the phonological structure of the nonword directly, addressing the phonological encoding system, or follows the auditory-to-motor planning pathway. In word and nonword repetition, a child is asked to repeat a word or nonword five times. This task aims to assess variability in speech production, which occurs when a child uses multiple productions of the same word or nonword. MRR is a pure motor task (articulomotor planning and programming) and does not require any knowledge of words, syllables, or phonemes. In the CAI, the evaluation of speech production is based on phonetic transcriptions and acoustic measurements. Both the tasks and speech analyses are computer implemented. Further explanation of the rationale of the speech tasks and administration procedures are presented in the Method section.

With the CAI, we sought to develop a speech assessment instrument that allows the detection of signs of delay or deviance in several speech production characteristics such that a norm-referenced speech profile for Dutch-speaking children could be obtained. Our ultimate goal for the CAI is to identify and classify children with SSDs. In this article, we will discuss the content of the instrument and the collection of normative data, as well as report on its psychometric properties in terms of interrater and test-retest reliability and its construct validity. Defining reliability as “the degree to which the measurement is free from measurement error” (Mokkink, Terwee, Patrick, et al.,

Figure 1. The speech production processes assessed in the four tasks of the Computer Articulation Instrument (based on Maassen & Terband, 2015, Figure 15.2).

2010), we examined the extent to which each constituent task measures the target construct consistently across time (test–retest) and across raters on the same occasion (interrater; Mokkink, Terwee, Patrick, et al., 2010).

The second aim of this study was to determine the validity of the CAI, which we defined as the degree to which it truly measures the construct it purports to measure (Mokkink, Terwee, Patrick, et al., 2010). In most situations, the first step in test construction is aimed at content or face validity, that is, whether the content of the instrument corresponds with the construct that the instrument is intended to measure. We will demonstrate the content validity of the CAI by the description of the test domain (articulation) and its four speech tasks. The second step in test construction is criterion validity, which refers to how well the scores of the instrument agree with the scores on a gold standard (De Vet, Terwee, Mokkink, & Knol, 2011). In situations in which there is no gold standard, as is the case for speech development in Dutch, one has to fall back on construct validity, which is defined by the COSMIN (Consensus-based Standards for the selection of health Measurement Instruments) panel as the degree to which the scores of a measurement instrument are consistent with hypotheses (Mokkink, Terwee, Knol, et al., 2010). In our study, we thus investigate two aspects of construct validity.

Because the primary aim of the CAI is to measure speech development, we investigated this first aspect of construct validity by comparing the raw scores or parameters of its tasks in a large sample of typically developing children aged between 2 and 7 years. One of the requirements

of a developmental test is that the outcomes show a correlation with age. We hypothesized that the selected parameters, such as the percentage of consonants correct (PCC), percentage of vowels correct (PVC), percentage of cluster reductions, and percentages of particular correctly produced syllable structures, would reflect typical speech development and would thus show a monotonous improvement with age. In order to examine the second aspect of construct or structural validity (De Vet et al., 2011), we conducted factor analyses based on the assumption that clusters of the selected parameters would reflect different aspects of speech production, either within or across tasks, with the parameters' factor structures contributing to the definition of individual speech profiles.

Method

Participants

A total of 1,524 typically developing children aged between 2;0 and 7;0 participated in the normative study. Stratifying for age, we created 14 groups with a range of 4 months for children aged 2;0–5;11 and a range of 6 months for those aged 6;0–6;11. Table 1 summarizes the characteristics of our sample. For the assessment of speech-language development, each age group of a normative sample should contain at least 100 individuals (Andersson, 2005). As Table 1 shows, all our age groups contained more than 100 children, except for the youngest age group ($n = 72$).

Table 1. Age, gender, and multilingualism for the 14 age groups of the normative sample.

Age group (years;months)	M_{age} (years;months)	Number of children	Gender (<i>n</i>)		Multilingual (<i>n</i>)
			Boys	Girls	
2;0–2;3	2;1	72	30	42	2
2;4–2;7	2;5	102	55	47	1
2;8–2;11	2;9	101	46	55	1
3;0–3;3	3;1	104	52	52	3
3;4–3;7	3;5	110	61	49	3
3;8–3;11	3;9	102	57	45	5
4;0–4;3	4;1	100	55	45	1
4;4–4;7	4;5	115	60	55	3
4;8–4;11	4;9	116	56	60	11
5;0–5;3	5;1	121	66	55	12
5;4–5;7	5;5	128	71	57	5
5;8–5;11	5;9	117	64	53	4
6;0–6;5	6;2	117	69	48	5
6;6–6;11	6;8	119	57	62	4
Total		1,524	799	725	60
% of sample		100	52.4	47.6	3.94

The participants were drawn from 47 nurseries and 71 elementary schools located in four different regions of the Netherlands (see Table 2). The nurseries and schools were sent a letter explaining the purpose of the study and inviting them to participate. All parents of the children in the participating nurseries and schools were given an information letter. After obtaining the signed parental consent form, the child was included in the study. To reach the required number of children for each age group in particular geographic regions, assessors randomly selected children from those for whom parental consent had been obtained.

Criteria for inclusion were no hearing loss and Dutch being the spoken language at the nursery or primary school. The parents and teachers of eligible children were asked to complete a questionnaire about the children's development. Another language than Dutch (e.g., Turkish, Arabic, or German) was spoken at home in 3.9% of the participants, with Dutch being the primary language for 96.7% of these children. To ensure the normative sample was representative of the Dutch population, we also included children with a history of speech and language difficulties ($n = 32$, 2.2%). The 4- to 7-year-old children were recruited between January 2008 and December 2014;

and the children in the younger age group (2–4 years), from March 2011 to April 2015.

Parental socioeconomic status (SES) was based on the social status of the district (zip code area) of the child's nursery or primary school as determined by the Netherlands Institute for Social Research (Knol, Boelhouwer, & Veldheer, 2012). The social status of a district was derived from a number of population characteristics, namely, education, income, and labor market position, with higher scores indicating a higher status for that particular district, with a mean of 0 (see Table 3).

The final sample was representative of the general Dutch population in terms of gender, geographic region, and degree of urbanization (see Tables 1 and 2). For example, in the north of the Netherlands, there are very few intermediately or densely populated areas, which is why all testing in that region was conducted in thinly populated areas.

Material: CAI

Tasks

The CAI consists of four tasks: picture naming, non-word imitation, word and nonword repetition, and MRR.

Table 2. Number of children tested per geographic region and degree of urbanization.

Region	Degree of urbanization			Total (%)
	Thinly populated area (index 1.0–2.6)	Intermediate density area (index 2.7–4.0)	Densely populated area (index 4.1–4.8)	
North	128	0	0	128 (8.40)
East	150	212	0	362 (23.8)
South	252	109	0	361 (23.7)
West	341	159	173	673 (44.2)
Total (%)	871 (57.2)	480 (31.5)	173 (11.4)	1,524 (100)

Table 3. Parental socioeconomic status (SES) of the normative sample.

SES	<i>n</i>	% of sample
< -1	182	11.9
≥ -1 and < 1	1,104	72.4
≥ 1	238	15.6
Total	1,524	100

The tasks were administered by (candidate) SLTs specifically trained in the administration of the CAI (for more details, see Procedure section). All utterances were audio-recorded and stored in the CAI database.

Picture naming. Picture naming consists of 60 items. For each item, the child's utterances are compared with the target words. Picture naming is often used for phonological assessment because of its simplicity and ease of administration. Compared to the assessment of conversational speech, a picture-naming task is more efficient and still provides a good index of phonological ability (Wolk & Meisler, 1998).

We used the 50 words of the Dutch revision of Hodson and Paden (1991) Assessment of Phonological Processes–Revised (Van de Wijer-Muris & Draaisma, 2000) that incorporates the full body of vowel, consonant, cluster, and syllable structure combinations of the Dutch language. The syllable shapes of the target words vary from simple to more complex. Because James, Ferguson, and Butcher (2016) suggest that multisyllabic words add value to picture-based speech testing, we decided to add 10 multisyllabic words with all phonemes occurring twice in different positions in different contexts. Comprising 40 one-syllable words, 13 two-syllable words, 6 three-syllable words, and 1 word with four syllables, our task assesses all Dutch phonemes in all possible syllable positions, except for /g/ because this consonant only occurs in loanwords in Dutch (see Appendix A). For the 4- to 7-year-olds, the words are presented in a random order, whereas for the 2- to 4-year-olds, the consonant–vowel–consonant (CVC) words are presented first, followed by the words with more complex syllable structures.

Both seated in front of a computer screen, the SLT asks the child to name what he or she sees on the color pictures that appear consecutively on the screen. Because it was crucial to elicit a sufficiently large speech sample, the computer reads out a sentence with a semantic cue when the child is unable to name the picture spontaneously. When this semantic cue does not elicit the target word, the computer reads out the target word and asks the child to repeat this out loud. It should be noted that, in the latter imitation condition, the lemma and word-form selection processes possibly play a different role than they do in the other two conditions.

Nonword imitation. Poor nonword imitation is widely used as a clinical marker of heritable forms of specific language impairment (Bishop, North, & Donlan, 1996).

The capacity to imitate nonwords has been largely attributed to phonological memory (Gathercole, Willis, Baddeley, & Emslie, 1994), but other cognitive and linguistic processes are also involved (Smith, 2006), including speech production (Shriberg et al., 2009; Vance, Stackhouse, & Wells, 2005). We included nonword imitation in the CAI to investigate the underlying processes of phonological encoding and motor programming (Vance et al., 2005). As the child needs to create new motor programs when reproducing nonwords, this task can be used to isolate motor programming skills (Vance et al., 2005).

The task required the children to reproduce prerecorded nonwords, with accompanying color pictures of “nonsense figures” being shown on the computer screen to make the task more attractive, especially for the younger children. To ensure that the pictures of the nonsense figures did not influence a visual processing component in recalling nonwords, we used pictures of unfamiliar nonsense figures (see an example in Figure 2). Forty-seven of the nonwords were derived from the “Dyspraxia Program” (Eurlings-van Deurse et al., 1993), an assessment comparable to that of the Nuffield Dyspraxia Program, and 23 from Scheltinga (1998). We added 10 more nonwords whose syllable structures were based on the words we had added to the picture-naming task. The frequency distribution of phonological features is shown in Appendix A. The full task comprises 29 one-syllable nonwords, 35 nonwords with two syllables, and 16 nonwords with three syllables, with the 2- to 3-year-old children being presented with the full set of 80 items, whereas the older children needed to reproduce a subset of 33 bisyllabic and trisyllabic items. If a child failed to respond to an item, an additional live-voice presentation of the stimulus was given.

Word and nonword repetition. Speech variability has been associated with certain types of speech disorders, such as childhood apraxia of speech (CAS; Davis, Jakielski, & Marquardt, 1998; Dodd, 1995; Forrest, 2003; Holm, Crosbie, & Dodd, 2007; Iuzzini-Seigel, Hogan, & Green, 2017) and inconsistent phonological disorders (Dodd, 1995). It has also been documented in typically developing 2- and 3-year-olds (Sosa, 2015).

