Journal of Communication Disorders 83 (2020) 105969

Contents lists available at ScienceDirect

Journal of Communication Disorders

journal homepage: www.elsevier.com/locate/jcomdis

Rating the overall speech quality of hearing-impaired children by means of comparative judgements

Nathalie Boonen*, Hanne Kloots, Steven Gillis

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp, Antwerp, Belgium

ARTICLE INFO

Keywords:
Language acquisition
Overall speech quality
Comparative judgement
Dutch
Children with a cochlear implant
Children with an acoustic hearing aid

ABSTRACT

Objective: Acoustic measurements have shown that the speech of hearing-impaired (HI) children deviates from the speech of normally hearing (NH) peers. The aim of the present study is to examine whether listeners with varying degrees of experience with (HI) children's speech perceive a difference in the overall speech quality of seven-year-old NH children and their HI peers who received a device before the age of two.

Method: Short speech samples of seven children with NH, seven children with an acoustic hearing aid (HA) and seven children with a cochlear implant (CI) were judged by three groups of listeners (audiologists, primary school teachers and inexperienced listeners) in a comparative judgement task. In this task, listeners compared stimuli in pairs and decided which stimulus sounded better, leading to a ranking of the stimuli according to their overall speech quality.

Results: The ranking showed that the overall speech quality differed considerably for HI and NH children. The latter group had a significantly higher overall speech quality than HI children. In the group of HI children, children with CI were ranked higher than children with HA. Moreover, length of device use was found to have a significant effect in the group of children with CI: longer device experience led to better ratings. This effect was significantly less strong in HA children. No significant differences were found between the three groups of listeners.

Conclusion: Listeners agree that the speech of NH children sounds better than the speech of HI children. This result indicates that even after almost seven years of device use, the speech of HI children still differs from the speech of NH children. The overall speech quality of CI children was better than that of HA children, and this effect increased with longer device use. No effect of listeners' experience with (NH and/or HI) children's speech was established.

1. Introduction

This study investigates the overall speech quality of hearing-impaired (HI) children in comparison to normally hearing (NH) children. Traditionally, the speech of HI children is assessed by means of acoustic studies (see §1.1.) or perceptual studies (see §1.2.). This study explores an alternative perceptual approach, viz. a comparative judgement task (see §1.3.).

* Corresponding author at: University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium.
E-mail addresses: nathalie.boonen@uantwerpen.be (N. Boonen), hanne.kloots@uantwerpen.be (H. Kloots), steven.gillis@uantwerpen.be (S. Gillis).

https://doi.org/10.1016/j.jcomdis.2019.105969
Received 25 September 2018; Received in revised form 20 November 2019; Accepted 1 December 2019
Available online 04 December 2019
0021-9924/

1.1. Acoustic studies on the speech characteristics of HI children

Approximately three per thousand newborns suffer from a hearing impairment. Due to the Universal Neonatal Hearing Screening (UNHS), auditory impairments are identified at an increasingly early age. As a consequence of this early diagnosis, rehabilitation can start at a very young age. Depending on aspects such as the locus and the severity of the hearing impairment, an acoustic hearing aid (HA) or a cochlear implant (CI) is fitted (Korver et al., 2017). Although these devices do not lead to perfect hearing, they result in the development of speech perception and speech production skills, which, for many years, were considered to be beyond what could be achieved by (severely to profoundly) HI children (Flipsen & Colvard, 2006). However, acoustic studies revealed that the speech of HI children differs from the speech of NH children. Their speech shows segmental deviations, i.e., deviations in the pronunciation of vowels and consonants (Baudonck, Dhooge, D'haeseleer, & Van Lierde, 2010; Verhoeven, Hide, De Maeyer, Gillis, & Gillis, 2016), as well as deviations at the suprasegmental level, such as inappropriate use of stress and a slower speech rate (Lenden & Flipsen, 2007; Vanormelingen, De Maeyer, & Gillis, 2016). Thus, there are marked, measurable differences between NH children's speech and that of HI peers.

In the group of HI children, the speech of children with CI and children with HA also differ. For example, their acoustic vowel spaces have discrepant magnitudes: compared to NH children, children with CI and HA exhibit a smaller vowel space, and the vowel space of children with CI is even smaller than that of children with HA (Verhoeven et al., 2016 - but see Baudonck, Van Lierde, Dhooge, & Corthals, 2011). HI children tend to prefer visible consonants such as labials, exhibit cluster reduction and have difficulties with high frequency sibilants. In these respects, children with CI seem to perform more accurately than children with HA (Van Lierde, Vinck, Baudonck, De Vel, & Dhooge, 2005). With respect to suprasegmental elements, nasality has been shown to be more deviant in children with HA than in children with CI (Baudonck, Van Lierde, D'haeseleer, & Dhooge, 2015).

Thus, the literature shows that even after several years of device use, particular aspects of the speech of HI children deviate acoustically from NH children's speech. But it is still unknown whether the subtle though significant acoustic differences are apparent to the human ear. In a previous study, using a categorisation task, it has already been reported that participants relatively accurately distinguished the speech of HI children from the speech of NH children. In other words: listeners predominantly categorised the utterances of HI children correctly, leading to the conclusion that the speech of HI children is identifiable (Boonen, Kloots, Verhoeven, & Gillis, 2019).

1.2. Perceptual studies and the role of listeners' experience

Whereas in acoustic studies particular characteristics of speech samples are measured, the focus of perceptual experiments shifts to listeners and their assessment of particular aspects of speech. Perceptual studies have already covered several characteristics of the speech of HI individuals, such as nasality (Baudonck et al., 2015), intelligibility (Chin, Bergeson, & Phan, 2012; Montag, AuBuchon, Pisoni, & Kronenberger, 2014), and prosody (Lenden & Flipsen, 2007). In these studies, listeners hear spoken language stimuli of HI children's speech and are asked to judge a specific characteristic or to make a phonetic or orthographic transcription of the speech they hear.

The main point is that perceptual studies do not rely on objective assessments such as duration or amplitude measures in a software package such as PRAAT (Boersma & Weenink, 2016), but on the judgements of listeners. As a result, perceptual assessments are by definition more subjective. For example, research has shown that experience with a particular type of speech has an influence on listeners' judgements or appreciation. Listeners who are used to hearing more variation in speech are also more sensitive to and more proficient in hearing subtle differences than listeners without experience (Munson, Johnson, & Edwards, 2012). Thus, the amount of experience is a factor that can explain results to some extent. Therefore, experienced and inexperienced listeners should be treated as separate groups.

