Inter-rater reliability in classification of canonical babbling status based on canonical babbling ratio in infants with isolated cleft palate randomised to Timing of Primary Surgery for Cleft Palate (TOPS)

DOI: 10.1080/02699206.2021.2012259

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Clinical Linguistics and Phonetics

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright, please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate the claim.
Trial registration number here: www.clinicaltrials.gov, identifier NCT00993551

Background

Infants typically start to produce well-formed speech-like syllables, known as canonical syllables, before 11 months of age (Koopmans-van Beinum & van der Stelt, 1986; Oller, 1980). Delayed onset of canonical babbling (CB) identifies a risk of delayed/disordered speech and language development (Lohmander et al., 2017; Oller et al., 1998).

An operational definition of canonical babbling onset as a ratio was first proposed by Oller and Eilers (1988). In research, perhaps the most commonly used criterion for presence of CB is a canonical babbling ratio (CBR) of at least 0.15 canonical syllables of all speech-like syllables produced in a speech sample. However, a number of variations have been used (See Lang et al., 2019; Molemans et al., 2012, for an overview).

The criterion of at least 0.15 canonical syllables has been empirically confirmed to match parents’ opinion that the child is babbling canonically (Lewedag, 1995). This criterion was supported by laboratory assessment in which assistants reviewed selected recordings of six infants of whom four were canonical and two not canonical. The lowest CBR of the canonical recordings observed was 0.15. Additionally, this criterion showed complete inter-rater agreement between seven raters in classification of CB status in 41/42 cases with disagreement for one recording with a CBR of 0.15 (Lynch et al., 1995). The 0.15 criterion has been found to distinguish effectively between typically developing infants and infants in clinical subgroups at risk of speech and language delay or disorder (Chapman et al., 2001; Chapman et al., 2003; Masataka et al., 2001; Overby et al., 2019a; Overby et al., 2019b; Patten et al., 2014).

Canonical babbling ratios have been derived using different approaches. Traditionally, many studies extracted the CBR from phonetic transcription of the speech samples (e.g., Chapman et al., 2001; Willadsen & Albrechtsen, 2006). More recently, studies...
have assessed speech-like syllables in a recording in real time to calculate the CBR (Belardi et al., 2017; Overby et al., 2019a; Patten et al., 2014; Willadsen et al., 2019). Variations in the method of assessment of CBR may influence the resulting CB status. Molemans et al. (2012) showed that if classification of CB status is based on a single speech sample, variation in sample size (number of utterances) may influence classification of CB status substantially. Based on longitudinal data from 40 typically developing infants from 6 to 9 months, Molemans et al. (2012), showed that if only one speech recording was included, a minimum of 425 syllables was needed to determine canonical babbling onset by means of CBR ≥0.15 with 95% confidence. The larger the utterance sample size, the more infants classified as having acquired CB, although the authors were unable to show a definite association between utterance sample size and age of CB onset. At 12 months of age, when all participants were canonical, a speech sample of 50 syllables was enough to classify CB status reliably (Molemans et al., 2012). The recording setting (home versus laboratory) has been shown to influence the number of utterances produced by infants with the home situation providing more than twice as many utterances as the laboratory setting, calculated as average utterances per minute in recordings including approximately 70 utterances (Lwedag et al., 1994). Lee et al. (2017) provided preliminary evidence into a complex three-way interaction between age, language/culture, and social circumstance on CBRs. They concluded that “the development of the capability for canonical babbling (…) seems to reflect a highly adaptive process involving infant responsivity to the social and linguistic environment, which itself appears to adapt to the infant across time” (p.150). Thus, this is a field of future study to increase our understanding of how these factors interact during the development of canonical syllables.
Variation and Reliability between Raters of CBR

Perceptual assessments are prone to human variation and error, as well as possible variation from coding procedures and subjects under assessment (Cordes, 1994). Thus, it is considered good practice to include more than one rater and to report inter-rater reliability (Sell, 2005), to ensure that results are as consistent across different raters as possible (Persson et al., 2020). As CBR is calculated from all canonical syllables divided by all speech-like syllables, variation in either measure will lead to a change in CBR. Therefore, assessment variation may change classification of CB status. Table 1 provides an overview of published studies reporting perceptual assessment of CBR as an outcome to determine CB status, and how inter-rater agreement was reported. Many of these provide limited information on inter-rater reliability. The majority included one principal rater with an additional rater assessing only a subset of the speech material. Most studies were based on phonetic transcription but only few included a measure of inter-rater variability of the key components: canonical syllables and the total number of syllables. Only one study (Lynch et al., 1995) based on phonetic transcription reported inter-rater agreement for classification of CB status. In all studies using real time assessment, with the exception of Nyman & Lohmander (2018), analyses were done from home videos collected retrospectively. This was because the infants had conditions that were diagnosed at an older age. Therefore, recordings varied greatly regarding factors including: recording length (approximately 1 to 10 minutes), situations filmed (play, meal time, bath), visibility of the infant, and age range (approximately 0-24 months of age).