In this task, children are requested to repeat five prerecorded words and as many nonwords five times (without accompanying pictures). Only one model is provided. Both the word and nonword conditions contain 3 two-syllable and 2 three-syllable items with equal, complex consonant structures (CVC-CCVC, CCVC-CVC, CVCC-CCVCC, CVC-CV-CCV, CV-CV-CVC).

MRR. Also known as diadochokinesis, this is one of the most commonly used oral motor assessments in clinical practice. As it is a pure motor task and does not require any knowledge of words, syllables, or phonemes (Maassen & Terband, 2015), the MRR is exploited to differentiate types of SSDs (Lewis, Freebairn, Hansen, Iyengar, & Taylor, 2004; Murray, McCabe, Heard, & Ballard, 2015; Preston & Edwards, 2009; Rvachew, Hodge, & Ohberg, 2005; Shriberg et al., 2010; Thoonen, Maassen, Gabreels, & Schreuder, 1999; Thoonen, Maassen, Wit, Gabreels, &

Figure 2. Two examples of the pictures of nonsense figures used for the nonword imitation task.

Schreuder, 1996). MRR is especially useful in the differential diagnosis of children with CAS. CAS is a disorder of speech motor programming and planning (Nijland, 2003), and MRR or diadochokinesis is one of the most important quantitative measures that can differentiate CAS from other types of SSDs (Murray et al., 2015; Thoonen et al., 1996).

The MRR requires the child to produce three monosyllabic sequences (/pa/, /ta/, /ka/), two bisyllabic sequences (/pata/, /taka/), and one multisyllabic sequence (/pataka/) as fast and accurately as possible. We used a similar protocol to that developed by Thoonen et al. (1996). The MRR is calculated as the number of syllables produced per second during the child's fastest correct attempt.

Scoring

The recordings of the children's speech productions were scored by the (student) SLTs that administered the test and took about 30–45 min, depending on the experience of the assessor and the number of speech errors of the child.

Picture naming and nonword imitation (phonetic transcription). Each utterance was transcribed using the Logical International Phonetics Programs (LIPP) software (Oller & Delgado, 2000), which allows for the transcription of International Phonetic Alphabet (IPA) via the traditional keyboard, along with user-designed analysis based on featural characterizations of segments. The assessors phonetically transcribed all speech recordings based on the correct target transcriptions by "editing in" the child's production errors. An example is given in Figure 3. To compare the child's performance with the targets, inventories of productions (or occurrences) of particular syllable structures, syllable-initial and syllable-final consonants, and vowel types as well as error counts were derived automatically based on a set of phonetic analysis rules, which are listed in Table 4. Percentages of correct productions were calculated by dividing the number of correctly produced phonemes or syllable structures by the total number of phonemes or syllable structures elicited in each task: PCC in syllable-initial position

(PCCI), PVC, and percentage of correct syllable structure (CVC and consonant–consonant–vowel–consonant [CCVC], respectively). All syllable-initial consonants and all vowels were considered when calculating PCCI and PVC. For PCCI, the percentage of correctly produced consonants was divided by the total number of consonants. Because in our investigations the focus was on phonological and not on phonetic development, both common and uncommon clinical consonant distortions were scored as correct, similar to the PCC-Revised calculation described by Shriberg, Austin, Lewis, McSweeney, and Wilson (1997). The PVC was calculated by dividing the vowels pronounced correctly by the total number of vowels. Error counts of cluster reductions were described as the percentage of reduction of initial consonant clusters from two consonants to one divided by the total number of initial consonant clusters of two consonants (RedClus). In addition, we calculated "Level 4" and "Level 5." As described by Beers (1995), these parameters reflect percentages of the correct production of the two highest phonological complexity levels in typical Dutch phonological development, with Level 4 containing the phonemes /b/, /f/, and /v/ and Level 5 containing the liquids /l/ and /R/, all in syllable-initial position. At least half of the typically developing children in the study by Beers reached Level 5 at the age of 2;6–2;8 (i.e., 75% correct responses).

Word and nonword repetition (variability). In the word trials, children had to repeat the following five true words five times: /kap-støk/, "kapstok," English: coat rack; /vilt-stift/, "viltstift," English: felt-tip pen; /vlixtuyx/, "vliegtuig," English: airplane; /pa-Ra-ply/, "paraplu," English: umbrella; and /te-lə-fon/, "telefoon," English: telephone. In the nonword trials, they had to repeat five items with similar structures (/təp-skIt/, "təpskit"; /xilt-stəxt/, "giltstecht"; /vluyx-tix/, "vluigtiex"; /po-Ro-pla/, "poorooplaa"; and /to-li-fan/, "tooliefaan"). For each (non)word, the number of different forms was determined. A production was identified as "different" when at least one of the phonemes of the target word was produced differently or deleted. For example, /airpane/ and /aiplane/ are two forms of the target word /airplane/. Consistency was

Figure 3. Example of the phonetic transcriptions of a target word and a recorded speech sample used in the Computer Articulation Instrument scoring procedure.

established by dividing the total number of forms (with a maximum of 25) by the total number of productions (maximum of 25). A score of 1 indicates maximum variability. This parameter is a revised version of the proportion of whole-word variability (PWV), as described by Ingram (2002). Only the trials with at least three productions of a (non)word were used for analysis.

MRR. For each trial, the MRR was calculated as the number of syllables produced per second, resulting

in six parameters: MRR-pa, MRR-ta, MRR-ka, MRR-pataka, MRR-pata, and MRR-taka. The fastest correctly produced syllable sequence was used for analysis. To determine the number of syllables and the duration of a trial, the sequence was displayed in a waveform using Praat (Boersma & Weenink, 2016). Only the trials with a minimum of five correct syllables were included in the analysis. Syllable boundaries were determined based on visual and auditory information. The burst of the voiceless plosives

Table 4. Parameters per speech task.

Task	Parameter	Description
PN	PCCI	Percentage of consonants correct in syllable-initial position
	PVC	Percentage of vowels correct
	Level 5	Percentage of correct consonants /l/ and /r/
	RedClus	Percentage of reduction of initial consonant clusters from two consonants to one
	CCVC	Percentage of correct syllable structure CCVC (C = consonant, V = vowel)
NWI	PCCI	Percentage of consonants correct in syllable-initial position
	PVC	Percentage of vowels correct
	Level 4	Percentage of correct consonants /b/, /f/, and /u/
	Level 5	Percentage of correct consonants /l/ and /r/
	RedClus	Percentage of reduction of initial consonant clusters from two consonants to one
WR	CVC	Percentage of correct syllable structure CVC
	CCVC	Percentage of correct syllable structure CCVC
	PWV word	Proportion of whole-word variability: word repetition
	NWR	Proportion of whole-word variability: nonword repetition
	MRR	
MRR	MRR-pa	Number of syllables per second of sequence /pa/
	MRR-ta	Number of syllables per second of sequence /ta/
	MRR-ka	Number of syllables per second of sequence /ka/
	MRR-pataka	Number of syllables per second of sequence /pataka/
	MRR-pata	Number of syllables per second of sequence /pata/
	MRR-taka	Number of syllables per second of sequence /taka/

Note. PN = picture naming; NWI = nonword imitation; WR = word repetition; NWR = nonword repetition; MRR = maximum repetition rate.

was used to localize the onset of a syllable. The first and last syllables were excluded from the analysis. Subsequently, the MRR was calculated by dividing the total number of syllables by the duration of the trial.

Procedure

CAI Administration

The children were tested individually in a quiet room in their own nursery or primary school. The SLT and child were seated side by side at a table on which a laptop computer was placed in a position comfortable for both. Both were wearing headsets, or a speaker and microphone were used. For standardization reasons, the tasks were presented in a fixed sequence: picture naming, word repetition, nonword imitation, nonword repetition, and MRR. Testing took approximately 30 min. Some children were seen twice because of their initial lack of interest or cooperation.

The tasks were administered in the younger age groups (2–4 years) by 14 SLTs and in the older children (4–7 years old) by 110 student SLTs (working in pairs) who were fulfilling their phonetics coursework in the third or fourth (final) year of their program. All were trained in the administration of the CAI by the first two authors, having received precise instructions and training in the scoring procedure (phonetic transcription, consistency evaluation, and MRR). Scoring was performed under the supervision of the first two authors and controlled for reliability. The scoring process was performed by the same SLT or SLT student who administered the test, under supervision of the first two authors. The assessors of the normative study were also used as raters for the reliability study.

Not all children completed all tasks for reasons of shyness or inattentiveness, among other causes. In Table 5, the number of children who completed a task is presented per age group. Incomplete tasks were excluded from the data set. The records of picture naming and nonword imitation were considered incomplete if the number of segments was less than 2 *SDs* below the mean number of segments for the age group. The data for word and nonword repetition were analyzed when a child had produced at least three words or nonwords per trial. For MRR, at least two of the three monosyllabic sequences needed to be correct. Table 5 shows that, from the age of 3;0, more than 60% of the children reached this criterion. Because of the high number of 2- and 3-year-olds not being able to perform the monosyllabic sequences, it was decided to set the lower age boundary for this task at 3;0. Thus, the MRR was calculated based on the data obtained in the children aged 3;0 onwards. After excluding the children who fell more than 2 *SDs* below the mean for picture naming and nonword imitation, the percentage of children with a history of speech-language difficulties ranged from 1.9% (MRR) to 2.1% (picture naming), similar to the percentage of the whole sample (2.2%).

Reliability and Validity Procedure

Interrater reliability and test–retest reliability were determined based on the data sets of subgroups of the total sample. Initially, it was our goal to use 10% of the normative data for these reliability evaluations, compared to other normative studies (Clausen & Fox-Boyer, 2017; Dodd, Holm, Hua, & Crosbie, 2003; Gangji, Pascoe, & Smouse, 2015). However, because of its big volume, we were only able to do so for 4.72%–7.02% of the data. We used a sample size of 63–107 children per parameter (interrater reliability: 67–103 children, test–retest reliability: 63–107 children), thereby far surpassing the sample sizes used in the studies mentioned and complying to the recommended minimum sample size of 50 (De Vet et al., 2011). Giraudeau and Mary (2001) conducted simulation studies, showing that a 95% confidence interval of ± 0.1 is reached with a sample size of 50–100, if the intraclass correlation coefficient (ICC) has a value between .7 and .8. Note that the number of children is not equal for all parameters (see Tables B1–B3 in Appendix B) because not all children completed all test items.

Interrater reliability. Each audio recording was scored independently by two raters: Besides an assessor/rater who had also been involved in the data collection of the normative study, we had an additional, independent rater score all the data selected for the reliability analyses. This rater received the same training as the others but had not taken part in the normative study. Interrater reliability was calculated by comparing the scores of the two raters.