1.3. Alternative perceptual approach: comparative judgement

Rating scales are commonly used when listeners evaluate speech in an experimental context, for instance when judging speech intelligibility (AlSanosi & Hassan, 2014; Calmels et al., 2004; Fang et al., 2014; Van Lierde et al., 2005). On a rating scale, listeners map how they perceive a specific characteristic of the speech they hear. For instance, on a numerical scale, participants indicate their rating by providing a number to assess the intelligibility of the speech sample varying from "very unintelligible" to "very intelligible" (Colman, 2009). Alternatively, on a verbal rating scale such as the SIR (Cox & McDaniel, 1989), each point on the scale has a verbal description going from "Connected speech is unintelligible" to "Connected speech is intelligible to all listeners". The rating procedure for all these types of scales is similar. Listeners hear a stimulus, reflect on a specific characteristic by using an implicit mental reference point and indicate their appreciation on the scale. This procedure is repeated for all stimuli separately.

Rating scales have advantages as well as disadvantages. The main advantages of using rating scales are that they allow a more holistic judgement and they are far less time consuming to administer than other types of tasks, such as a transcription task. One of the disadvantages of rating scales is that, because each participant's approach in scoring on a rating scale may differ, rating scales are fairly subjective which results in low reliability scores and high intra- and interjudge variability (Miller, 2013; Munson et al., 2012). Moreover, rating specific speech characteristics on a scale is usually perceived as a difficult task especially when the rating involves hearing subtle differences, which makes them more suited for experienced listeners (Chin & Kuhns, 2014; Miller, 2013; Schiavetti, 1992).

In the tasks described above, participants hear the stimuli one at a time and they make their judgements by comparing each stimulus with an implicit mental representation. In the present study, a methodology is used that involves the direct perceptual comparison of two stimuli. In a comparative judgement task, listeners hear two stimuli, one after the other. They are instructed to

compare the two stimuli and decide which stimulus has the higher overall quality or possesses a certain attribute more than the other stimulus (Bramley, 2015; Thurstone, 1927). The idea behind this type of task was described by Thurstone (1927), who argued that stimuli can be judged by comparison if they share a characteristic that can be ranked. He gave the example of assessing the "greyness" of grey color chips. Rating a single grey value on a typical numerical scale is quite a challenge. But comparing two grey color chips and indicating, for instance, the darker one, is a far less challenging task (Bejar, 2012). By repeating this procedure for several pairs with different shades of grey, a ranking can be easily obtained from light grey to dark grey. Because the comparisons are quite straightforward and less arbitrary than the traditional rating scales, a higher reliability and consistency is obtained (Lesterhuis, Verhavert, Coertjens, Donche, & De Maeyer, 2017).

1.4. Hypotheses

1.4.1. Ranking: which child sounds better?

Similar to Boonen et al. (2019), the present study addresses HI children's speech identifiability. However, whereas in Boonen et al. (2019), the listeners explicitly categorised the children by identifying their hearing status, the present study approaches speech identifiability more implicitly by judging the children's overall perceived speech quality, i.e., the general impression of the quality of their speech (Kondo, 2012). More specifically, the overall perceived speech quality of NH and HI children is examined using a comparative judgement task. This is a relatively new method in the field of language acquisition research. The outcome of a comparative judgement task is a ranking. In our experiment, each stimulus receives a ranking score according to its overall speech quality (independent of the hearing status of the speaker). In this study, stimuli of NH and HI children are included. The latter group consists of children with CI and children with HA. This structure implies two separate analyses of the ranking, that is, an analysis in two steps. Firstly, we focus on the differences in the overall speech quality of NH vs. HI children. More specifically, the first research question is: Do listeners perceive a difference in the overall speech quality of NH and HI children? Secondly, we differentiate between children with CI and HA. The research question here is: Do listeners also perceive differences in the overall speech quality of children with CI and children with HA?

For the first step, we hypothesize that, if the overall speech quality of NH and HI children differs, this results in a ranking in which the two hearing statuses (NH and HI) are clearly separated. In other words, the stimuli of NH children are expected to be on one side of the ranking, whereas the stimuli of HI children are expected to be found on the other side of the ranking. If listeners do not pick up differences in the speech of the two groups, both groups' stimuli are distributed over the whole ranking, indicating that the acoustic differences are too small or too subtle for listeners to differentiate between NH and HI children's speech.

In a second step, possible differences between children with CI and children with HA are investigated. It is difficult to predict which hearing status is ranked higher since some studies found better results in CI children (Baudonck, Dhooge, & Van Lierde, 2010; Lejeune & Demanez, 2006; Van Lierde et al., 2005), whereas other studies found better results in children with HA (Verhoeven et al., 2016). Especially in children with CI, length of device use appears to be a significant predicting variable in previous research, as it has been shown that their speech continues to improve after implantation (Fang et al., 2014; Gillis, 2017; Tomblin, Spencer, Flock, Tyler, & Gantz, 1999; Yoshinaga-Itano, Baca, & Sedey, 2010). Consequently, CI children with longer device use are expected to be ranked higher than children with shorter device use.

1.4.2. Effect of listener group

In the present study, three listener groups differing in the amount of experience with the speech of (HI) children participated, viz. audiologists, primary school teachers and inexperienced listeners. With regard to the effect of listener groups, the question is whether experienced listeners, i.e., primary school teachers and audiologists, judge the overall speech quality differently from inexperienced listeners. Earlier studies using comparative judgements suggested that judgements of experienced as well as inexperienced participants led to a reliable and comparable ranking (Jones & Alcock, 2014). Consequently, we expect the rankings of the audiologists, primary school teachers and inexperienced listeners to be not markedly different. However, other studies have shown that experience with a specific type of speech does influence the speech perception of listeners and their rating behaviour in an experimental context (Beukelman & Yorkston, 1980; Munson et al., 2012). This implies that it could equally reasonably be hypothesized that the different backgrounds of the listeners lead to different rankings. Considering that audiologists and primary school teachers are familiar with the speech of children, we assume that these listeners are more likely to hear a difference in the overall speech quality of NH and HI children. If this is the case, the rankings of the audiologists and primary school teachers are expected to resemble each other, whereas the ranking of the inexperienced listeners is expected to differ. Moreover, when considering HI children as two separate groups, it can be hypothesized that audiologists are more successful in making a distinction between HA and CI children. In that sense, their ranking is expected to differ from that of the other two listener groups.