Insert Table 1 here
Table 1 shows that studies reporting inter-rater-reliability most commonly used intraclass correlation coefficients (ICCs). The ICC is often used in reliability studies where typically a random sample of entities are rated independently by two or more raters. An ICC takes the range of possible values from 0 to 1, with higher scores indicating greater reliability. Intraclass correlation coefficients may either show if scores agree in absolute value (absolute agreement) or agree in rank ordering (average agreement; Hallgren, 2012). Unfortunately, none of the studies specified the ICC measure used and only one study provided confidence intervals, both of which complicate the interpretation of results with most studies reassessing relatively few recordings. For studies assessing inter-rater reliability for canonical syllables (n=7), it was reported as very good to excellent (ICC 0.9 to 0.98). Only one study reported inter-rater agreement of the total number of syllables, which was slightly lower (ICC 0.87) with a wide confidence interval (0.61-0.95). Likewise, point-by-point agreements for identification of canonical syllables reported in two studies showed an overall mean of at least 91%.

Only one of the studies used real time analysis to assess inter-rater reliability of CBR and reported ICC in the excellent range (0.89) however, a wide 95% confidence interval was reported (0.69-0.96;Patten et al., 2014). Importantly, few studies reported the impact of inter-rater variation on CB status ranging between 90% and 100% agreement. The sample sizes on which these studies were based were typically small. They included inter-rater agreement of 6-20 speech samples of varying, sometimes unspecified, length, including a maximum of 10 infants assessed by 2 or 3 raters (Belardi et al., 2017; Overby et al., 2019b; Patten et al., 2014) except for Lynch et al. (1995) who included 7 raters.
Thus, there is limited knowledge about inter-rater reliability in assessment of CB status based on CBR assessed in real time. Furthermore, information regarding the inter-rater differences in CBR, and the range of variation expected in CBR assessments, is limited.

This study addresses this gap in knowledge. As a subproject in an international randomised controlled trial, Timing of Primary Surgery for Cleft Palate (TOPS; Shaw et al., 2019), we used blinded data that were collected from the participants at 12-month of age. The purpose of the study was to explore inter-rater reliability of the:

1) Classification of CB status based on CBR (≥0.15);
2) CBR;
3) The total number of syllables per infant;
4) Association between the total number of syllables and the CBR and CB status

We hypothesize an inter-rater agreement of at least 90% for classification of CB status by means of CBR (≥0.15), based on previous reports from small scale studies (Belardi et al., 2017; Lynch et al. 1995; Overby et al., 2019b; Patten et al., 2014). No further hypotheses could be generated due to lack of, or very limited information reported on the remaining measures.

--

**Method**

**Speech and Language Therapist Participants**

Twenty-nine experienced speech and language therapists (SLTs), with expertise in cleft palate speech, rated the 12-month assessment video recordings of infants in the TOPS trial. All SLTs were native speakers of one of the five languages included in the TOPS trial:
Brazilian Portuguese, Danish, English, Norwegian, and Swedish. Each SLT had participated in a 3-day training and calibration workshop (Shaw et al., 2019) on assessment of prelinguistic vocalizations, including recognition of canonical syllables. At the end of the workshop, all 29 SLTs were assessed using four video recordings of 12-month-old infants with cleft palate who were not TOPS participants. We examined the impact on CBRs when using four raters compared to three (Persson et al, 2020). The impact of the fourth rater was minimal, the maximum absolute difference of the mean CBR across the four recordings varied between 6% and 13%. Furthermore, the assessment of prelinguistic vocalizations revealed no influence of the SLTs' first language on assessment of the infants' syllable/consonant inventory. Thus, three raters were randomly chosen from the pool of 29 SLTs to rate each recording (Persson et al., 2020).