- Picture naming and nonword imitation: The audio recordings of 99 children were randomly selected (6.50% of the full sample with all age ranges being included) and transcribed by 35 (picture naming) and 34 (nonword imitation) raters and compared to the 99 transcriptions of the independent rater.
- Word repetition and nonword repetition: The audio recordings of 72 children were randomly selected (4.72% of the full sample with all age ranges being included) and scored by one rater, whose scores were compared to those of the independent rater.
- MRR: The audio recordings of 103 children were randomly selected (6.76% of the full sample with all age ranges being included) and scored by 33 raters. Their scores were compared to those of the independent rater.

Test–retest reliability. A total of 107 children randomly selected from one nursery and five elementary schools were tested twice (7.02% of the full sample). The subsample included children from all age ranges and geographic regions. The data of 11 of these children were also used in the interrater reliability analysis. To avoid learning effects and effects resulting from natural speech development, Kirk and Vigeland (2014) recommend 1–3 weeks as the preferred time interval between the two tests. In our study, the average interval between the initial test (T1) and

Table 5. Number of children who completed a task, per task and age group.

Age group (years;months)	Normative sample	PN	NWI	WR	NWR	MRR	
	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>	Pass (% <i>n</i> MRR)
2;0–2;3	72	72	65	42	43	59	18 (30.5)
2;4–2;7	102	101	90	56	65	83	26 (31.3)
2;8–2;11	101	101	96	69	73	88	55 (62.5)
3;0–3;3	104	102	97	70	80	93	68 (73.1)
3;4–3;7	110	107	105	80	87	99	65 (65.7)
3;8–3;11	102	101	98	81	88	97	86 (88.7)
4;0–4;3	100	99	99	85	84	88	77 (87.5)
4;4–4;7	115	111	113	102	103	96	90 (93.8)
4;8–4;11	116	112	113	97	97	94	93 (98.9)
5;0–5;3	121	117	117	101	103	107	103 (96.3)
5;4–5;7	128	128	128	112	116	115	111 (96.5)
5;8–5;11	117	116	116	106	108	105	104 (99.0)
6;0–6;5	117	117	117	102	105	108	108 (100)
6;6–6;11	119	119	118	108	109	110	109 (99.1)
Total	1,524	1,503	1,472	1,211	1,261	1,342	1,113 (82.9)
% of sample	100	98.6	96.6	79.5	82.7	88.1	73.0

Note. PN = picture naming; NWI = nonword imitation; WR = word repetition; NWR = nonword repetition; MRR = maximum repetition rate; Pass = correct production of at least two of the three monosyllabic sequences.

the retest (T2) was 3.4 weeks, with a range of 1–13 weeks and a median of 3 weeks, comparable with other studies (Abou-Elsaad, Baz, & El-Banna, 2009; Kirk & Vigeland, 2014; Tresoldi et al., 2015), with 89.4% of the children being retested within 1–5 weeks. All children were in the same age group during the first and second administrations. Both tests were administered by the same assessor, and all four tasks were repeated. Two raters scored the randomized T1 and T2 audio recordings, with the same rater scoring the tests of the same child.

Construct validity. Age trends for all the extracted parameters mentioned in Table 4 were determined, and means per age group were calculated. Next, continuous normalized standard scores were computed based on the model developed by Tellegen and Laros (1993), in which the cumulative proportions of the raw scores across the age groups were simultaneously fitted as a higher-order function of raw score and age. With the exception of MRR, this model was applied separately to the six age groups including those aged 2;0 to 3;11 and the eight groups with children aged 4;0–6;11 for the three speech tasks, mainly because the parameters were not identical for the younger and older age groups. For MRR, the model was applied in one run to the 11 age groups starting from the age of 3;0.

We conducted two factor analyses to determine the component structure of the selected parameters across tasks and the factor scores per task, in which latter scores were obtained such that a test result could also be obtained in cases in which not all tasks had been completed.

Statistical Analyses

Interrater reliability and test–retest reliability were studied by estimating ICCs. For interrater reliability, two-way

random-effects models were fitted with the parameter of interest as the dependent variable and the independent variables as random intercepts to allow for different levels per child and rater. No fixed effects were included in the model. For every parameter, the ICC was estimated, and the corresponding 95% confidence intervals were constructed with the bootstrap method. For the test–retest reliability estimation, three-way random-effects models were fitted; a random intercept for time point (T1 or T2) was also included. Again, ICCs were estimated, and bootstrap confidence intervals were constructed.

The Dutch Committee on Tests and Testing (COTAN) considers reliability coefficients below .70 as insufficient, those between .70 and .80 as sufficient, and estimates higher than .80 as good (Evers, Lucassen, Meijer, & Sijsma, 2010). Besides ICCs, the interrater reliability of the phonetic transcriptions (picture naming and nonword imitation) was examined by calculating point-to-point agreement, with a mean percentage of agreement being reported. For test–retest reliability, Wilcoxon signed-ranks tests were performed to compare the T1 and T2 means.

Means and standard deviations of all parameters were calculated per age group to describe age trends. In addition, error bar graphs were plotted. Parameters were selected based on clinical relevance and monotonous age trends; for these parameters, percentile scores were calculated with the Tellegen and Laros (1993) regression formula containing the raw parameter score, its square, its cube, age, age squared, age cubed, all interaction factors except for the interaction of both cubes, and score cubed with age squared. To see how much variance was explained by the differences between age groups, *R*-squared values were calculated.

To determine the factor structure of the parameters across tasks and per task, a principal component analysis

(PCA) with varimax rotation was conducted on all CAI parameters to identify clusters of items. The Kaiser–Meyer–Olkin (KMO) measure was performed prior to the PCA. Values greater than 0.5 were considered acceptable (Field, 2009). In the PCA, the criterion for eigenvalues to be greater than 1 was used. A factor loading of an absolute value of more than .4 is considered important (Field, 2009). The relationship between tasks was examined with Pearson product–moment correlations.

SPSS Version 20 for Windows (SPSS Inc.) and R Version 3.2.3 were used for all statistical analyses. The estimated ICCs and their bootstrap confidence intervals were computed in RStudio.

Results

Interrater Reliability

Table B1 (see Appendix B) shows the results of the interrater reliability evaluation. For picture naming, the interrater reliability was good for the parameters PCCI, Level 5, RedClus, and CCVC, with ICCs ranging from .80 to .90. The ICC (.59) for PVC was insufficient.

With an ICC of .82, the interrater reliability of the nonword-imitation task was good for RedClus and sufficient for the parameters PCCI, Level 4, Level 5, CVC, and CCVC, with ICCs ranging from .71 to .77. Interrater reliability (ICC of .62) was insufficient for PVC. With .73, the ICC for word repetition (PWV word) was sufficient but insufficient for nonword repetition (PWV nonword: .25).

Interrater reliability was good for the monosyllabic sequences MRR-pa and MRR-ka (ICCs of .81 and .83, respectively) and sufficient for MRR-ta (ICC of .77). With ICCs ranging from .41 to .62, interrater reliability estimates for MRR-pataka, MRR-pata, and MRR-taka were insufficient.

Table B2 shows the results of the point-to-point interrater agreement of the phonetic transcriptions of the picture naming and nonword imitation responses. The agreement of segments (total number of consonants, vowels, and word and syllable boundaries), consonants, and vowels was high for both tasks.

Test–Retest Reliability

Table B3 shows the results of the test–retest reliability analysis. Picture naming showed sufficient reliability for the parameters Level 5, RedCLus, and CCVC (ICCs of .71, .75, and .76, respectively), but this was insufficient for PCCI and PVC (ICCs of .51 and .31, respectively).

The nonword-imitation task showed a good test–retest reliability for CVC (ICC of .88) and sufficient estimates for PCCI (ICC of .74), PVC (ICC of .77), and Level 5 (ICC of .73). The other parameters scored insufficient, with ICCs ranging from .41 to .60.

Insufficient test–retest values were found for word and nonword repetition (ICCs of .66 and .39, respectively) as well as for MRR, where ICCs ranged from .18 to .60,

except for MRR-pa with a sufficient reliability score (ICC of .70). As Table B3 also shows, in picture naming, the T2 scores for PCCI, PVC, and Level 5 were significantly higher than the T1 scores, as was the case for the nonword-imitation parameters PCCI, PVC, Level 4, Level 5, CVC, and MRR-pataka.

Validity

Age Trends and Continuous Norms

The means and standard deviations of the picture-naming and nonword-imitation parameters are presented in Table B4 for the younger age group (2;0–3;11) and in Table B5 for the older age group (4;0–6;11). Monotonous increases with age were found for all parameters, except for the age range between 4;4 (52 months) and 5;7 (67 months), during which plateaus or only minor increases were observed. Apparently, during this stage, little development takes place.

The means and standard deviations of the parameters of word repetition and nonword repetition are presented in Tables B6 (younger age group) and B7 (older age group). Minor decreases with age were found for the PWV in both tasks. Across age groups, the children were more consistent in producing words than nonwords.

As mentioned in the Method section, the MRR task was completed by children from the age of 3;0 onwards. The mean percentage of children who could produce at least two monosyllabic repetitions was 75.8% at the age of 3;0–3;11, 93.5% at the age of 4;0–4;11, 97.2% at the age of 5;0–5;11, and 99.5% at the age of 6;0–6;11. Table B8 shows steady increases in the rates for monosyllabic, bisyllabic, and trisyllabic sequences with increasing age.

Appendix C shows examples of the modeling results for the main parameters. Criteria to accept the regression model are (a) an increase of percentile scores with increasing raw scores given a particular age and (b) a decrease of percentile scores with increasing age given a particular raw score. In the graph, lines per age group must thus show monotonous increases in percentile scores (*y*-axis) with increasing raw scores (*x*-axis) and should not cross. Because at the tails of the distribution these conditions are not always met, minimum and maximum values were determined for the raw scores such that the model is adequate within these limits. This implies that, within the first to fifth percentile range and within the 95th–100th percentile range, fine discrimination is lost, in which loss is considered clinically acceptable. *R*-squared values per parameter were all .96 or higher, indicating that the model-predicted percentile score based on raw score and age explained 96% of the variance in the raw scores, which can be considered an excellent match.

Factor Analysis

A PCA with orthogonal rotation (varimax) was conducted on the parameters of all four tasks (see Table 6). The KMO measure verified the sampling adequacy for the analysis (KMO = 0.78), and all Cronbach's alphas were

Table 6. Summary of the factor analysis results for picture naming (PN), nonword imitation (NWI), word (WR) and nonword (NWR) repetition, and maximum repetition rate (MRR).