2. Method

In this study, short stimuli of children with normal hearing (NH), children with an acoustic hearing aid (HA) and children with a cochlear implant (CI) were judged by three groups of listeners in a comparative judgement task. This study was approved by the Ethics Committee for the Social Sciences and Humanities (SHW_15_37) of the University of Antwerp. The participating listeners as well as the (caregivers of the) children were informed about the goal of the study and gave their written informed consent.

2.1. Stimuli

2.1.1. Audio recordings

In the present study, a selection of existing speech samples was used. The samples originated from recordings of an imitation task, which were made as part of an earlier study on the speech of NH and HI children (Hide, 2013). In the imitation task, speech of one hundred-eleven children was collected: 11 children with CI, 10 children with HA, and 90 NH children. They were all native speakers of Dutch and enrolled in the mainstream education system in Flanders, the northern, Dutch speaking part of Belgium. The children were instructed to imitate a carrier sentence in which a disyllabic pseudo-word was embedded ("Ik heb X gezegd", "I have said X"), where X represents /lVlV/ (with V = /a/, /e/ or /o/). All recordings were made in a quiet setting in the comfort of the children's homes or schools.

2.1.2. Selection of the experimental stimuli

The speech samples used in the present study came from seven children with CI, seven children with HA and seven NH children. The children were randomly selected from the study discussed above. The sample contained six utterances of each child. This resulted in a set of 126 stimuli that was used in the experiment. The same stimuli were used in Boonen et al. (2019). Detailed information on the individual children with CI and HA can be found in Boonen et al. (2019).

a Children with CI

The average age of the children with CI (four girls, three boys) at the time of the recording was 7;10 (years;months) (SD = 1;1). The mean age of implantation was 12 months (SD = 0;6). On average, the children had 6;9 of device use at the time of the recording (SD = 1;5). Six children were implanted bilaterally and had on average 3;11 of bilateral device experience. Before implantation, the children's mean unaided hearing loss level was 116 dB (SD = 7 dB), which evolved to an average of 29 dB (SD = 7 dB) at the time of the recordings. At the moment of the recording, the teachers and/or caregivers were explicitly asked about additional disabilities and they confirmed that the children did not have any disabilities apart from the hearing loss.

b Children with HA

The average age of the children with bilateral HAs (four girls, three boys) at the time of the recording was 7;9 years (SD = 0;11 years) which is not significantly different from the chronological age of the children with CI (Wilcoxon Rank Sum Test: z = 0.00, p = 1.0). They received their HAs around the age of 0;11 (SD = 0;7). The children had on average 6;10 years of device use at the time of the recording (SD = 1;6) with a minimum of four years. Before receiving HAs, the children's mean unaided hearing loss level was 66 dB (SD = 15 dB), which evolved to an average of 33 dB at the time of the recording (SD = 7 dB). The CI and HA children's aided hearing levels were comparable (Wilcoxon Rank Sum Test: z = 0.91, p = 0.37). At the moment of the recording, the teachers and/or caregivers were explicitly asked about additional disabilities and they confirmed that the children did not have any disabilities apart from the hearing loss.

c Children with NH

Seven NH children (four girls, three boys), who attended the same primary schools as the CI children, participated in this study. These children were matched on gender, age and regional background with the HI children. Their hearing was assessed in the first month of life with an automated auditory brainstem response test (AABR) or otoacoustic emissions (OAE) as part of the Universal Neonatal Hearing Screening. At the moment of the recording, the teachers and/or caregivers were explicitly asked about disabilities and they confirmed that the children did not have any disabilities.

2.2. Listeners

The participating listeners (n = 60) were all native speakers of Dutch and lived in the same region of Belgium (province of Limburg). The listeners self-reported to have no hearing problems.

Three groups of 20 listeners with varying degrees of experience with the speech of (HI) children participated in the perception experiment: audiologists, primary school teachers and inexperienced listeners. The first group consisted of speech and language therapists with a specialisation in audiology, henceforth audiologists. They were on average 36 years old (SD = 7 years). Their mean experience as an audiologist was 12 years (SD = 7 years) in which they gained theoretical and practical experience with the speech of HI children. The second group consisted of primary school teachers. They were on average 40 years old (SD = 8 years), had a mean of 17 years of experience as a teacher (SD = 8 years), and were obviously familiar with the speech of NH children. The third group were naïve listeners without any specific experience with the speech of (HI) children. They were on average 41 years old (SD = 12 years) and will henceforth be referred to as inexperienced listeners.

2.3. Procedure

The listeners sat in front of a computer screen and listened to the stimuli through high quality headphones (type: Bowers & Wilkens P5) set at a comfortable volume. Each listener made 65 comparisons of two stimuli (see last section of this paragraph). For

each comparison, the listeners compared the two stimuli and decided "which one sounded better". The instruction was deliberately phrased in general terms in order not to guide the listeners into a specific direction. They were stimulated to take into account whatever aspect they thought was decisive for each comparison. The decision was made by clicking the appropriate box (child A or child B) on the computer screen. Throughout the experiment, the stimuli could be repeated as many times as the listener wanted.

Prior to the experiment, two text boxes with the instructions for the experiment were presented. The first introduced the task and specified how to complete it. The second text box mentioned two possibly misleading aspects that should not influence the responses. The first point concerned regional variation. Considering that the children in the sample lived in various regions of Flanders, some regional variation was present in the speech samples. The listeners were asked not to let regional variation lead their decision. Secondly, listeners were instructed not to pay attention to the loudness and the sound quality of the recordings. The listeners were informed that they would hear sentences spoken by primary-school aged children with CI, children with HA and NH children. No additional information about the typical speech characteristics of these children was provided. The participants were also informed that for each comparison two stimuli were randomly paired so that samples of children with different hearing statuses as well as identical hearing statuses could be paired.

