Speech intelligibility measurement

A latent variable approach on utterances' transcriptions

Jose Rivera[1], Sven de Maeyer[2], and Steven Gillis[3]

[1] Department of Training and Education Sciences,
University of Antwerp, Antwerp, Belgium
E-mail: JoseManuel.RiveraEspejo@uantwerpen.be

(corresponding author)

[2] Department of Training and Education Sciences,
University of Antwerp, Antwerp, Belgium
E-mail: sven.demaeyer@uantwerpen.be

[3] Computational Linguistics, and Psycholinguistics Research Centre
University of Antwerp, Antwerp, Belgium
E-mail: steven.gillis@uantwerpen.be

June 9, 2022

Abstract

Contents

List of Figures

List of Tables

1. Introduction

Intelligible speech can be defined as the extent to which the elements in an speaker's acoustic signal, e.g. phonemes or words, can be correctly recovered by a listener [47, 73, 70, 36]. Because intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered [36], its attainment carries an important societal value, as it is a milestone in children's language development, the ultimate checkpoint for the success of speech therapy, and has been qualified as the 'gold standard' for assessing the benefit of cochlear implantation [13].

The literature suggest two perspectives from which *speech intelligibility* can be assessed: the message and listener's perspective [4, 5]. The first, also known as acoustic studies, is focused on assessing separately particular characteristics of the speech samples, e.g. their pitch, duration or stress (supra segmental characteristics), or the articulation of vowels and consonants (segmental characteristics) [63]. Whereas the second, also known as perceptual studies, is centered on making holistic assessments of the speech stimuli, e.g. measure their perceived quality [4, 5]. On both instances, the stimuli (children's utterances) can be generated from reading at loud, contextualized utterances, or spontaneous speech tasks[1].

Furthermore, perceptual studies can use multiple approaches to measure intelligibility, but they can be largely grouped into two: objective and subjective ratings [45]. In *objective rating* methods, listeners transcribe the children's utterances orthographically or phonetically, and use such information to construct a score. In that sense, in the transcription task, intelligibility can be inferred from the extent a set of transcribers can identify the word contained in an utterance [5]. In contrast, under *subjective rating* methods, listeners directly infer the utterance's intelligibility score through specific procedures, e.g. absolute holistic, analytic, or comparative judgments, among others.

The accoustic studies seems the most objective to me …

It is easy to deduce that *objective rating* methods produce more valid[2] and reliable[3] scores than its *subjective* counterpart, and as a result, are usually used as an objective measure of intelligibility [5, 27].

From a reader perspective, it's not clear why these factors matter?

Accompanying the intelligibility assessment methods, the literature supply a myriad of factors that are thought also contribute to the (under)development of intelligible spoken language [57, 6, 40, 28]. Among these are audiology related factors, such chronological age, age at implantation, the duration of device use, *hearing age*, bilateral or contralateral cochlear implantation, and the children's preoperative and postoperative hearing levels. On the other hand, there are also child related factors, such as the cause of the hearing impairment (genetic, infections), additional disabilities (mental retardation, speech motor problems), and gender. Finally, there are also environmental factors, such as communication modality.

Considering all of the above, this paper seeks to investigates the speech intelligibility levels of normal hearing (NH) versus hearing-impaired children with cochlear implants (HI/CI). For that purpose, ten utterances recordings, from thirty two NH and HI/CI children, were selected from a large corpus of *spontaneously spoken speech* collected by the CLiPS research center. Additionally, we set up an experiment, where one hundred language students transcribed each stimuli to the Qualtrics environment [77]. Finally, the transcriptions were transformed into an entropy measure per utterance, which served as our outcome variable.

Swap paragraphs

These three issues are at the core of you work (added value of your work to the literature). I would make that clear from the start (or at least earlier on in the introduction).