Task	Parameter	Rotated factor loadings				
		Component				
		1	2	3	4	5
PN	PCCI	.24	.019	.75	-.014	.15
	PVC	.21	-.017	.67	-.053	.031
	RedClus	.040	.025	.77	.25	-.060
	Level 5	.22	.044	.72	-.016	.21
	CCVC	.026	.014	.80	.20	-.090
NWI	PCCI	.87	.031	.17	.12	.16
	PVC	.75	.038	.084	.091	-.011
	RedClus	.21	.023	.15	.85	.092
	Level 4	.76	.033	.086	.097	.14
	Level 5	.67	-.001	.26	.026	.17
	CVC	.67	.008	.14	.17	-.17
	CCVC	.21	.014	.098	.86	.081
	PWV word	.014	.059	.096	.11	.76
	PWV nonword	.15	-.051	.033	.028	.77
MRR	MRR-pa	.035	.73	.002	-.021	-.056
	MRR-ta	-.015	.77	.034	.075	.016
	MRR-ka	-.004	.74	.024	.046	.010
	MRR-pataka	.076	.59	-.004	-.063	.025
	MRR-pata	-.025	.71	.026	-.047	-.017
	MRR-taka	.023	.69	-.006	.075	.036
	Eigenvalues	4.62	2.99	1.89	1.33	1.23
% of variance		23.1	14.9	9.45	6.66	6.16
Cronbach's α		.83	.80	.82	.81	.45
Extraction method: principal component analysis						
Rotation method: varimax with Kaiser normalization						
Rotation converged in five iterations						

Note. PCCI = percentage of consonants correct in syllable-initial position; PVC = percentage of vowels correct; RedClus = percentage of reduction of initial consonant clusters from two consonants to one; Level 5 = percentage of correct consonants /l/ and /r/; CCVC = percentage of correct syllable structure CCVC; Level 4 = percentage of correct consonants /b/, /f/, and /v/; CVC = percentage of correct syllable structure CVC; PWV word = proportion of whole-word variability: word repetition; PWV nonword = proportion of whole-word variability: nonword repetition; MRR-pa = number of syllables per second of sequence /pa/; MRR-ta = number of syllables per second of sequence /ta/; MRR-ka = number of syllables per second of sequence /ka/; MRR-pataka = number of syllables per second of sequence /pataka/; MRR-pata = number of syllables per second of sequence /pata/; MRR-taka = number of syllables per second of sequence /taka/.

.80 or higher, except for PWV (.45). The five-component solution explained 60.3% of the variance. All picture-naming parameters are reflected by the third component. The first component mainly reflects the segmental parameters of nonword imitation (NWI-seg); and the fourth component, the nonword imitation cluster-reduction and syllable-structure parameters. The MRR trials form an independent component (Component 2), with loadings on the other components smaller than .1. Adding the PWV did not change the structure of these four components; the PWV measures form an additional component (Component 5) and do not load on the other components. Factor analyses without the MRR parameters and without the PWV measures resulted in an identical structure of picture-naming and nonword-imitation components.

Because in the complete factor analysis there were relatively many missing values, and to obtain composite performance scores, a second series of factor analyses

for each task separately was conducted with the same three components being found for picture naming and nonword imitation, explaining 63.9% of the variance. Word and nonword repetition (PWV) and monosyllabic and multisyllabic MRRs both yielded a one-component solution, explaining 64.4% and 46.9% of the variance, respectively. The lower percentage for MRR is partly due to missing values in the multisyllabic sequences, which had to be estimated; in the factor analysis of the three monosyllabic sequences only, 68.4% of the variance was explained.

In order to test any remaining influence of gender and demographic variables on the factor scores, separate analyses of variance were conducted with each of the five factors as the dependent variables and gender (girl, boy), age group (14 categories), and covariate SES (standardized value) as the independent variables. Significant gender effects were found for the factors picture naming, $F(1, 1321) = 8.16$, $p = .004$; MRR, $F(1, 989) = 12.3$, $p < .001$; and consistency

of repetition (PWV), $F(1, 1155) = 5.02, p = .025$, with the girls performing slightly better on the picture naming and consistency of repetition (PWV) factors and the boys performing slightly better on the MRR. The factor SES was significant for picture naming, $F(1, 1321) = 31.8, p < .001$, and consistency of repetition (PWV), $F(1, 1155) = 9.84, p = .002$.

Correlation Coefficients

Pearson correlations for the different tasks are shown in Table 7. Weak but significant correlations were found between picture naming and nonword imitation (NWI-seg and syllabic parameters [NWI-syl]) and between picture naming and consistency of repetition (PWV), as well as between nonword imitation (NWI-seg and NWI-syl) and consistency of repetition (PWV). No significant correlations were found between NWI-seg and NWI-syl or between MRR and the other tasks.

Discussion

In this study, we report on the psychometric properties of the CAI, a test battery assessing the development of speech production skills, based on the performance of 1,524 children aged between 2 and 7 years, making it the first norm-referenced standardized test designed for process-oriented diagnostics of spoken Dutch. Based on a selection of the scores on its four constituent tasks, namely, picture naming, nonword imitation, word and nonword repetition, and MRR, we examined the interrater and test-retest reliability of the CAI and its construct validity (factor analyses and correlation coefficients) across age groups, with norm values being established for each parameter separately.

Interrater Reliability of the CAI Tasks

The overall findings of the reliability study (interrater and test-retest) were sufficient to good. The majority of the parameters met an ICC minimum of .70.

Phonetic transcription of children's speech using the IPA is widely used, although the reliability of the method is questioned, especially the transcriptions of the speech of young children, and different agreement percentages are described (Preston & Koenig, 2011; Shriberg & Lof, 1991). This variation is due to different factors, such as the experience of the transcriber and environmental conditions. In many speech assessments, especially in those of nonword-imitation performance, rather than phonetic transcription, a correct-incorrect score is used because, in such dichotomous ratings, the interference of variation or measurement errors is less pronounced. We hence were expecting some variation in our phonetic transcriptions and hence the ICCs. Still, the interrater reliability of most of the picture-naming and nonword-imitation parameters was sufficient to good, with the percentages for point-to-point agreement being high (above 80%) for all measures. We think that the high reliability values we obtained are the result of the use of speech recordings and a broad phonetic transcription, with target transcriptions being provided and access to the acoustic signal for verification purposes being easy (Shriberg & Lof, 1991).

The interrater reliability of the PVC was insufficient for both the picture-naming and nonword-imitation tasks. Arguably, the transcription of vowels is more challenging than the transcription of consonants, which may have two causes: (a) There is greater dialectal variation in Dutch vowels than there is in Dutch consonants, and (b) the categorization of vowels is in general more difficult than that of consonants, although phoneme boundaries are less clearly defined (Howard & Heselwood, 2012; Pollock & Berni, 2001). Lower reliability scores for vowel transcriptions have been reported in other studies (Dodd et al., 2003), but others describe no difference in consonant and vowel agreement (Shriberg & Lof, 1991). Noteworthy here is that these latter studies describe interrater reliability in terms of point-to-point agreement only, which was also good for vowels in our evaluation. Whereas point-to-point agreement only reflects agreement, ICCs reflect both agreement and correlation (Shrout & Fleiss, 1979). The calculation

Table 7. Pearson correlations between tasks.

Task	Correlation	PN	NWI-seg	NWI-syl	PWV	MRR
PN	Pearson correlation	1	.34**	.22**	.15**	.049
	<i>n</i>	1,466	1,351	1,351	1,168	991
NWI-seg	Pearson correlation	.34**	1	-.001	.19**	.045
	<i>n</i>	1,351	1,373	1,373	1,180	980
NWI-syl	Pearson correlation	.22**	-.001	1	.12**	.039
	<i>n</i>	1,351	1,373	1,373	1,180	980
PWV	Pearson correlation	.15**	.19**	.12**	1	.018
	<i>n</i>	1,168	1,180	1,180	1,184	937
MRR	Pearson correlation	.049	.045	.039	.018	1
	<i>n</i>	991	980	980	937	1,012

Note. PN = picture naming; NWI-seg = segmental parameters of nonword imitation; NWI-syl = syllabic parameters of nonword imitation; PWV = proportion of whole-word variability; MRR = maximum repetition rate.

**Correlation of factor scores is significant at the .01 level (two-tailed).

of ICC involves dividing the between-subjects variability (BV) by the total variability (similar to analysis of variance). The total variability can be modeled as consisting of BV plus the within-subject variability or—in case of a reliability study—error variance (EV). In formula: $ICC = BV / [BV + EV]$. This implies that, given a particular value for EV (due to interrater variability or test–retest variability), the ICC is lower if BV is relatively small, as was the case in our normalization data. The PVCs in the typically developing children we tested were high, with little variation (small standard deviation). If the outcomes of the sample show large variability, which we expect to occur when also children with speech disorders are tested, it is easier to distinguish the children from each other despite any measurement errors than when the outcomes differ very little. A next step in the psychometric evaluation of the CAI will be to add children with speech disorders in the reliability study. Because of the important role of PVC in distinguishing typical speech development from SSDs (Forrest, 2003; Iuzzini-Seigel & Murray, 2017), the PVC remains included as a parameter of picture naming and nonword imitation in the CAI.

The ICC for PWV in the word repetition task was sufficient but insufficient for nonword repetition, which may be due to the fact that it is more difficult for raters to judge unfamiliar phonological items than known words. Because the nonword reproductions were not transcribed, the raters had to rely on auditory information only. Here, phonological short-term memory plays a large role, placing higher demands on the rater's working memory capacities. Because of the unfamiliarity of nonwords, it is hypothesized that raters are inclined to listen for more detailed information similarly to what happens with narrow phonetic transcriptions, whereas this manner of transcription proved unreliable (Shriberg & Lof, 1991). In the future, raters are accordingly advised to qualify the differences between target and reproduced nonwords as they do based on broad transcriptions without paying attention to small diacritic differences.

Whereas the few studies that have been conducted previously (Gadesmann & Miller, 2008) describe poor interrater reliability for this task, in our study, the scoring of the monosyllabic items of the MRR was reliable. In most studies, MRR is typically measured by counting syllables and reading the time with a stopwatch. Kent, Kent, and Rosenbek (1987) suggested that interrater reliability would increase if some method of instrumentation were to be used that displayed acoustic waveforms, while standardized instructions and procedures would help reduce variability within and across children. In our study, we indeed used a standardized measurement protocol and an acoustic waveform. The raters judged the speech samples in three steps, supported by a display of the acoustic signal, determining (a) the onset of the second syllable and (b) the onset of the final syllable and then (c) counting the number of syllables, with the duration of a sequence calculated automatically. This procedure potentially explains the high reliability in the scoring of monosyllabic MRRs. In contrast,

the interrater reliability for the bisyllabic and trisyllabic items was insufficient. Several factors might underlie this result. First, we noticed that the younger children had difficulties performing the MRR task, likely because they had difficulties understanding the instructions, whereas a large number of children were not able to perform the task at all. Conversely, there was high agreement in the raters' judgments whether the attempts were successful or not. The data of the children who failed to perform the task were not included in the reliability study. If we had, this would have resulted in higher ICCs. Another influencing factor might be that judging whether the sequences of the bisyllabic and trisyllabic items were produced correctly is more difficult than it is for the monosyllabic items because the younger children made more age-specific errors of pronunciation, as Yaruss and Logan (2002) also noted.

Because we used many different raters for our interrater reliability evaluation, our study may be a good reflection of the professional field, with the ICCs we obtained being representative of clinical practice. On the other hand, it has caused more variability in the ratings and lowered the ICCs.

Test–Retest Reliability of the CAI Tasks

Except for the PCCI and PVC, test–retest reliability was sufficient for the other parameters of picture naming, which means that performance on these parameters was stable over time. Because of the important role of PCCI and PVC in the diagnosis of SSDs, as discussed above for interrater reliability, PCCI and PVC remain included as parameters of picture naming.