**Cohort**

All participants had a diagnosis of non-syndromic isolated cleft palate and were participants of the TOPS trial (Shaw et al., 2019). They were randomised to receive standardised primary surgery, using the Sommerlad technique (Sommerlad, 2003) for closure of the cleft at either six or twelve months of gestational corrected age. Full term was defined as 40 weeks.

Infants randomised to surgery at 12 months undertook their 12-month outcome assessment prior to surgical repair. This means that one half of the participants had a repaired cleft palate while the other half had an unrepaired cleft palate. However, this does not affect results of the present study as the purpose was to assess inter-rater agreement in a large sample of infants with expectedly large variation in prelinguistic vocalisations. Data were collected at 12 months of age, irrespective of age of surgery, which was blinded to the
research team. Information on participants’ age at recording and gender are provided in Table 2.

Insert Table 2 here

Canonical babbling is included as a secondary endpoint in the TOPS trial, whose primary analysis is adhering to the intention to treat principle. Intention to treat means the participant data is part of the surgical group they were initially allocated, irrespective of unexpected changes according to the protocol. This is to avoid excluding cases who may have different characteristics from the others.

Of the 558 participants in the TOPS trial, the 12-month speech follow-up was conducted and a recording taken for 501 (89.8%) infants. Reasons for not having a recording included withdrawal from the trial, non-attendance, or no recording taken (n=57). Of the 501 video recordings available 17 (3.4%) were not eligible as the infant produced fewer than 50 utterances. This cut-off was chosen to allow the infant “enough” vocalizations to assess CB status, a number in agreement with previous studies (e.g., Lang et al., 2020; Moeller et al., 2007; Rvachew et al., 1999; ). In total, 484 (58% female) infants were assessed at a median age of 12.3 months (SD 1.2; range 10.4 to 20.7).

Speech Samples

At 12 months of age, every infant was video recorded at a local cleft site by a trained Site SLT for approximately 45 minutes, during a play session with their parent. The infant wore a custom-made vest hosting a microphone. All sites used identical equipment: video camera (JVC-GY-HM100E Solid state) and microphone (Rode NT4). The SLT instructed the parent to play with their infant as they would commonly do and select activities that made the infant more talkative.
All video recordings were divided into two parts of approximately equal length. All speech-like vocalizations were assessed, while vegetative vocalizations such as burps, coughs and hiccups were excluded as well as laughter, cry, and whispered vocalizations (e.g., Lynch et al., 1995). Vocalizations produced while mouthing objects (e.g., toys or food) were excluded (Patten et al., 2004), as well as vocalizations with significant adult vocal or noise overlay.

**Procedure**

As mentioned previously, a total of 29 experienced SLTs assessed the recordings independently, with three randomly selected SLTs assessing each recording. This took place in silent rooms, at the same place and time, using the same type of laptop (Elite Book 840) and headphones (AKG K271 Mark II). To reduce rater fatigue, comfort breaks were encouraged between recordings. As far as practically possible, SLTs did not assess infants from their own site.1

The SLTs assessed all recordings in *real time without pausing a recording and without taking notes*. They watched the video recordings on a laptop, wearing good quality headphones and using a custom-made software, TimeStamper (Willadsen et al., 2018). Before an assessment started, a pop-up window asked the rater if she/he was ready to assess the entire recording (22 minutes): pausing was not possible and would delete prior annotations. Timestamper registers annotations of two default keys for non-canonical and canonical syllables respectively. Thus, SLTs stroke a default key each time they heard a

---

1 In 66 (4.5%) of the 1452 assessments (3*484) an infant was assessed by an SLT from their own site.
Syllables were defined as rhythmic units of speech-like vocalizations (e.g., Patten et al., 2014), while non-speech-like sounds (hicups, burps, sneezes, laughter, crying, and sounds of effort) were intentionally not coded. A non-canonical syllable included a vowel with/without a glottal consonant; a canonical syllable included a fast and smooth transition between a vowel and at least one supra-glottal consonant (e.g., Belardi et al., 2017). Overall, transitions that are too fast to be heard as a “transition” are perceived as gestalt syllables (Patten et al., 2014). During training it was emphasized that common canonical syllables such as [ma] and [dada] sound like words in adult speech (Oller, 2000; p. 100).

Following the assessment, TIMESer automatically generated an output file with a count of non-canonical and canonical syllables and the CBR; in other words the percentage of canonical syllables out of all syllables (canonical plus non-canonical) identified by the SLT.