The comparative judgement task was implemented in the online tool D-PAC (Digital Platform for the Assessment of Competence) (Lesterhuis et al., 2017). In this task, the judgements of each listener group led to a ranking of the set of 126 stimuli with respect to their overall speech quality. An exhaustive pairing of all stimuli would lead to a total of 7875 possible pairs for each listener group. Obviously, such a large number of pairs would undermine the practical feasibility of the experiment. However, in order to arrive at a reliable ranking, not all possible combinations of stimuli had to be assessed. More specifically, each stimulus had to be judged by the listeners of each particular listener group – together, not individually – at least 20 times. This number was established in a previous study that found that the number of times a stimulus is judged is an important contributing factor to the reliability of the eventual ranking and that, after 20 rounds, the reliability of the ranking reached a ceiling level (Verhavert, 2018). Thus, each listener group judged each stimulus 20 times. For each of the three listener groups, this resulted in 63 comparisons for each listener (rounded off to 65, see formula (1)), and 1300 comparisons per listener group.

$$number\ \ of\ \ comparisons\ \ per\ \ listener = \frac{number\ \ of\ \ stimuli \times 20}{number\ \ of\ \ listeners}/2 \tag{1}$$

2.4. Data analysis

The basic building block in a comparative judgement task is the individual comparison of two stimuli. A participant indicates whether $stimulus_a$ "wins" the competition of $stimulus_b$. Over all the comparisons made by the listeners of a particular listener group, it is calculated how often a particular stimulus "won" a competition and how often it lost. These calculations provide the input for the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952), which is used to compute the likelihood that a particular stimulus "wins" a competition. Mathematically, the BTL model is formulated in Eq. (2) (Bradley & Terry, 1952; Verhavert, De Maeyer, Donche, & Coertjens, 2018).

$$p(x_{ij} = 1|\nu_i, \nu_j) = \frac{e^{(\nu_j - \nu_i)}}{1 + e^{(\nu_j - \nu_i)}} \tag{2}$$

where $x_{ij} = 1$ if stimulus j is considered to exhibit better overall speech quality than stimulus i, and ν_i and ν_j are the estimated logit scores of the respective stimuli.

This model led to scores expressed in logits. In other words, the logit scores that were attributed to each stimulus were calculated at the end of the assessment, rather than for each listener individually. The lowest logit score indicated the stimulus with the lowest overall speech quality, whereas the stimulus with the highest overall speech quality had the highest logit score. Numerically ordering the logits from low to high resulted in a ranking that represented the stimuli according to their overall speech quality. For the statistical analysis, these logit scores were used since the distance between two stimuli in the ranking was variable, whereas an ordinal scale would suggest a constant distance between two stimuli. Next, the reliability of the ranking was calculated using the Scale Separation Reliability (SSR) measure, which assessed the likelihood that the rank order was due to a measuring error (Andrich, 1982). The SSR of this experiment resulted in a mean score of .87, which meant that the likelihood that this ranking was the result of measuring errors was fairly small (Andrich, 1982; Lesterhuis et al., 2017).

Statistical analyses were performed by means of multilevel mixed-effect modeling (MLM) in the open source software R (packages *lme4* and *lmerTest*) (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017; R Core Team, 2016). Crucial in MLM is the distinction between fixed effects and random effects in a model: the fixed effects represent the variables with repeatable levels, such as the distinction between HI and NH children. The random effects represent the variables with levels randomly sampled from a population, such as the particular children whose speech samples are judged and the particular speech samples used in the experiment. Building the best fitting model is an iterative process. First, random effects are added to the null model. In the next step, fixed effects are added one after the other.

In this study, the fixed effects were the factor Hearing status (with values NH and HI or the values NH, CI and HA depending on the analysis), Length of device use (for HI children, this is the amount of time between the moment when they started using their device and the moment of the recording of the speech samples) and Listener group (audiologists, primary school teachers and inexperienced listeners). The random part consisted of the individual children and the individual utterances. Considering that the fixed effect Hearing status is most relevant for this study, this factor was entered as the first fixed effect. Next, the factor Listener

group was added (1) as a main effect and (2) in interaction with Hearing status. This order also applied for the second analysis, but here, the factor Length of device use was also added. Similarly to the factor Listener group, the factor Length of device use was first entered as a main effect and next, it was entered in interaction with Hearing status.

At each step, it is assessed through a likelihood ratio test if the resulting model yields a better fit. Only the predictors that significantly improve the model fit are retained and only the best fitting model is reported in the result section. In addition, the random effects model is compared with the best fitting model by means of a likelihood ratio test. More specifically, the likelihood ratio test takes into account the difference between the negative square log-likelihood ratios of both models (expressed as Δ -2LL) and the difference between the degrees of freedom (expressed as Δ df) in order to assess whether one model provides a significantly better fit than another model. In addition, both models are compared in terms of the difference in AIC (Akaike information criterion) values (Burnham & Anderson, 2004; McElreath, 2018).

The tables in this results section are expressed in logits, but for reasons of familiarity and readability, they are further discussed in terms of probabilities. Each fixed effect in a model is assigned a reference category, which is also mentioned in the tables. A significance level of $p < 0.05$ was set.

3. Results

This study investigates whether listeners with a varying degree of experience with children's speech perceive a difference in the overall speech quality of normally hearing (NH) and hearing-impaired (HI) children. In accordance with the three main research questions, this section will be subdivided into three parts: (1) comparing the overall speech quality of HI children, treated as one group, to NH children, (2) investigating and comparing the overall speech quality of children with CI and children with HA as two separate groups, and (3) investigating the role of listeners' experience in the overall speech quality judgements.

This section contains the results of 56 listeners. A total of 60 adult listeners participated in our experiment. However, a misfit analysis (Lesterhuis et al., 2017) showed deviant responses (> 2 SD) in the results of four participants (one audiologist, one primary school teacher and two inexperienced listeners). These participants were excluded from the statistical analyses.

3.1. Overall speech quality of NH and HI children

Firstly, we compare the position on the ranking of NH and HI children. This ranking – ordering all stimuli according to their overall speech quality based on the outcomes of the BTL model – is a representation of all the comparisons made by all listeners. At this point, children with a cochlear implant (CI) and children with an acoustic hearing aid (HA) will be treated as one group. The dependent variable in this analysis is numerical: a logit score which reflects the likelihood of each stimulus to "win" a comparison. The score is also representative for the overall speech quality: a higher logit score indicates a higher overall speech quality and vice versa.