Comparable to other studies (Gadesmann & Miller, 2008; Gray, 2003), test–retest reliability proved insufficient for most of the parameters of the other three tasks (nonword imitation, word and nonword repetition, and MRR), with the scores having improved significantly the second time round. Four factors seem to be involved. First, as reflected by the normative data, the children's speech variables changed rapidly over time, becoming more accurate especially in the younger age groups (Dodd et al., 2003; McIntosh & Dodd, 2008). The factor “normal speech development” plays an unexpected major role here and impacted the level of the reliability coefficients directly, even within the relatively short space of 3 weeks (mean time interval between T1 and T2: 3.4 weeks; range: 1–13 weeks), while the increase in the raw scores between T1 and T2 of the five children who repeated the tasks 13 weeks after the first administration did not differ from that observed in the other children.

Second, besides a developmental effect, the effect of learning might be more pronounced in tasks assessing imitation and repetition of words and nonwords than it is in picture-naming paradigms. At the first assessment, most of the children will have had no experience with these kinds of tasks, which thus test new skills. During the retest, the children are more familiar with the procedure, which may have boosted their performance.

Third, the tasks were administered by the same assessor at both time points, and especially, the younger children might have felt more at ease with the assessor during the retest, which may have positively influenced their T2 scores. A fourth factor that might have negatively affected the test–retest reliability in our study is that, in most previous studies, nonword imitation was assessed using whole-word scoring procedures, which tend to yield higher reliability results. Because the purpose of our study was to investigate speech production, we chose to use phoneme-by-phoneme scoring (phonetic transcription), which has the disadvantage that scoring will show more variation among raters. Besides the variability likely attributable to developmental and learning effects, also attention and motivation differences among tests may underlie (part of) the variability. The sample size of both reliability studies (interrater and test–retest reliability) was less than 10% of the normative data, and this is a weakness of this study. Our goal was to use 10% of the data, as was recommended in the literature (Clausen & Fox-Boyer, 2017; Dodd et al., 2003; Gangji et al., 2015). However, we used a sample size of 63–107 children per parameter, thereby far surpassing the sample sizes used in the studies mentioned and complying to the recommended minimum sample size of 50 (De Vet et al., 2011; Giraudeau & Mary, 2001).

Another limitation of the study is the presence of children with a history of speech and language difficulties and children with another primary language than Dutch, in the normative sample. Children with a speech-language delay or different linguistic background are less reliable to transcribe (Shriberg & Lof, 1991). However, it is important to include these groups of children in order to achieve a representative sample, as stated by Dodd et al. (2003).

Validity of the CAI

The content validity of the CAI was demonstrated by the description of the constituent tasks and their items. The distribution of the consonants, vowels, clusters, and syllable structures included is representative of the Dutch language. The items of the four tasks each measure different aspects of speech production. Because of the lack of another comprehensive, norm-referenced speech production assessment instrument in Dutch, we were not able to establish the criterion validity of our test battery.

The increase we reported in the mean scores of almost all parameters of the CAI with chronological age supports its construct validity: Speech production abilities improved with increasing age, as was reported in other studies (Abou-Elsaad et al., 2009; Dodd et al., 2003; Lousada, Mendes, Valente, & Hall, 2012; McIntosh & Dodd, 2008; Priester, Post, & Goorhuis-Brouwer, 2011; Tresoldi et al., 2015; Vance et al., 2005). As Tables B4 and B5 show, the largest increase occurred during the preschool years (i.e., in the 2- to 4-year-olds). Because speech development is generally assumed to be completed by the age of 5 years (Dodd, 2011), lower variation is expected when children grow older. Most parameters of picture naming and nonword

imitation indeed showed a decrease in standard deviations with increasing age, with the higher standard deviations recorded for the younger age groups, where intersubject variation is typical (Dodd, 1995; Stoel-Gammon & Cooper, 1984). In contrast, standard deviations of the MRR parameters were quite stable for all age groups. Here, with the number of syllables per second increasing, variation did not decrease with age, which indicates that speech or phonological development, as gauged with these two tasks, progresses differently from speech motor development, with the processes representing two different aspects of speech development. We have no clear explanation for the relatively low score on nonword imitation PCCI in those aged 4;0–4;3 ($M_{\text{age}} = 49$ months), except that the children aged 4 years and over were offered more complex material. Because modeling was conducted for the children aged 2;0–3;11 and 4;0–6;11 separately, this decrease in scores should not have complicated the interpretation of the normative scores. Noteworthy is the PWV variability score for word and nonword repetition, which only showed minor decreases with age and little between-subjects variation (Cronbach's $\alpha = .448$). We propose that, rather than gauging speech development, this component gauges speech pathology, where the normative data can help to differentiate different types of speech disorders.

Continuous normalized standard scores were calculated for all parameters, and all showed an increase of percentile scores with increasing raw scores for all ages. SLTs can thus use these normative data to discriminate between typically developing children and children with a speech delay or potential speech disorder.

Factor analyses revealed five meaningful factors, based on which as many constructs of speech production could be determined. One component represented all picture-naming parameters; a second component, the segmental parameters of nonword imitation; and a third, the nonword-imitation parameters RedClus and CCVC, both measuring cluster reduction and syllable structure, with the PWV parameters forming an additional, fourth component, best characterized as “(non)word-repetition consistency,” and the fifth component reflecting all MRR parameters. Factor analyses for each task separately confirmed the five factors.

Gender-related effects were found for picture naming, MRR, and consistency of repetition (PWV) but not for nonword imitation. The girls performed slightly better on picture naming and PWV, which is comparable to other studies that also reported gender differences in phonological acquisition, with most finding girls to outperform boys on speech accuracy measures (Dodd et al., 2003; Smit, Hand, Freilinger, Bernthal, & Bird, 1991). As to MRR, the boys performed better than the girls and especially so in the younger age groups, a finding also consistent with another study on speech motor performance, where the oral diadochokinetic rates of boys tended to be faster than those recorded for girls (Prathanee, Thanaviratnanich, & Pongjanyakul, 2003).

CAI's construct validity is underlined by the correlation between picture naming and nonword segmental

accuracy as well as between the picture-naming and non-word-imitation parameters and consistency of repetition (PWV). Monosyllabic and multisyllabic repetition was not correlated with any of the other measures, confirming that MRR is a distinct task.

Future Perspectives

In this study, CAI outcomes were scored based on manual phonetic transcription using LIPP (Oller & Delgado, 2000) and Praat for acoustic measurements (Boersma & Weenink, 2016). To make CAI more user-friendly, in the future version for use in clinical practice (Maassen et al., 2019), all scoring procedures will be automated, where, for instance, the LIPP set of phonetic analysis rules and acoustic waveforms for MRR scoring will have been integrated into the software.

Studies are currently being conducted to determine the role of the CAI in diagnosing speech disorders. We are investigating whether the instrument can differentiate children with typical speech from peers showing signs of an SSD (known-group validity) while allowing pathology profiles to be made for a differential diagnosis, where the subtest scores of children with a suspected deviant speech development are compared with the normative data. Further steps in the implementation of an automated, process-oriented diagnosis of abnormal speech development will include the addition of objective (acoustic) measures of speech production and a process analysis of the outcomes of our assessment battery on the basis of data collected in different speaking conditions (Maassen & Terband, 2015).

A limitation of this study is that the CAI is currently only available in Dutch. However, the CAI has an open structure in that all stimulus materials (spoken instructions, the pictures for the naming task, audio targets for the word and nonword repetition tasks, and audio targets for the MRR task) are stored in separate files. The phonetic transcriptions of the target items in IPA and the rules for analyzing transcribed utterances in relation to the targets are also stored in separate files. A strong asset of the CAI software is the interpreter, which allows for a comprehensive and versatile analysis of transcriptions. This allows the software to be adapted to test the speech development of children speaking other languages than Dutch. When translating the CAI into other languages, new reliability and validity studies should be carried out. In addition, when choosing new test items, the distribution of phonemes of the other language must be taken into account. International collaboration may then contribute to further evaluation and refinement of the instrument.

Conclusion

In this article, we reported the results of a normative study of CAI, a newly developed computer-based speech assessment instrument. The test battery incorporates a picture-naming task and word and nonword reproduction tasks, along with a task assessing MRRs. With these tasks

in the CAI, different aspects of speech production can be evaluated and compared with each other in order to obtain a complete speech profile. The analyses of the phonological measurements, syllabic structures, and speech motor skills yielded indices of typical speech development in Dutch-speaking children in the ages between 2 and 7 years, based on which norm-referenced estimates of speech delay were determined.

Reliability and validity evaluations overall yielded sufficient to good values for interrater reliability. ICCs for test-retest reliability were low due to natural development and learning effects but good for construct validity, indicating that the CAI can be used to gauge typical and atypical speech development.

Acknowledgments

The reported research was partly funded by BOOM Test Publishers, Amsterdam, the Netherlands, and the Damsté-Terpstra Fonds, Zeist, the Netherlands. We thank the children and their parents, nurseries, primary schools, speech-language therapists, and speech-language therapy students who/that participated in this project.

References

- Abou-Elsaad, T., Baz, H., & El-Banna, M. (2009). Developing an articulation test for Arabic-speaking school-age children. *Folia Phoniatrica et Logopaedica*, 61(5), 275–282. <https://doi.org/10.1159/000235650>
- Andersson, L. (2005). Determining the adequacy of tests of children's language. *Communication Disorders Quarterly*, 26(4), 207–225. <https://doi.org/10.1177/15257401050260040301>
- Baarda, D., de Boer-Jongsma, N., & Jongsma, B. (2013). *LOGO-Art. Nederlands Articulationonderzoek [LOGO-Art Dutch Articulation Assessment]*. Dronten, the Netherlands: Uitgeverij LOGO-Start.
- Beers, M. (1995). *The phonology of normally developing and language-impaired children*. IFOTT, University of Amsterdam, Amsterdam, the Netherlands.
- Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *The Journal of Child Psychology and Psychiatry*, 37(4), 391–403. <https://doi.org/10.1111/j.1469-7610.1996.tb01420.x>
- Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer* [Computer program]. Version 6.0.21. Retrieved from <http://www.praat.org/>
- Clausen, M. C., & Fox-Boyer, A. (2017). Phonological development of Danish-speaking children: A normative cross-sectional study. *Clinical Linguistics & Phonetics*, 31(6), 440–458. <https://doi.org/10.1080/02699206.2017.1308014>
- Davis, B. L., Jakielski, K. J., & Marquardt, T. P. (1998). Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics & Phonetics*, 12(1), 25–45.
- De Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. New York, NY: Cambridge University Press.
- Diepeveen, S., Van Haften, L., Terband, H., De Swart, B., & Maassen, B. (submitted). *Clinical reasoning in the field of SSD: The what's and why's regarding diagnosis and intervention in SLTs' daily practice*.