**Speech Outcomes**

*Classification of CB Status*

Each assessment (including both parts of a recording, 45 min in total) was reviewed and the number of canonical syllables was divided by the total number of syllables produced by an infant, as identified by the assessing SLT. To classify an infant as canonical (CBR≥0.15) or not canonical (<0.15) the three individual assessments were reviewed and an infant was classified according to the majority assessment, see Table 3.

*Insert Table 3 here*

**Inter-Rater Variation in CBR and Total Syllable Count**
Although classifications according to the CBR may agree, the underlying CBR assessment may vary. We examined variation in the CBR overall and within each of the four agreement groups that were divided according to complete agreement (3/3 SLTs) or majority agreement (2/3 SLTs) for classification as canonical or not canonical, as seen in Table 3. Similarly, we explored the variation in the total number of syllables per infant in relation to the CBR and the classification of the CB status as based on CBR.

Ideally to assess CB status, the total number of syllables (canonical plus non-canonical) should be consistent across participants to ensure an equal number of opportunities to produce canonical syllables. Given the high number of participants this was not possible in the present study, however, we were able to explore the possible effect of number of syllables identified on CBR and on classification of CB status.

Three sets of results were produced for each infant; one per SLT assessing the recordings. For each infant the SLTs’ results were ordered from smallest to largest for the total number of syllables identified and the CBR. The difference between the smallest and largest observed values was also calculated.
The results across infants are summarised descriptively. The median and inter quartile range are calculated for the smallest recorded value per infant and this is then repeated for higher ordered values. The minimum and maximum are also calculated. This is undertaken for the cohort as a whole and then for each subgroup as defined in Table 3. Variation across the average number of syllables identified is compared for each classification of CB status and the average CBR per infant.

**Results**

**Inter-Rater Agreement in Classification of Canonical Babbling Status**

The three assessing SLTs agreed completely in their classification of CB status in 423 (87.4%) infants, with higher complete agreement for classification as canonical (326/358; 91%) than for classification as not canonical (97/126; 77%), see Table 4, which presents the classifications of the ratios and agreement.

Insert Table 4 here

Table 5 provides details of the SLTs assessments split according to their classification of CB status. This allows variation in individual SLT assessments to be considered against the majority of the three assessing SLTs per recording. Row 1 provides the results for SLT 1 and shows SLT 1 made 17 assessments and classified 13 as canonical and 4 as not canonical. For the canonical classifications, SLT 1 agreed with the majority in 11 cases (disagreed with majority in 2 cases) and for the non-canonical assessments SLT 1 agreed with the majority in all 4 cases. Of the total of the 29 assessing SLTs, 5 had no
disagreements with the majority classification of any of the recordings they assessed.

Furthermore, 7 and 9 SLTs respectively had no disagreements with the majority classification of canonical and not canonical.

*Insert Table 5 here*

**Inter-Rater Variation in Canonical Babbling Ratio and -Total Number of Syllables**

Table 6 provides summary statistics of the three SLT assessments made for each infant’s CBR and total number of syllables.

*Insert Table 6 here*

**CBR Variation for Infants Classified as Canonical by All SLTs**

Of the 484 infants studied, 326 (67.4%) were classified as canonical (CBR≥0.15) by all three TOPS SLTs, see Table 4. Table 7 provides the summary statistics for CBR and the total number of syllables for these infants.

*Insert Table 7 here*

Supplemental information Figure 1 and Figure 2 present the three SLT assessments of the infants with complete SLT agreement of ‘canonical’ with the smallest and largest difference in CBR respectively.

**CBR Variation for Infants Classified as Not Canonical by All SLTs**

Of the 484 infants, 97 (20%) were classified as not canonical (CBR<0.15) by all three of the TOPS SLTs, see Table 4. Table 8 provides the summary statistics for CBR and the total number of syllables for these infants.
Supplemental information Figure 3 and Figure 4 present the three SLT assessments of the infants with complete SLT agreement of ‘not canonical’ with the smallest and largest difference in CBR respectively.

**Insert Table 8 here**

**CBR Variation for Infants Classified as Canonical by Partial SLT Agreement**

Of the 484 infants 32 (6.6%) were classified as canonical (CBR≥0.15) by only two of the TOPS SLTs, with the third SLT classifying as not canonical. Table 9 provides the summary statistics for CBR and the total number of syllables for these infants.