We believe this paper make three specific contributions to the understanding of the factors that drive the intelligibility of spoken language. First, we develop a novel analysis using a latent variable approach [26]. More specifically, we model *speech intelligibility* as a latent variable that can be inferred from the entropy replicates. This method offers three specific benefits. On the one hand, the method 'constructs' an intelligibility score, which in turn, allow us to test different hypothesis and even make individual comparisons at the appropriate level. On the other hand, it allow us to control for different sources of variation. This is particularly important as, by failing to account for the appropriate hierarchies in the data, we could be 'manufacturing' false confidence in the parameter estimates, leading us to incorrect inferences [53]. Finally, the method also provides a 'criterion' on how reliable are the entropy replicates to measure speech intelligibility.

I agree with Marijn. Readers might also not be that familiar with 'appropriate hierarchies' and entropy

Second, we use Directed Acyclic Graph (DAG) [59, 17] to depict all the relevant variables though to influence speech intelligibility. We describe in detail our causal and non-causal hypothesis, and supplement our description with a causal diagram. The benefit of the method lies, not only, in that it makes the assumptions of our hypothesis more transparent, but also allow us to derive statistical procedures from the aforementioned causal assumptions [53, 79, 62].

As these are key concepts for your research, I would add them in the main text and not as foot notes

[1] ordered on increasing level of ecological validity [32, 25]

[2] the extent to which scores are appropriate for their intended interpretation and use [50, 68]

[3] the extend to which a measure would give us the same result over and over again [68], i.e. measure something, free from error, in a consistent way.

Third and final, we wrap the analysis procedure under the Bayesian framework, providing the assumptions, and the steps required to reproduce the computational implementation of the models.

2. Materials and Methods

What is/are your research question(s)? I would explicitly state them before starting this section. This will help you in explaining the the operationalisation of concepts, analyses, … to the reader.

We set up an experiment where speech samples were transcribed by a group of listeners. The current section succinctly describes the participating children, the stimuli used, and the experimental setup, while also delve into the causal and statistical framework of analysis.

2.1. Children

Thirty two children were selected using a large corpus of *spontaneously spoken speech*, collected by the Computational Linguistics, Psycholinguistics and Sociolinguistics research center (CLiPS). The selection followed a two step procedure [27]. First, a sample of sixteen hearing-impaired children (ten boys, six girls) were selected based on the quality of their registered stimuli. Second, an additional matched sample of sixteen normal hearing children was also selected (six boys, ten girls), and served as a comparison group.

For the first group, all the hearing-impaired children with cochlear implants (HI/CI) were native speakers of Belgian Dutch, living in Flanders, the Dutch speaking area of Belgium. They were all raised orally using monolingual Dutch, with a limited support of signs. All of the children were screened by the Universal Neonatal Hearing Screening (UNHS), using an automated auditory brainstem response hearing tests for newborns, and all received the cochlear implantation before the age of two. Their medical and audiological records did not ascertain any additional health or developmental issues. Hence, no known additional comorbidities were though present. Finally, at the date of the measurement, they were all enrolled in the mainstream educational system.

For the second group, the sixteen normal hearing children (NH) were closely matched to the HI/CI group based on chronological age. All children were also native speakers of Belgian Dutch, and enrolled in the mainstream educational system. None reported hearing loss or additional disabilities, judged from the UNHS screening procedure and their respective parental report.

The characteristics of the selected children is detailed in Table 3, in the supplementary section A.1.

2.2. Stimuli

The stimuli consisted of children's utterances, i.e. sentences of similar length, recovered from previously mentioned CLiPS corpus. More specifically, we use a portion of the corpus that consisted of ten utterances recordings for each of the thirty two selected children, adding to a total of 320 stimuli.

The stimuli were documented when the child was telling a story cued by the picture book 'Frog, where are you' [52] to a caregiver 'who does not know the story'.

The recordings were orthographically transcribed with the CLAN editor in CHAT format [51]. The quality of the stimuli was ensured by selecting utterances with no syntactically ill-formed or incomplete sentences, any background noise, cross-talk, long hesitations, revisions or non-words [5]. The aforementioned transcriptions were used only in the selection process of the stimuli for the experiment.

2.3. Experimental setup

It is not so clear for me why an experimental set-up is needed? What are the experimental/control conditions?

The experiment was setup to perform a transcription task in the Qualtrics environment [77]. One hundred language students from the University of Antwerp participated. The participants were native speakers of Belgian Dutch, without any particular experience with the speech of hearing-impaired children.