- Dodd, B.** (1995). *Differential diagnosis and treatment of children with speech disorder*. London, United Kingdom: Wiley.
- Dodd, B.** (2011). Differentiating speech delay from disorder: Does it matter. *Topics in Language Disorders*, 31(2), 96–111. <https://doi.org/10.1097/TLD.0b013e318217b66a>
- Dodd, B., Holm, A., Hua, Z., & Crosbie, S.** (2003). Phonological development: A normative study of British English-speaking children. *Clinical Linguistics & Phonetics*, 17(8), 617–643. <https://doi.org/10.1080/0269920031000111348>
- Eurlings-van Deurse, M., Freriks, A., Goudt-Bakker, K., Van der Meulen, S., & Vries, L. D.** (1993). *Dyspraxieprogramma: Therapieprogramma voor kinderen met kenmerken van een verbale ontwikkelingsdyspraxie [Dyspraxia Program: A therapy program for children with childhood apraxia of speech] (Dyspraxia Program)*. Lisse, the Netherlands: Swets & Zeitlinger.
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K.** (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests [COTAN Assessment system for the quality of tests]*. Amsterdam, the Netherlands: NIP.
- Field, A.** (2009). *Discovering statistics using SPSS*. London, United Kingdom: Sage.
- Flipsen, J. P., & Ogiela, D. A.** (2015). Psychometric characteristics of single-word tests of children's speech sound production. *Language, Speech, and Hearing Services in Schools*, 46, 166–178. https://doi.org/10.1044/2015_lshss-14-0055
- Forrest, K.** (2003). Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists. *American Journal of Speech-Language Pathology*, 12(3), 376–380. [https://doi.org/10.1044/1058-0360\(2003\)083](https://doi.org/10.1044/1058-0360(2003)083)
- Gadesmann, M., & Miller, N.** (2008). Reliability of speech diadochokinetic test measurement. *International Journal of Language & Communication Disorders*, 43(1), 41–54. <https://doi.org/10.1080/13682820701234444>
- Gangji, N., Pascoe, M., & Smouse, M.** (2015). Swahili speech development: Preliminary normative data from typically developing pre-school children in Tanzania. *International Journal of Language & Communication Disorders*, 50(2), 151–164. <https://doi.org/10.1111/1460-6984.12118>
- Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H.** (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, 2(2), 103–127. <https://doi.org/10.1080/09658219408258940>
- Giraudeau, B., & Mary, J.** (2001). Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistics in Medicine*, 20(21), 3205–3214. <https://doi.org/10.1002/sim.935>
- Gray, S.** (2003). Diagnostic accuracy and test–retest reliability of nonword repetition and digit span tasks administered to pre-school children with specific language impairment. *Journal of Communication Disorders*, 36(2), 129–151. [https://doi.org/10.1016/S0021-9924\(03\)00003-0](https://doi.org/10.1016/S0021-9924(03)00003-0)
- Hayden, D.** (2004). PROMPT: A tactually grounded treatment approach to speech production disorders. In I. Stockman (Ed.), *Movement and action in learning and development: Clinical implications for pervasive developmental disorders* (pp. 255–297). San Diego, CA: Elsevier-Academic Press.
- Hodson, B. W., & Paden, E. P.** (1991). *A phonological approach to remediation: Targeting intelligible speech*. Austin, TX: Pro-Ed.
- Holm, A., Crosbie, S., & Dodd, B.** (2007). Differentiating normal variability from inconsistency in children's speech: Normative data. *International Journal of Language & Communication Disorders*, 42(4), 467–486. <https://doi.org/10.1080/13682820600988967>
- Howard, S., & Heselwood, B.** (2012). The contribution of phonetics to the study of vowel development and disorders. In M. Ball, J. Damico, M. Ball, & F. Gibbon (Eds.), *Handbook of vowels and vowel disorders* (pp. 61–112). Abingdon, United Kingdom: Routledge.
- Ingram, D.** (2002). The measurement of whole-word productions. *Journal of Child Language*, 29(4), 713–733. <https://doi.org/10.1017/S0305000902005275>
- Iuzzini-Seigel, J., Hogan, T. P., & Green, J. R.** (2017). Speech inconsistency in children with childhood apraxia of speech, language impairment, and speech delay: Depends on the stimuli. *Journal of Speech, Language, and Hearing Research*, 60(5), 1194–1210. https://doi.org/10.1044/2016_JSLHR-S-15-0184
- Iuzzini-Seigel, J., & Murray, E.** (2017). Speech assessment in children with childhood apraxia of speech. *Perspectives of the ASHA Special Interest Groups*, 2(2), 47–60. <https://doi.org/10.1044/persp2.SIG2.47>
- James, D. G. H., Ferguson, W. A., & Butcher, A.** (2016). Assessing children's speech using picture-naming: The influence of differing phonological variables on some speech outcomes. *International Journal of Speech-Language Pathology*, 18(4), 364–377. <https://doi.org/10.3109/17549507.2015.1101159>
- Kent, R. D., Kent, J. F., & Rosenbek, J. C.** (1987). Maximum performance tests of speech production. *Journal of Speech and Hearing Disorders*, 52(4), 367–387. <https://doi.org/10.1044/jshd.5204.367>
- Kirk, C., & Vigeland, L.** (2014). A psychometric review of norm-referenced tests used to assess phonological error patterns. *Language, Speech, and Hearing Services in Schools*, 45(4), 365–377. https://doi.org/10.1044/2014_lshss-13-0053
- Knol, F., Boelhouwer, J., & Veldheer, V.** (2012). *Statusontwikkeling van wijken in Nederland 1998–2010 [Neighbourhood status development in the Netherlands 1998–2010]*. The Hague, the Netherlands: Institute for Social Research.
- Leijdekker-Brinkman, W.** (2002). *Metaphonbox (Dutch version of the Metaphon Screening Assessment)*. Lisse, the Netherlands: Swets & Zeitlinger B.V.
- Lewis, B. A., Freebairn, L. A., Hansen, A. J., Iyengar, S. K., & Taylor, H. G.** (2004). School-age follow-up of children with childhood apraxia of speech. *Language, Speech, and Hearing Services in Schools*, 35(2), 122–140.
- Lousada, M., Mendes, A. P., Valente, A. R., & Hall, A.** (2012). Standardization of a phonetic–phonological test for European–Portuguese children. *Folia Phoniatrica et Logopaedica*, 64(3), 151–156. <https://doi.org/10.1159/000264712>
- Maassen, B., & Terband, H.** (2015). Process-oriented diagnosis of childhood and adult apraxia of speech (CAS and AOS). In M. A. Redford (Ed.), *The handbook of speech production* (1st ed., pp. 331–350). Hoboken, NJ: Wiley.
- Maassen, B., van Haften, L., Diepeveen, S., Terband, H., van den Engel-Hoek, L., Veenker, T., & De Swart, B.** (2019). *Computer Articulation Instrument*. Amsterdam, the Netherlands: Boom Uitgevers.
- McCauley, R. J., & Strand, E. A.** (2008). A review of standardized tests of nonverbal oral and speech motor performance in children. *American Journal of Speech-Language Pathology*, 17(1), 81–91. [https://doi.org/10.1044/1058-0360\(2008\)007](https://doi.org/10.1044/1058-0360(2008)007)
- McIntosh, B., & Dodd, B. J.** (2008). Two-year-olds' phonological acquisition: Normative data. *International Journal of Speech-Language Pathology*, 10(6), 460–469. <https://doi.org/10.1080/17549500802149683>
- McLeod, S., & Verdon, S.** (2014). A review of 30 speech assessments in 19 languages other than English. *American Journal*

- of *Speech-Language Pathology*, 23(4), 1–16. https://doi.org/10.1044/2014_ajslp-13-0066
- Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., ... de Vet, H. C. W. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology*, 10(22), 1–8. <https://doi.org/10.1186/1471-2288-10-22>
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
- Murray, E., McCabe, P., Heard, R., & Ballard, K. J. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 58, 43–60. https://doi.org/10.1044/2014_jslhr-s-12-0358
- Nijland, L. (2003). *Developmental apraxia of speech: Deficits in phonetic planning and motor programming*. Wageningen, the Netherlands: Ponson & Looijen.
- Oller, D., & Delgado, R. (2000). Logical International Phonetics Program (Version 2.02) [Computer software]. Miami, FL: Intelligent Hearing Systems.
- Pollock, K. E., & Berni, M. C. (2001). Transcription of vowels. *Topics in Language Disorders*, 21(4), 22–40.
- Prathane, B., Thanaviratnanich, S., & Pongjanyakul, A. (2003). Oral diadochokinetic rates for normal Thai children. *International Journal of Language & Communication Disorders*, 38(4), 417–428. <https://doi.org/10.1080/1368282031000154042>
- Preston, J. L., & Edwards, M. L. (2009). Speed and accuracy of rapid speech output by adolescents with residual speech sound errors including rhotics. *Clinical Linguistics & Phonetics*, 23(4), 301–318. <https://doi.org/10.1080/02699200802680833>
- Preston, J. L., & Koenig, L. L. (2011). Phonetic variability in residual speech sound disorders, exploration of subtypes. *Topics in Language Disorders*, 31(2), 168–184. <https://doi.org/10.1097/TLD.0b013e318217b875>
- Priester, G. H., Post, W. J., & Goorhuis-Brouwer, S. M. (2011). Phonetic and phonemic acquisition: Normative data in English and Dutch speech sound development. *International Journal of Pediatric Otorhinolaryngology*, 75(4), 592–596. <https://doi.org/10.1016/j.ijporl.2011.01.027>
- Rvachew, S., Hodge, M., & Ohberg, A. (2005). Obtaining and interpreting maximum performance tasks from children: A tutorial. *Journal of Speech-Language Pathology and Audiology*, 29(4), 146–157.
- Scheltinga, F. (1998). *Nonword repetition test*. Universiteit van Amsterdam & CED Groep Rotterdam, Rotterdam, Netherlands.
- Shriberg, L. D., Austin, D., Lewis, B. A., McSweeney, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC): Metric extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, 40(4), 708–722. <https://doi.org/10.1044/jslhr.4004.708>
- Shriberg, L. D., Fourakis, M., Hall, S. D., Karlsson, H. B., Lohmeier, H. L., McSweeney, J. L., ... Wilson, D. L. (2010). Extensions to the speech disorders classification system (SDCS). *Clinical Linguistics & Phonetics*, 24(10), 795–824. <https://doi.org/10.3109/02699206.2010.503006>
- Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*, 5(3), 225–279. <https://doi.org/10.3109/02699209108986113>
- Shriberg, L. D., Lohmeier, H. L., Campbell, T. F., Dollaghan, C. A., Green, J. R., & Moore, C. A. (2009). A nonword repetition task for speakers with misarticulations: The syllable repetition task (SRT). *Journal of Speech, Language, and Hearing Research*, 52(5), 1189–1212. [https://doi.org/10.1044/1092-4388\(2009/08-0047\)](https://doi.org/10.1044/1092-4388(2009/08-0047))
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1991). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Research*, 55(4), 779–798. <https://doi.org/10.1044/jslhr.3402.446>
- Smith, B. (2006). Precautions regarding nonword repetition tasks. *Applied Psycholinguistics*, 27(4), 584–587. <https://doi.org/10.1017/S0142716406310399>
- Sosa, A. V. (2015). Intra-word variability in typical speech development. *American Journal of Speech-Language Pathology*, 24(1), 24–35. https://doi.org/10.1044/2014_AJSLP-13-0148
- Stoel-Gammon, C., & Cooper, J. A. (1984). Patterns of early lexical and phonological development. *Journal of Child Language*, 11(2), 247–271. <https://doi.org/10.1017/S0305000900005766>
- Tellegen, P., & Laros, J. (1993). The construction and validation of a nonverbal test of intelligence: The revision of the Snijders-Oomen tests. *European Journal of Psychological Assessment*, 9(2), 147–157.
- Thoonen, G. (1998). *Developmental apraxia of speech in children: Quantitative assessment of speech characteristics*. Enschede, the Netherlands: Print Partners Ipskamp.
- Thoonen, G., Maassen, B., Gabreels, F., & Schreuder, R. (1999). Validity of maximum performance tasks to diagnose motor speech disorders in children. *Clinical Linguistics & Phonetics*, 13(1), 1–23. <https://doi.org/10.1080/026992099299211>
- Thoonen, G., Maassen, B., Wit, J., Gabreels, F., & Schreuder, R. (1996). The integrated use of maximum performance tasks in differential diagnostic evaluations among children with motor speech disorders. *Clinical Linguistics & Phonetics*, 10(4), 311–336. <https://doi.org/10.3109/02699209608985178>
- Tresoldi, M., Ambrogi, F., Favero, E., Colombo, A., Barillari, M. R., Velardi, P., & Schindler, A. (2015). Reliability, validity and normative data of a quick repetition test for Italian children. *International Journal of Pediatric Otorhinolaryngology*, 79(6), 888–894. <https://doi.org/10.1016/j.ijporl.2015.03.025>
- Van de Wijer-Muris, I., & Draaisma, Y. (2000). *Cursusreader 'Fonologische stoornissen,' version October 2000 (Dutch version of the Hodson Assessment of Phonological Patterns)*. Unpublished document.
- Vance, M., Stackhouse, J., & Wells, B. (2005). Speech-production skills in children aged 3–7 years. *International Journal of Language & Communication Disorders*, 40(1), 29–48. <https://doi.org/10.1080/13682820410001716172>
- Verhoeven, L., & Vermeer, A. (2001). *Taaltoets alle kinderen (Dutch language proficiency test for all children)*. Arnhem, the Netherlands: Citogroep.
- Verhoeven, L., & Vermeer, A. (2006). *Verantwoording taaltoets alle kinderen (Dutch language proficiency test for all children)*. Arnhem, the Netherlands: Citogroep.
- Wolk, L., & Meisler, A. W. (1998). Phonological assessment: A systematic comparison of conversation and picture naming. *Journal of Communication Disorders*, 31(4), 291–313. [https://doi.org/10.1016/S0021-9924\(97\)00092-0](https://doi.org/10.1016/S0021-9924(97)00092-0)
- Yaruss, S. J., & Logan, K. J. (2002). Evaluating rate, accuracy, and fluency of young children's diadochokinetic productions: A preliminary investigation. *Journal of Fluency Disorders*, 27(1), 65–86. [https://doi.org/10.1016/S0094-730X\(02\)00112-2](https://doi.org/10.1016/S0094-730X(02)00112-2)