**Insert Table 9 here**

Supplemental information Figure 5 presents the three SLT assessments of the infants CBR and total number of syllables with partial SLT agreement of ‘canonical’.

**CBR Variation for Infants Classified as Not Canonical by Partial SLT Agreement**

Of the 484 infants, 29 were classified as not canonical (CBR<0.15) by only two of the TOPS SLTs, with the third SLT classifying as canonical. Table 10 provides the summary statistics for CBR and the total number of syllables for these infants.

**Insert Table 10 here**

**Total Number of Syllables Per Infant in Relation to the Classification of CB status**

Figure 1 presents the average total number of syllables identified by the three assessing SLTs, per infant, split by classification of CB status. Across all infants, the minimum number of syllables identified is below 50 for both subgroup classifications. The
average total number of syllables identified per infant is generally fewer for the not canonical than the canonical group (Figure 1). The average total number of syllables per infant according to the number of SLTs classifying the child as canonical is summarized in Table 11. The range and interquartile range by the two groups with an SLT in disagreement with the majority are smaller than those where SLTs are in complete agreement in CB classification by CBR.

Insert Figure 1 here

Insert Table 11 here

Figure 2 presents a scatter plot of the SLT average of the total number of syllables identified (x-axis), per infant, against the SLT average CBR for the infant (y-axis). The red horizontal line represents a CBR=0.15 and the red vertical line represents the average total number of syllables equal to 50. Figure 2 shows that 3 and 7 infants classified as canonical and not canonical by CBR (CBR≥0.15) respectively, had an average total number of syllables identified less than 50.

Discussion

Through this study we examined inter-rater reliability between three independent SLTs in: classification of CB status based on CBR (>0.15), CBRs, and total number of syllables per infant in relation to CBR and to the classification of CB status based on CBR (>0.15). Each recording was assessed independently by three SLTs that were randomly drawn
They assessed a total of 484 12-month-old participants in the TOPS trial.

**Insert Figure 2 here**

**Inter-Rater Agreement in Classification of CB Status**

Results showed complete agreement between three assessing SLTs in classification of CB status, based on CBR, in 423 (87.4%) infants. Moreover, disagreement with the majority decision (n=61) was distributed among 24 of the 29 participating SLTs. In total, 12 of the 29 SLTs agreed with the majority classification of canonical in all of their canonical assessments, similarly, 14 of the 29 SLTs agreed with the majority classification of not canonical in all of their not canonical assessments. Accordingly, this large study based on analyses of speech recordings obtained by trained SLTs following a strict sampling procedure, provides new evidence that trained raters (here SLTs) can reliably assess CB status based on CBR. Furthermore, the strengths of this study are:

- The high number of infants assessed
- The large range in CBR (0 to 0.93)
- Inclusion of three independent assessments per infant
- The high number of SLTs (n=29) included
- The random allocation of three SLTs rating each speech recordings (resulting in a high number of rater combinations)

Previously, results of inter-rater agreement for classification of CB status in real time have come from retrospective studies based on home videos. Therefore, the video recordings have very apparent limitations such as being short, of different length (1-10 minutes), and
quality. In addition, inter-rater reliability of these videos were based on a small number of recordings (6-20), from a few infants (6-10), with a wide age range (0-24 months) and assessed by two or three raters (Belardi et al., 2017; Overby et al., 2019a; Patten et al., 2014). These studies reported very high reliability levels (90%-100%), which was not found in this large study with a high level of control of recording and assessment conditions.

Complete inter-rater agreement occurred more frequently for classification of CB status as canonical (326/358; 91%) than for classification of not canonical (97/126; 77%) (Table 4). This finding is presumably explained by a number of factors: canonical syllables are regarded as building blocks for speech (e.g., Stoel-Gammon, 2011) and are therefore easily recognised by mature speakers; canonical syllables show a high level of inter-rater agreement (e.g., Patten et al. 2014; Von Hapsburg et al., 2006) and are easily recognised by parents (e.g., Oller et al., 2001). Inter-rater agreement for total number of syllables (canonical and non-canonical syllables) has been reported as lower than for total number of canonical syllables (e.g., Patten et al., 2014; Willadsen & Albrechtsen, 2006) indicating that it is more difficult to agree on presence/absence of non-canonical syllables. This is not surprising as some of the non-canonical syllables are affected by factors such as slow transitions between segments and imprecise articulation (+ref). Likewise, a number of studies using either phonetic transcription or real time assessment reported better agreement for assessment of canonical syllables than non-canonical syllables (Overby et al., 2019a; 2019b; Patten et al., 2014; Willadsen & Albrechtsen, 2006). These consistent findings support the view that parents and raters intuitively recognise canonical syllables (Oller et al., 2001; Ramsdell et al., 2012). Nevertheless, although results showed high inter-rater agreement in classification of CB status, inter-rater variation in CBRs and total number of syllables identified was observed.
Variation in Assessment of CBR Across Raters