The best fitting model is reported in Table 1. Compared to the model containing merely random effects, the likelihood ratio test showed that the reported model's fit is significantly better (Δ -2LL = 13.01 with Δ df = 1, $p < 0.001$; Δ AIC = 11). The results indicate that the logit scores of NH and HI children differ significantly. More precisely, NH children's score is significantly higher ($p < 0.001$) than that of HI children, meaning that NH children in general sound "better" and thus have a higher overall speech quality than HI children. This result is also visualised in Fig. 1 (left panel), which shows that NH children are more likely to "win" in a pairwise comparison and, hence, to be ranked higher than HI children.

3.2. Overall speech quality of CI and HA children

Do listeners appreciate the overall speech quality of children with CI in a different way than the overall quality of HA children's speech? In order to answer this question, the HI children are subdivided into CI and HA children. In addition, the factor Length of device use is often considered to be an important predicting variable for speech and language outcomes, especially for CI children. Therefore, this variable is included in the analysis of both HI groups. The best fitting model (Table 2) contains the fixed effects Hearing status and Length of device use and a likelihood ratio test showed that this model is a significantly better fit than the model containing only random effects (Δ -2LL = 10.98 with Δ df = 3, $p = 0.012$; Δ AIC = 4.98). The model indicates that, at intercept (which is set at the mean Length of device use of 85 months), CI children's score differs significantly from that of HA children ($p = 0.018$). More specifically, children with a CI score higher than children with HA. This result is also visualised in the righthand panel of Fig. 1, which shows that the overall speech quality of both HI groups is lower than that of NH children, but also that the

Table 1
Parameter estimates of the MLM model estimating the overall speech quality for NH and HI children.

Random effects	**Variance**	**Std. dev.**		**p**
Individual children	1.445	1.202		< 0.001
Individual utterances	0.025	0.159		0.036
Fixed effects	**Estimate**	**Std. error**	**t-value**	**p**
Intercept	1.566	0.466	3.360	< 0.001
Hearing status [HI]	−2.399	0.565	−4.245	< 0.001

NH = reference category.

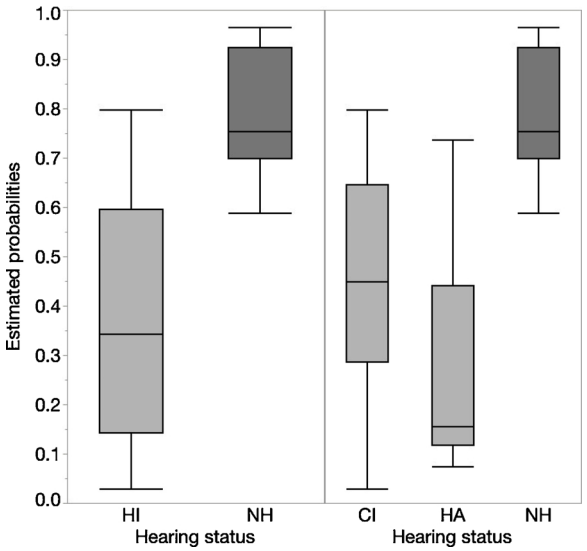

Fig. 1. Comparison of the overall speech quality judgements of NH and HI children (as one group in the left panel and subdivided into children with CI and HA in the right panel), the y-axis represents the estimated probabilities that the speakers are judged to exhibit better overall speech quality).

Table 2
Parameter estimates of the MLM model estimating the overall speech quality of CI and HA children.

Random effects	**Variance**	**Std. dev.**		**p**
Individual children	0.787	0.887		< 0.001
Individual utterances	0.057	0.240		0.001
Fixed effects	**Estimate**	**Std. error**	**t-value**	**p**
Intercept	−0.098	0.365	−0.268	0.789
Hearing status [HA]	−1.173	0.495	−2.371	0.018
Length of device use	0.081	0.022	3.648	< 0.001
Hearing status [HA] * Length of device use	−0.079	0.031	−2.549	0.011

CI = reference category.

overall speech quality of children with CI is higher than that of children with HA. Moreover, the effect of Length of device use is significant, indicating that children's overall speech quality is higher when they have more experience with their device. The interaction effect of Hearing status and Length of device use is visualised in Fig. 2, which clearly shows that an increase of the length of device use has a different effect on the overall speech quality of CI and HA children. For CI children, the overall speech quality increases as the length of device use increases. For children with HA, this evolution is barely noticeable.

3.3. Effect of the degree of listeners' experience

In both analyses, the best fitting model did not contain the factor Listener group (audiologists, primary school teachers and inexperienced listeners). This should be interpreted as follows: in the process of building the best fitting models, the factor Listener

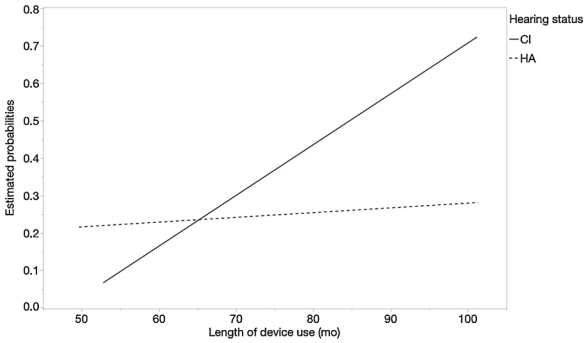

Fig. 2. Comparison of the overall speech quality judgements of CI and HA children as a function of length of device use (the y-axis represents the estimated probabilities that the speakers are judged to exhibit better overall speech quality).