Appendix A

Frequency Distributions of the Phonological Features of Picture Naming and Nonword Imitation

Class	Feature	Number of syllable-initial features		
		Picture naming	Nonword imitation	
			2–3 years	4–7 years
Consonants	p	4	9	7
	b	2	5	3
	t	9	5	3
	d	3	7	5
	k	3	6	4
	g	0	0	0
	ŋ	0	0	0
	m	3	9	7
	n	2	8	6
	l	6	9	7
	R	4	6	4
	f	6	14	10
	v	3	6	4
	s	5	12	8
	z	3	3	1
	ʃ	2	1	1
	ʒ	1	0	0
	j	3	3	1
	x	3	7	5
	h	2	4	2
Vowels	u	4	4	2
	i	8	22	3
	y	2	4	1
	e	4	8	0
	ø	4	0	0
	a	7	14	5
	o	5	15	5
	u	4	7	0
	ɪ	9	16	16
	ɛ	3	20	16
	ɑ	11	16	11
	ʊ	2	11	11
	ə	12	1	1
	ɔ	9	11	11
Diphthongs	ɛi	5	0	0
	au	2	1	1
	ɔy	3	1	1
Syllable structures	V	3	16	1
	CV	17	42	7
	VC	0	15	0
	CVC	40	66	66
	CCV	3	1	1
	CVCC	6	2	2
	CCVC	15	3	3
	CCVCC	3	2	2
	CCVCV	3	0	0
	Initial consonant clusters	23	6	6

Appendix B (p. 1 of 4)

Data of the Reliability and Validity Studies

Table B1. Intraclass correlation coefficients for the interrater reliability per task.

Task	Parameter	<i>n</i>	ICC	CI lb	CI ub
PN	PCCI	99	.80	.74	.88
	PVC	99	.59	.45	.72
	Level 5	99	.84	.80	.91
	RedClus	99	.90	.87	.95
	CCVC	99	.85	.80	.92
NWI	PCCI	98	.77	.70	.87
	PVC	98	.62	.50	.75
	Level 4	98	.75	.68	.85
	Level 5	98	.76	.69	.86
	RedClus	96	.82	.76	.90
	CVC	96	.76	.68	.85
	CCVC	94	.71	.62	.82
WR	PWV word	72	.73	.65	.85
NWR	PWV nonword	67	.25	.05	.47
MRR	MRR-pa	100	.81	.75	.89
	MRR-ta	95	.77	.70	.87
	MRR-ka	103	.83	.77	.90
	MRR-pataka	67	.41	.23	.62
	MRR-pata	90	.62	.51	.77
	MRR-taka	91	.52	.38	.67

Note. ICC = intraclass correlation coefficient; CI lb= confidence interval, lower bound; CI ub= confidence interval, upper bound; PN = picture naming; PCCI = percentage of consonants correct in syllable-initial position; PVC = percentage of vowels correct; Level 5 = percentage of correct consonants /l/ and /r/; RedClus = percentage of reduction of initial consonant clusters from two consonants to one; CCVC = percentage of correct syllable structure CCVC; NWI = nonword imitation; Level 4 = percentage of correct consonants /b/, /f/, and /u/; CVC = percentage of correct syllable structure CVC; WR = word repetition; PWV word = proportion of whole-word variability: word repetition; NWR = nonword repetition; PWV nonword = proportion of whole-word variability: nonword repetition; MRR = maximum repetition rate; MRR-pa = number of syllables per second of sequence /pa/; MRR-ta = number of syllables per second of sequence /ta/; MRR-ka = number of syllables per second of sequence /ka/; MRR-pataka = number of syllables per second of sequence /pataka/; MRR-pata = number of syllables per second of sequence /pata/; MRR-taka = number of syllables per second of sequence /taka/.

Table B2. Point-to-point agreement for interrater reliability for phonetic transcription.

Task	<i>n</i>	% Agreement		
		Segments	Consonants	Vowels
PN	100	95.7	95.5	95.8
NWI	97	94.7	95.1	94.4

Appendix B (p. 2 of 4)

Data of the Reliability and Validity Studies

Table B3. Intraclass correlation coefficients and descriptives (means and standard deviations) for test–retest reliability per task.

Task	Parameter	n	ICC	CI lb	CI ub	T1		T2		Sig. ^a
						M	SD	M	SD	
PN	PCCI	106	.51	.39	.67	97.5	4.25	98.6	3.38	.002
	PVC	106	.31	.15	.52	97.9	3.47	98.9	2.44	.000
	Level 5	106	.71	.63	.81	95.6	13.2	97.6	9.94	.043
	RedClus ^b	105	.75	.68	.86	−3.42	8.49	−2.93	10.4	.10
	CCVC	105	.76	.69	.85	95.1	11.5	96.3	12.4	.055
NWI	PCCI	107	.74	.63	.88	89.1	9.70	92.6	7.76	.000
	PVC	107	.77	.69	.89	94.2	6.59	96.5	5.33	.000
	Level 4	107	.52	.37	.72	90.4	11.8	93.4	9.00	.007
	Level 5	107	.73	.65	.84	87.7	16.0	91.3	13.4	.001
	RedClus ^b	105	.41	.26	.59	−5.71	9.76	−6.57	13.8	.51
	CVC	106	.88	.84	.94	93.9	10.4	96.1	8.59	.000
	CCVC	105	.60	.50	.76	87.0	21.4	84.9	25.9	.71
	WR	86	.66	.56	.82	0.26	0.051	0.25	0.059	.65
NWR	PWV nonword	75	.39	.21	.64	0.32	0.093	0.32	0.076	.69
MRR	MRR-pa	65	.70	.61	.85	4.63	0.73	4.69	0.64	.66
	MRR-ta	63	.53	.35	.75	4.50	0.61	4.52	0.61	.75
	MRR-ka	63	.60	.47	.79	4.25	0.55	4.34	0.54	.29
	MRR-pataka	54	.18	−.070	.36	3.91	0.86	4.35	0.83	.004
	MRR-pata	59	.46	.28	.68	4.48	0.86	4.71	0.93	.072
	MRR-taka	54	.43	.25	.67	4.63	0.88	4.57	0.70	.81

Note. ICC = intraclass correlation coefficient; CI lb= confidence interval lower bound; CI ub= confidence interval upper bound; PN = picture naming; PCCI = percentage of consonants correct in syllable-initial position; PVC = percentage of vowels correct; Level 5 = percentage of correct consonants /l/ and /r/; RedClus = percentage of reduction of initial consonant clusters from two consonants to one; CCVC = percentage of correct syllable structure CCVC; NWI = nonword imitation; Level 4 = percentage of correct consonants /b/, /f/, and /u/; CVC = percentage of correct syllable structure CVC; WR = word repetition; PWV word = proportion of whole-word variability: word repetition; NWR = nonword repetition; PWV nonword = proportion of whole-word variability: nonword repetition; MRR = maximum repetition rate; MRR-pa = number of syllables per second of sequence /pa/; MRR-ta = number of syllables per second of sequence /ta/; MRR-ka = number of syllables per second of sequence /ka/; MRR-pataka = number of syllables per second of sequence /pataka/; MRR-pata = number of syllables per second of sequence /pata/; MRR-taka = number of syllables per second of sequence /taka/.

^aStatistical significance of the T1 and T2 differences were calculated using the Wilcoxon signed-ranks test. The significance level was set at .05. ^bThe percentage cluster reductions (RedClus) was inverted (multiplied by −1) such that higher values reflected a more accurate performance.

Appendix B (p. 3 of 4)

Data of the Reliability and Validity Studies

Table B4. Descriptives (means and standard deviations) of the extracted parameters for the picture-naming and nonword-imitation tasks for the younger age groups (2;0–3;11).