This is the first study to explore inter-rater variation in assessment of CBR in real time in a large sample. Results revealed that 5 of the 29 SLTs always agreed with the majority classification. Inter-rater variation in assessment of CBR has only been reported sporadically in previous studies. Two studies reported a mean difference between two raters of approximately 10% for the CBRs obtained (Oller & Eilers, 1988; Patten et al., 2014). Oller and Eilers (1988) used phonetic transcription and assessed 11 normally hearing infants between 11-14 months of age who were all canonical. Patten et al., (2014) used real time assessment and assessed CBR in 10 infants (2 x 5 minutes from each infant) including infants with and without autism spectrum disorder from two age groups 9-12 months and 15-18 months. Distribution of infants according to group and age was not reported. The different methodologies used, the low number of infants assessed, and the short video recordings in one study, make direct comparisons with the present study difficult. However, both studies seem to underestimate inter-rater variation in assessment of CBR and none of the studies reported the range of differences in CBR observed. The present study provides evidence of inter-rater variation in CBR ranging between perfect agreements to substantial differences (0%-53%). Exploration of this finding revealed the least variation in CBR (0.04) across the 97 infants that were unanimously classified as not canonical (CBR<0.15). This is expected as CBR can only differ between 0-15% for infants classified as not canonical as opposed to a possible variation between 15-100% for infants unanimously classified as canonical. However, while the maximum average variation (between the 3 assessing SLTs) across the 30 infants with least variation was 0.02, the maximum variation across the 30 infants with the most marked variation was 0.15. This finding probably reflects the pronounced variability in
normal vocal development and highlights the fact that some infants were more difficult to assess reliably than others.

Across the four classification groups (Table 3), the largest average variation was observed for the two groups with majority decision, in other words where only two of the three assessing SLTs agreed in classification of CB status (n=61). These variations may be related to well-formedness of the syllables produced by the infants, since canonical syllables have been found to be assessed with greater confidence and to show better inter-rater agreement in phonetic transcription than non-canonical syllables (Ramsdell et al., 2007). Thus, more variation in CBRs might be related to more syllables perceived as less well-formed. Additionally, Willadsen and Albrechtsen (2006) suggested that the nasalized vowel quality in (some) infants with cleft palate made it difficult to distinguish an isolated vowel production from a complaint and might explain a lower inter-rater agreement for syllable identification in infants born with cleft palate than for infants without cleft palate. This might also have influenced results in the present study.

Although the SLTs were trained to assess pre-speech vocalizations and to use the Timestamper software, it cannot be ruled out that mistakes may have occurred in some of the pronounced cases of differences in CBR and syllable count. For example, SLTs may have had attention lapses, or may have hit a wrong key during part of an assessment. But whilst human error cannot be ruled out, the training and calibration was a strength of the study.

**Variation in Total Number of Syllables Across Raters**

Results revealed a median difference of 95 syllables in the total number of syllables identified between the three assessing SLTs, averaged across all infants. The variation in total number of syllables identified between the three assessing SLTs ranged from almost perfect
agreement (1 syllable), to extensive disagreement (801 syllables). Thus, the recordings clearly varied in difficulty of assessment, as well as some of the variation being due to human error. Nevertheless, the three raters assessed number of syllables similarly in many infants. As expected, these infants may primarily have produced well-formed speech-like syllables and relatively few nasalized vowels in isolation. Therefore, identification of speech-like syllables is challenging in some recordings but may only have influenced results to a limited degree as results showed high inter-rater agreement in classification of CB status (87.4%).