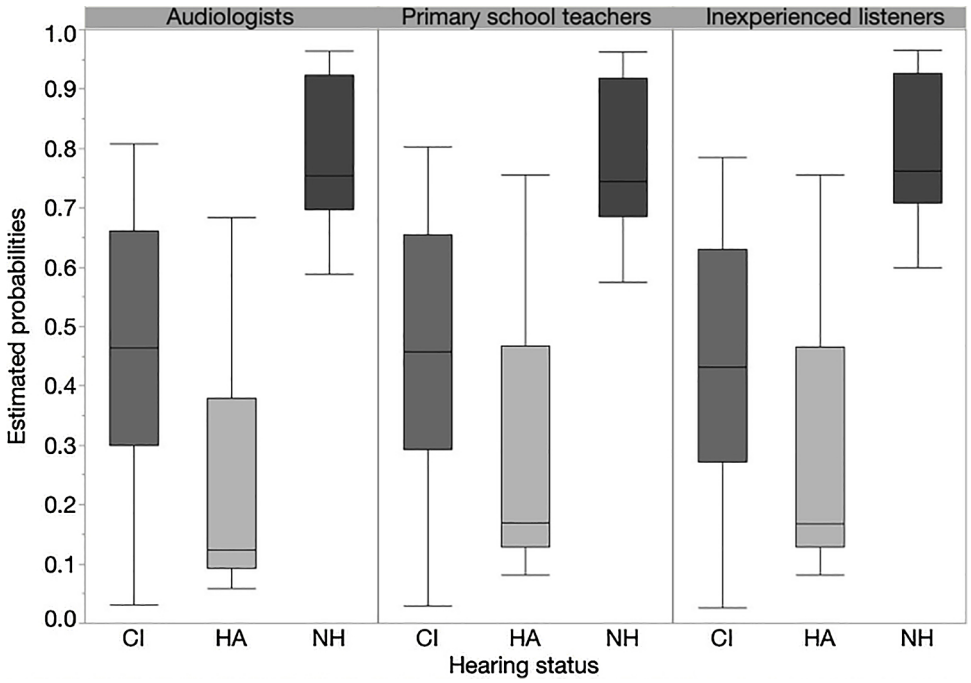

Fig. 3. Comparison of the overall speech quality judgements of NH, CI and HA children in the three listener groups (the y-axis represents the estimated probabilities that the speakers are judged to exhibit better overall speech quality).

group was added as a fixed effect. However, this factor did not lead to a better model fit. This result implies that the likelihood that a certain hearing status is selected as exhibiting the higher overall speech quality is comparable for audiologists, primary school teachers and inexperienced listeners, as is also visualised in Fig. 3. In other words: independent of their degree of experience with children's speech, listeners assess the overall speech quality of NH, CI and HA children in a similar way.

4. Discussion

The purpose of this study was to compare the overall perceived speech quality of hearing-impaired (HI) children with that of normally hearing (NH) children by means of a comparative judgement task completed by listeners with varying degrees of experience with children's speech.

4.1. Ranking: which child sounds better?

By repeatedly comparing speech stimuli, a ranking representing the overall speech quality is established. In this ranking, NH children are ranked higher – and thus were judged to sound better – than HI children. This result strongly suggests that listeners hear a qualitative difference between the two types of speech, which corroborates the findings of Boonen et al. (2019) who used the same speech samples but in a different experimental setup. They found that listeners are able to reliably distinguish the speech of NH children from the speech of HI children. After several years of device use, the speech of HI children apparently still sounds different from that of NH children. This result suggests that listeners possibly pick up some of the deviant characteristics that are found in the speech of HI children (Baudonck, Dhooge, D'haeseleer et al., 2010; Lenden & Flipsen, 2007; Vanormelingen et al., 2016; Verhoeven et al., 2016). Further research is needed to determine which characteristics guided listeners in their decision process.

In the HI group, children with a cochlear implant (CI) and children with an acoustic hearing aid (HA) were represented. The present research demonstrates that children with CI are ranked higher than children with HA, which means that the former sound significantly better than the latter. This result is not completely in agreement with the study of Boonen et al. (2019), who found that the speech of CI and HA children could not be differentiated above chance level, but they also found that children with CI were more often than HA children categorised as NH, thus showing a qualitative difference. Moreover, the overall quality of children's speech is partly determined by the factor length of device use for children with CI and HA. But this factor is more outspoken in children with CI. In this group, the score for overall speech quality considerably increases with longer device use. This result indicates that children with longer device use exhibit better overall speech quality than children with less experience. This significant progress is in agreement with several other studies. For example, Svirsky, Robbins, Kirk, Pisoni, and Miyamoto (2000) found that the rate at which children with CI acquire their speech and language skills is comparable to that of NH children. For the HA children in this study, this rapid improvement is barely observed. On the contrary, the overall speech quality for children with less experience is similar to the overall speech quality of children with more device experience. This result is similar to the study of Bat-Chava, Martin, and Kosciw

(2005) in which the speech of children with HA improved significantly slower than that of children with CI. Other studies comparing the children with CI and HA also found better results in CI children, for example with respect to speech intelligibility (Baudonck, Dhooge, Van Lierde et al., 2010; Lejeune & Demanez, 2006; Van Lierde et al., 2005).

4.2. Differences in listener groups with varying degree of experience

Concerning the three listener groups, no significant differences in the ranking and the individual comparisons are found. This result indicates that experience does not seem to influence a listener's notion of which stimulus has the highest overall speech quality. This finding is not in line with previous research stating that listeners' experience and knowledge about a specific type of speech influences their judgement of that particular type of speech (Beukelman & Yorkston, 1980; Munson et al., 2012). The contradicting findings can possibly be ascribed to the methodologically differing approaches of the studies. In the previous studies, listeners rated speech stimuli separately and therefore needed an implicit mental reference point. For experienced listeners, this reference point was established by their prior experiences, whereas for inexperienced listeners, this reference point may be lacking. In contrast, in the present study, the reference point is explicitly present since the stimuli are judged in pairs rather than separately. Therefore, the listeners did not require any prior knowledge or experience with the speech of HI children. In summary, whereas rating speech samples was considered to be a task for experienced individuals in previous studies (Chin & Kuhns, 2014; Miller, 2013; Schiavetti, 1992), comparative judgements seem to provide an alternative that is feasible for inexperienced listeners as well. This is in line with the study of Jones and Alcock (2014), who also found that a reliable ranking can be obtained by inexperienced as well as experienced listeners.

4.3. Clinical implications

In this paragraph, we present some thoughts on the clinical implications of our experiment. In this study, the speech of NH children is judged to exhibit a better overall quality than the speech of HI children. This finding adds to the idea that the speech of HI children differs from that of NH children, even after several (almost seven in this study) years of device use. To clinicians, these results are of importance when informing parents and other caregivers about the long-term expectations with respect to the speech of HI children.