Age range (years; months)	2;0–2;3		2;4–2;7		2;8–2;11		3;0–3;3		3;4–3;7		3;8–3;11	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
PN												
PCCI	76.0	14.1	81.4	14.1	87.7	10.0	90.9	7.8	92.0	7.2	93.4	6.3
PVC	87.6	9.2	89.8	7.6	93.3	4.9	94.9	4.1	95.2	3.8	96.6	3.1
RedClus ^a	–49.6	26.1	–34.6	26.3	–20.0	21.5	–13.2	17.1	–11.1	16.2	–7.1	14.5
Level 5	49.7	27.2	59.1	27.2	69.6	24.7	76.7	21.6	81.6	18.8	86.1	18.1
CCVC	39.7	28.0	56.9	28.4	75.4	24.9	82.0	19.3	84.4	18.5	88.6	17.2
NWI												
PCCI	67.2	12.7	72.1	11.9	77.3	10.3	79.7	10.3	83.7	9.1	84.6	10.1
PVC	81.8	9.2	86.5	6.7	89.3	6.4	91.9	5.6	93.5	4.8	94.1	4.9
RedClus ^a	–54.5	29.1	–42.2	28.1	–35.1	26.5	–26.3	21.4	–23.1	23.1	–13.5	16.6
Level 4	26.0	36.0	45.6	40.1	55.6	35.6	60.9	33.9	64.0	32.9	74.9	28.4
Level 5	63.1	17.6	68.4	17.0	75.4	14.9	76.3	15.8	81.5	12.7	82.5	13.8
CVC	93.0	7.3	93.5	5.2	96.0	4.4	96.0	4.0	97.7	3.1	97.4	4.1
CCVC	66.7	47.7	72.2	45.1	77.8	41.9	85.0	35.9	85.9	35.0	94.1	23.7

Note. PN = picture naming; PCCI = percentage of consonants correct in syllable-initial position; PVC = percentage of vowels correct; RedClus = percentage of reduction of initial consonant clusters from two consonants to one; Level 5 = percentage of correct consonants /l/ and /r/; CCVC = percentage of correct syllable structure CCVC; NWI = nonword imitation; Level 4 = percentage of correct consonants /b/, /f/, and /u/; CVC = percentage of correct syllable structure CVC.

^aThe percentage of cluster reductions (RedClus) was inverted (multiplied by –1) such that higher values reflect a more accurate performance.

Table B5. Descriptives (means and standard deviations) of the extracted parameters for the picture-naming and nonword-imitation tasks for the older age groups (4;0–6;11).

Age range (years; months)	4;0–4;3		4;4–4;7		4;8–4;11		5;0–5;3		5;4–5;7		5;8–5;11		6;0–6;5		6;6–6;11	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
PN																
PCCI	95.2	5.2	96.7	3.3	96.9	3.4	96.8	3.7	97.0	5.8	97.9	3.3	98.2	2.5	98.4	2.2
PVC	97.0	3.9	97.9	2.8	98.0	2.2	97.7	3.1	97.6	5.5	98.5	2.4	98.4	2.3	98.6	1.8
RedClus ^a	–4.8	7.1	–3.2	5.9	–2.8	6.8	–3.0	6.2	–2.3	6.2	–1.7	5.1	–1.7	3.3	–0.8	2.0
Level 5	91.9	12.4	91.9	13.6	95.3	8.9	93.6	11.0	95.5	9.2	96.5	10.7	97.3	5.7	98.1	4.9
CCVC	92.6	10.6	95.3	8.0	95.2	8.1	94.4	9.9	94.9	9.4	96.8	7.0	96.9	5.5	97.5	4.5
NWI																
PCCI	82.5	10.1	87.2	9.8	86.9	9.6	87.7	6.9	89.0	7.2	89.6	7.4	90.9	6.2	91.8	7.0
PVC	90.5	7.5	92.4	6.5	93.0	6.7	93.5	4.8	94.4	4.5	94.3	4.7	96.1	3.8	94.8	5.7
RedClus ^a	–12.8	13.7	–8.0	12.2	–9.3	12.9	–7.7	11.8	–6.6	11.9	–6.8	11.8	–6.4	10.5	–4.2	8.5
Level 4	79.2	15.0	87.0	12.0	85.8	14.3	87.8	11.8	90.0	9.9	88.9	10.3	92.3	8.6	92.8	9.6
Level 5	80.3	16.2	84.3	16.3	84.1	15.7	87.2	12.8	86.3	12.9	87.9	14.6	89.4	11.7	89.7	11.2
CVC	91.1	8.4	92.6	7.4	92.7	7.6	93.3	5.7	94.4	5.5	95.0	5.0	96.6	4.0	95.2	5.8
CCVC	74.9	25.7	82.1	22.3	79.8	25.7	83.0	25.2	85.6	21.7	84.1	24.4	86.1	21.6	90.8	16.2

Note. PN = picture naming; PCCI = percentage of consonants correct in syllable-initial position; PVC = percentage of vowels correct; RedClus = percentage of reduction of initial consonant clusters from two consonants to one; Level 5 = percentage of correct consonants /l/ and /r/; CCVC = percentage of correct syllable structure CCVC; NWI = nonword imitation; Level 4 = percentage of correct consonants /b/, /f/, and /u/; CVC = percentage of correct syllable structure CVC.

^aThe percentage of cluster reductions (RedClus) was inverted (multiplied by –1) such that higher values reflect a more accurate performance.

Appendix B (p. 4 of 4)

Data of the Reliability and Validity Studies

Table B6. Descriptives (means and standard deviations) of the extracted parameters for the word and nonword repetition tasks for the younger age groups (2;0–3;11).

PWV		2;0–2;3	2;4–2;7	2;8–2;11	3;0–3;3	3;4–3;7	3;8–3;11
WR	<i>M</i>	0.30	0.31	0.30	0.30	0.28	0.28
	<i>SD</i>	0.11	0.09	0.07	0.09	0.09	0.08
NWR	<i>M</i>	0.36	0.35	0.33	0.34	0.33	0.30
	<i>SD</i>	0.11	0.10	0.12	0.11	0.11	0.07

Note. PWV = proportion of whole-word variability; WR = word repetition; NWR = nonword repetition.

Table B7. Descriptives (means and standard deviations) of the extracted parameters for the word and nonword repetition tasks for the older age groups (4;0–6;11).

PWV		4;0–4;3	4;4–4;7	4;8–4;11	5;0–5;3	5;4–5;7	5;8–5;11	6;0–6;5	6;6–6;11
WR	<i>M</i>	0.24	0.24	0.24	0.23	0.23	0.23	0.23	0.23
	<i>SD</i>	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.04
NWR	<i>M</i>	0.27	0.28	0.27	0.28	0.27	0.26	0.26	0.25
	<i>SD</i>	0.07	0.08	0.07	0.08	0.07	0.07	0.07	0.06

Note. PWV = proportion of whole-word variability; WR = word repetition; NWR = nonword repetition.

Table B8. Maximum repetition rates (MRRs) per sequence and age group (number of syllables per second).

Sequence		3;0–3;3	3;4–3;7	3;8–3;11	4;0–4;3	4;4–4;7	4;8–4;11	5;0–5;3	5;4–5;7	5;8–5;11	6;0–6;5	6;6–6;11
MRR-pa	<i>M</i>	4.00	4.02	4.08	4.29	4.54	4.57	4.64	4.80	4.82	4.92	5.01
	<i>SD</i>	0.54	0.49	0.54	0.69	0.50	0.64	0.61	0.57	0.61	0.51	0.58
MRR-ta	<i>M</i>	3.88	3.99	4.06	4.15	4.41	4.44	4.44	4.69	4.69	4.85	4.93
	<i>SD</i>	0.57	0.58	0.62	0.65	0.55	0.60	0.60	0.54	0.61	0.67	0.61
MRR-ka	<i>M</i>	3.61	3.75	3.85	3.98	4.15	4.21	4.34	4.35	4.42	4.50	4.60
	<i>SD</i>	0.51	0.56	0.53	0.56	0.57	0.58	0.51	0.45	0.49	0.50	0.55
MRR-pataka	<i>M</i>	3.48	3.70	3.64	3.74	3.93	3.96	4.09	4.14	4.37	4.39	4.55
	<i>SD</i>	0.71	0.98	0.89	0.76	0.77	0.97	0.82	0.88	0.89	0.98	0.93
MRR-pata	<i>M</i>	4.11	3.93	4.04	4.37	4.42	4.47	4.49	4.71	4.58	4.77	4.80
	<i>SD</i>	0.72	0.61	0.75	0.91	0.78	1.02	0.73	0.75	0.83	0.90	0.82
MRR-taka	<i>M</i>	3.82	4.04	4.09	4.28	4.37	4.47	4.37	4.54	4.74	4.63	4.96
	<i>SD</i>	0.72	0.85	0.82	0.75	0.80	0.78	0.75	0.61	0.77	0.74	0.78

Note. MRR-pa = number of syllables per second of sequence /pa/; MRR-ta = number of syllables per second of sequence /ta/; MRR-ka = number of syllables per second of sequence /ka/; MRR-pataka = number of syllables per second of sequence /pataka/; MRR-pata = number of syllables per second of sequence /pata/; MRR-taka = number of syllables per second of sequence /taka/.

Figure 1. Continuous norms per age group for picture naming percentage of consonants correct in syllable-initial position, age 2;0–3;11.

Figure 2. Continuous norms per age group for picture naming percentage of consonants correct in syllable-initial position, age 4;0–6;11.

Figure 3. Continuous norms per age group for picture naming percentage of vowels correct, age 2;0–3;11.

Figure 4. Continuous norms per age group for picture naming percentage of vowels correct, age 4;0–6;11.

Figure 5. Continuous norms per age group for nonword imitation percentage of consonants correct in syllable-initial position, age 2;0–3;11.

Figure 6. Continuous norms per age group for nonword imitation percentage of consonants correct in syllable-initial position, age 4;0–6;11.

Figure 7. Continuous norms per age group for nonword imitation percentage of vowels correct, age 2;0–3;11.

Figure 8. Continuous norms per age group for nonword imitation percentage of vowels correct, age 4;0–6;11.

Figure 9. Continuous norms per age group for word repetition; proportion whole-word variability, age 2;0–3;11.

Figure 10. Continuous norms per age group for word repetition; proportion whole-word variability, age 4;0–6;11.

Figure 11. Continuous norms per age group for nonword repetition; proportion whole-word variability, age 2;0–3;11.

Figure 12. Continuous norms per age group for nonword repetition; proportion whole-word variability, age 4;0–6;11.

Figure 13. Continuous norms per age group for number of syllables per second of sequence /pa/, age 3;0–6;11.

Figure 14. Continuous norms per age group for number of syllables per second of sequence /taka/, age 3;0–6;11.

Figure 15. Continuous norms per age group for number of syllables per second of sequence /pataka/, age 3;0–6;11.

Figure 16. Continuous norms per age group for number of syllables per second of sequence /taka/, age 3;0–6;11.

Copyright of Journal of Speech, Language & Hearing Research is the property of American Speech-Language-Hearing Association and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.