Overall, the per-infant average total number of syllables identified for infants classified as canonical was higher than for infants classified as not canonical. Thus, had larger speech samples been obtained it is possible that more infants might have been classified as canonical (Molemans et al., 2012). In total, 116 recordings had not been checked to contain at least 50 utterances; a cut-off chosen to allow enough speech to assess CB status and used in previous studies (e.g., Lang et al., 2020; Moeller et al., 2007; Rvachew et al., 1999). A total of ten recordings were assessed to include less than 50 syllables, of which seven were classified as not canonical and three as canonical. Hence, not all infants with few syllables were classified as not canonical. From longitudinal speech samples of 40 typically developing infants from 6-12 months of age, results of a bootstrapping procedure showed that at 12 months of age a sample size of 50 syllables confirmed CB in all the participants with 95% confidence (Molemans et al., 2012). Accordingly, the possible influence of total number of syllables on CBR, and classification of CB status in some infants in the present study, indicates that canonical babbling may not have been stable enough to be detected in one speech sample, as has been reported in previous studies (Lynch et al., 1995; Molemans et al., 2012). Therefore, although some of the infants may have been misclassified as not canonical
by the CBR criterion, they seemed to be (less stable and therefore) delayed in their
development of well-formed syllables.

In conclusion, this study shows new evidence of high inter-rater agreement between
three independent assessments of CB status by 29 trained SLTs based on real time assessment
of CBR across 484 infants born with cleft palate. As disagreements were distributed across
the SLTs, this means that one (trained) SLT/rater can reliably classify CB status, based on
CBR, in real time. Due to the reduced costs of real time assessment including the need for
only one trained rater, this methodology may be useful in longitudinal studies and in large
scale studies such as registry studies. The study also reveals new results by demonstrating
inter-rater variation in assessment of CBR and total number of syllables produced. Thus, we
recommend to include three raters in assessment of CBR and total number of syllables, if the
purpose is to compare these measures across infants or studies. Future study should include
assessment of intra-rater reliability to test stability of assessments over time, especially if
only one rater is included in classification of CB status.

Acknowledgements

Collaborators of the TOPS Speech Group:

Liz Albery, Bristol, UK; Helene Søgård Andersen, Copenhagen, Denmark; Malin, Appelqvist, Uppsala, Sweden; Pia Bodling, Linköping, Sweden; Melanie Bowden, Manchester, UK; Karin Brunnegård, Umeå, Sweden; Stephanie Van Eeden, Newcastle, UK; Josefin Enfalt, Umeå, Sweden; Ana Paula Fukushiro, Bauru, Brazil; Cristina Guedes de Azevedo Bento Goncalves, Bauru, Brazil; Line Dahl Jørgensen, Copenhagen, Denmark; Jorunn Lemvik, Oslo, Norway; Louise Leturgie, Århus, Denmark; Eva Liljerehn, Uppsala,
The authors wish to thank Joan Bogh Nielsen, Århus, Denmark; Lucy Rigby, Newcastle, UK; Nina-Helen Pedersen, Bergen, Norway, and Ragnhild Aukner, Oslo, Norway for contributing to data collection and analyses. We also wish to thank the participating clinical sites, the Data Coordinating Center, Liverpool, UK, and the Administrative Coordinating Center, Manchester, UK, for help with data collection, management of the twelve months visit data and preparation of the speech material.

**Disclosure of Interest**

The authors report no conflict of interest. This publication was made possible by Grants Number R21DE15128, U01DE018664 and U01DE018837 from the Institute of Dental and Craniofacial Research (NIDCR). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIDCR.
References


569_2003_040_0453_tiocto_2.0.co_2


https://doi.org/10.1597/09-105


https://doi.org/10.1111/j.1469-8749.2012.04123.x


https://doi.org/10.1016/j.ridd.2013.01.014


https://doi.org/10.1111/1467-7687.00161
Moeller, M.P., Hoover, B., Putman, C., Arbataitis, K., Bohnenkamp, G., Peterson, B., …
Stelmachowicz, P. Vocalizations of infants with hearing loss compared with infants
with normal hearing; Part II-transition to words. *Ear and Hearing, 28*(5), 628-642.
doi:10.1097/AUD.0b013e31812564c9

Molemans, I., van den Berg, R., Van Severen, L., & Gillis, S. (2012). How to measure
doi:10.1017/S0305000911000171

disability and validity of a simplified way of measuring canonical babbling
https://doi.org/10.1080/02699206.2017.1320588

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. Yeni-
Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology: Vol 1: Production*


sounds by parents. *Developmental Science, 4*(1), 49-60.
https://doi.org/10.1111/1467-7687.00148


doi:10.1017/S0305000910000425


https://doi.org/10.3109/02699206.2011.636499