Rating children's speech in clinical practise has a long tradition of using scales. Reliably using such scales requires experience and has been shown to exhibit high intra- and interjudge variability (Miller, 2013; Munson et al., 2012). The use of comparative judgements may constitute a more reliable alternative in speech and language practices. Moreover, since this study has shown that the ranking of listeners with varying degrees of experience with the speech of HI children does not differ significantly, keeping track of speech improvements would not be limited to experienced listeners such as speech and language pathologists, but could also be extended to parents or school teachers. More research is certainly desirable into the practical applicability of a comparative judgement task in a clinical context.

4.4. Limitations

The research reported in the present paper has some limitations that need to be acknowledged. First of all, the number of children that were recorded and the number of recordings used as stimuli in this study are relatively small. Replicating this study with a larger sample is required to show the robustness of the findings. Moreover, the composition of the sample is also a matter of further attention. For example, in the present study, listeners found the overall speech quality of children with CI better than that of children with HA. The two groups of HI children had a comparable aided hearing loss, which was crucial to create comparable groups. However, for the children with HA, it led to a group that is not representative for the whole population of HA users. Our HA group had an average unaided threshold of 66 dB HL. This hearing loss was treated with a traditional acoustic HA, yet recent research suggests that children with a hearing loss above 65 dB HL could reach better results when treated with a CI (Leigh, Moran, Hollow, & Dowell, 2016). Considering that the hearing loss in our sample was moderate to severe, children with slight to mild hearing losses were left out. However, these degrees of hearing loss are very common. For example, mild hearing loss constitutes 42% of all cases of hearing loss in Australian children (Russ et al., 2009). Possibly different results would be observed in a group of HA children with lower PTA levels. For children with CI, less variation is found in the unaided hearing thresholds since most children start with a severe to profound hearing loss. Thus, in contrast to the HA children in our sample, the children with CI are representative to the complete group of child CI users.

5. Conclusion

The present study shows that listeners with varying degrees of experience judge the overall speech quality of normally hearing (NH) and hearing-impaired (HI) children to be different when comparing the stimuli pairwise. The speech of NH children is considered to be better sounding and is preferred by listeners. This implies that the acoustic differences in the speech of HI children are also perceived by listeners. In the HI group, children with a cochlear implant have a higher overall perceived speech quality than children with an acoustic hearing aid, especially with a longer length of device use. Listeners, irrespectively of their degree of experience with (HI) children's speech, completed the task similarly, indicating that a comparative judgement task does not require special expertise and is feasible for different types of listeners.

CRediT authorship contribution statement

Nathalie Boonen: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Hanne Kloots:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision. **Steven Gillis:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgment

This project was funded by a predoctoral research grant of the Research Foundation – Flanders (FWO) to the first author (1100316N).

References

AlSanosi, A., & Hassan, S. M. (2014). The effect of age at cochlear implantation outcomes in Saudi children. *International Journal of Pediatric Otorhinolaryngology, 78*(2), 272–276. https://doi.org/10.1016/j.ijporl.2013.11.021.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(1), 95–104.

Bat-Chava, Y., Martin, D., & Kosciw, J. G. (2005). Longitudinal improvements in communication and socialization of deaf children with cochlear implants and hearing aids: Evidence from parental reports. *Journal of Child Psychology and Psychiatry, 46*(12), 1287–1296. https://doi.org/10.1111/j.1469-7610.2005.01426.x.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Baudonck, N., Dhooge, I., D'haeseleer, E., & Van Lierde, K. (2010). A comparison of the consonant production between Dutch children using cochlear implants and children using hearing aids. *International Journal of Pediatric Otorhinolaryngology, 74*(4), 416–421. https://doi.org/10.1016/j.ijporl.2010.01.017.

Baudonck, N., Dhooge, I., & Van Lierde, K. (2010). Intelligibility of hearing impaired children as judged by their parents: A comparison between children using cochlear implants and children using hearing aids. *International Journal of Pediatric Otorhinolaryngology, 74*(11), 1310–1315. https://doi.org/10.1016/j.ijporl.2010.08.011.

Baudonck, N., Van Lierde, K., D'haeseleer, E., & Dhooge, I. (2015). Nasalance and nasality in children with cochlear implants and children with hearing aids. *International Journal of Pediatric Otorhinolaryngology, 79*(4), 541–545. https://doi.org/10.1016/j.ijporl.2015.01.025.

Baudonck, N., Van Lierde, K., Dhooge, I., & Corthals, P. (2011). A comparison of vowel productions in prelingually deaf children using cochlear implants, severe hearing-impaired children using conventional hearing aids and normal-hearing children. *Folia Phoniatrica et Logopaedica, 63*(3), 154–160. https://doi.org/10.1159/000318879.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x.

Beukelman, D. R., & Yorkston, K. M. (1980). Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of Communication Disorders, 13*(1), 33–41. https://doi.org/10.1016/0021-9924(80)90019-2.

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer (Version 5.3).* Retrieved fromwww.praat.org.

Boonen, N., Kloots, H., Verhoeven, J., & Gillis, S. (2019). Can listeners hear the difference between children with normal hearing and children with a hearing impairment? *Clinical Linguistics & Phonetics, 33*(4), 316–333. https://doi.org/10.1080/02699206.2018.1513564.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. *Biometrika, 39*(3-4), 324–345. https://doi.org/10.1093/biomet/39.3-4.324.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment*Cambridge, UK: Cambridge Assessment Research Report.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644.

Calmels, M.-N., Saliba, I., Wanna, G., Cochard, N., Fillaux, J., Deguine, O., et al. (2004). Speech perception and speech intelligibility in children after cochlear implantation. *International Journal of Pediatric Otorhinolaryngology, 68*(3), 347–351. https://doi.org/10.1016/j.ijporl.2003.11.006.

Chin, S. B., Bergeson, T. R., & Phan, J. (2012). Speech intelligibility and prosody production in children with cochlear implants. *Journal of Communication Disorders, 45*(5), 355–366. https://doi.org/10.1016/j.jcomdis.2012.05.003.

Chin, S. B., & Kuhns, M. J. (2014). Proximate factors associated with speech intelligibility in children with cochlear implants: A preliminary study. *Clinical Linguistics & Phonetics, 28*(7-8), 532–542. https://doi.org/10.3109/02699206.2014.926997.

Colman, A. M. (2009). *A dictionary of psychology* (3rd ed.). Oxford: Oxford University Press.

Cox, R. M., & McDaniel, D. M. (1989). Development of the speech intelligibility rating (SIR) test for hearing aid comparisons. *Journal of Speech, Language, and Hearing Research, 32*(2), 347–352. https://doi.org/10.1044/jshr.3202.347.

Fang, H. Y., Ko, H. C., Wang, N. M., Fang, T. J., Chao, W. C., Tsou, Y. T., et al. (2014). Auditory performance and speech intelligibility of Mandarin-speaking children implanted before age 5. *International Journal of Pediatric Otorhinolaryngology, 78*(5), 799–803. https://doi.org/10.1016/j.ijporl.2014.02.014.

Flipsen, P., & Colvard, L. G. (2006). Intelligibility of conversational speech produced by children with cochlear implants. *Journal of Communication Disorders, 39*(2), 93–108. https://doi.org/10.1016/j.jcomdis.2005.11.001.

Gillis, S. (2017). Speech and language in congenitally deaf children with a cochlear implant. In A. Bar-On, & D. Ravid (Eds.). *Handbook of communication disorders: Theoretical, empirical, and applied linguistic perspectives* (pp. 763–790). Berlin: Mouton De Gruyter.

Hide, Ø. (2013). *Acoustic features of speech by young cochlear implant users. A comparison with normal-hearing and hearing-aided age mates. (Unpublished doctoral dissertation).* Antwerp, Belgium: University of Antwerp.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974.

Kondo, K. (2012). *Subjective quality measurement of speech.* Berlin/Heidelberg: Springer.

Korver, A. M., Smith, R. J., Van Camp, G., Schleiss, M. R., Bitner-Glindzicz, M. A., Lustig, L. R., et al. (2017). Congenital hearing loss. *Nature Reviews Disease Primers, 3*, 16094. https://doi.org/10.1038/nrdp.2016.94.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Leigh, J. R., Moran, M., Hollow, R., & Dowell, R. C. (2016). Evidence-based guidelines for recommending cochlear implantation for postlingually deafened adults. *International Journal of Audiology, 55*(S2), S3–S8. https://doi.org/10.3109/14992027.2016.1146415.

Lejeune, B., & Demanez, L. (2006). Speech discrimination and intelligibility: Outcome of deaf children fitted with hearing aids or cochlear implants. *B-ENT, 2*(2), 63–68.

Lenden, J. M., & Flipsen, P. (2007). Prosody and voice characteristics of children with cochlear implants. *Journal of Communication Disorders, 40*(1), 66–81. https://doi.org/10.1016/j.jcomdis.2006.04.004.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgment as a promising alternative to score competences. In E. Cano, & G. Ion (Eds.). *Innovative practices for higher education assessment and measurement* (pp. 119–138). Hershey: IGI Global.

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan.* Boca Raton, FL: CRC Press.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders, 48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061.

Montag, J. L., AuBuchon, A. M., Pisoni, D. B., & Kronenberger, W. G. (2014). Speech intelligibility in deaf children after long-term cochlear implant use. *Journal of Speech Language and Hearing Research, 57*(6), 2332–2343. https://doi.org/10.1044/2014_JSLHR-H-14-0190.

Munson, B., Johnson, J. M., & Edwards, J. (2012). The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners. *American Journal of Speech-Language Pathology, 21*(2), 124–139. https://doi.org/10.1044/1058-0360(2011/11-0009).

R Core Team (2016). *R: A language and environment for statistical computing.* Retrieved fromwww.R-project.org.

Russ, S. A., Poulakis, Z., Barker, M., Wake, M., Rickards, F., Sounders, K., et al. (2009). Epidemiology of congenital hearing loss in Victoria, Australia. *International Journal of Audiology, 42*(7), 385–390. https://doi.org/10.3109/14992020309080047.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.). *Intelligibility in speech disorders: Theory, measurement and management* (pp. 11–34). Amsterdam: John Benjamins Publishing Company.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science, 11*(2), 153–158. https://doi.org/10.1111/1467-9280.00231.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.

Tomblin, J. B., Spencer, L., Flock, S., Tyler, R., & Gantz, B. (1999). A comparison of language achievement in children with cochlear implants and children using hearing aids. *Journal of Speech Language and Hearing Research, 42*(2), 497–509. https://doi.org/10.1044/jslhr.4202.497.

Van Lierde, K. M., Vinck, B. M., Baudonck, N., De Vel, E., & Dhooge, I. (2005). Comparison of the overall intelligibility, articulation, resonance, and voice characteristics between children using cochlear implants and those using bilateral hearing aids: A pilot study. *International Journal of Audiology, 44*(8), 452–465. https://doi.org/10.1080/14992020500189146.

Vanormelingen, L., De Maeyer, S., & Gillis, S. (2016). A comparison of maternal and child language in normally-hearing and hearing-impaired children with cochlear implants. *Language, Interaction and Acquisition, 7*(2), 145–179. https://doi.org/10.1075/lia.7.2.01van.

Verhavert, S. (2018). *Beyond a mere rank order: The method, the reliability and the efficiency of comparative judgment. (Unpublished doctoral dissertation).* Antwerp, Belgium: University of Antwerp.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement, 42*(6), 428–445. https://doi.org/10.1177/0146621617748321.

Verhoeven, J., Hide, Ø., De Maeyer, S., Gillis, S., & Gillis, S. (2016). Hearing impairment and vowel production. A comparison between normally hearing, hearing-aided and cochlear implanted Dutch children. *Journal of Communication Disorders, 59*, 24–39. https://doi.org/10.1016/j.jcomdis.2015.10.007.

Yoshinaga-Itano, C., Baca, R. L., & Sedey, A. L. (2010). Describing the trajectory of language development in the presence of severe-to-profound hearing loss: A closer look at children with cochlear implants versus hearing aids. *Otology & Neurotology, 31*(8), 1268–1274. https://doi.org/10.1097/MAO.0b013e3181f1ce07.