The participants and stimuli were divided into five groups, where each group of 20 students transcribed 64 stimuli on their series. The stimuli were presented to the listeners in a random order. As a result, the setup produced 20 transcriptions per utterance, adding to a total of 6400 transcriptions. The steps that comprised the task are detailed in the supplementary section A.2.1.

The data resulting from the transcription task was then processed and converted into one entropy measure per utterance (H), which served as our outcome variable.

I would start with this explanation.

The entropy measure is bounded in the unit interval $[0, 1]$, and it was used as a quantification of (dis)agreement between listeners' transcriptions: utterances yielding a high degree of agreement between transcribers were considered highly intelligible, and therefore, registered a lower entropy ($H \to 0$). In

contrast, utterances yielding a low degree of agreement were considered as exhibiting low intelligibility, and therefore, registered a higher entropy ($H \to 1$) [5, 27]. The procedure followed to calculate the entropies is detailed in the supplementary section A.2.2.

Given that entropy is your outcome variable, I would incorporate the basics of its calculation and meaning here. (I saw that you use examples in the appendix. I think that is a good approach to explain it to your readers.)

2.4. Causal framework

The analysis was informed by a preliminary work aimed at describing the causal and non-causal factors influencing speech intelligibility. More specifically, the current research uses a Directed Acyclic Graph (DAG) [59, 17] to describe all the relevant variables though to influence intelligibility. A DAG is a type of *structural causal model* that can be represented, among other ways, by a *causal diagram*.

Much more explanation is needed on the goal and background of a DAG because it's one of the assets of your study.
Also, the expected relations between the variables (now on page 6 to 8) should be outlined in the introduction (now I understand why these factors were mentioned!) then you can focus here on the DAG itself

Is there a causal relation between hearing status and age?

I would try to use meaningful words in the DAG instead of abbreviations (to decrease the cognitive load on readers :-))
The measures themselves should also already be clear before diving into the DAG.

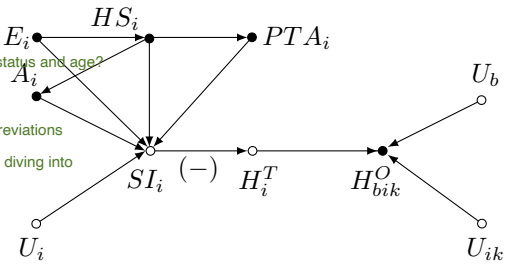

Figure 1: DAG: causal diagram describing the relationships among the analyzed variables. Open circles describe latent/unobserved variables, full circles describe observed variables.

Figure 1 shows the *causal diagram* for our research hypothesis. In the figure, H_{bik}^O denote the *observed* entropy replicates nested within children (Uik) and experimental blocks, (Ub) where $k = 1, \ldots, 10$ utterances, $i = 1, \ldots, 32$ children, and $b = 1, \ldots, 5$ blocks. Moreover, H_i^T and SI_i denotes the child's *true* entropy and speech intelligibility scores, respectively. In addition, HS_i denotes the children's hearing status group, A_i their *hearing age*, PTA_i their unaided post-implant pure tone average, and E_i the etiology of the disease that led to the hearing impairment.

Three main features can be emphasized from the figure. First, the children's *speech intelligibility* and *true* entropy scores are thought to be latent/unobservable variables [26] (drawn with open circles, see supplementary section A.3 about the appropriate interpretations of the scores). The figure also shows the scores are thought to be inferable from the *observed* entropy replicates. More specifically, we are asserting the *observed* entropy replicates H_{bik}^O represent multiple realizations of a child's *true* entropy H_i^T. Finally, as expected from our the theory, we assume the *intelligibility* score SI_i is ~~inversely~~ negatively related to the *true* entropy H_i^T, i.e. the lower the intelligibility the higher the entropy, and vice versa.

I would add that information here

Second, the figure reflects the expected multilevel structure of variability present in our data. This is particularly important as, by failing to account for the appropriate dependencies in the data, we could be 'manufacturing' false confidence in the parameter's estimates, leading us to incorrect inferences [53].

Based on the experimental setup described in section 2.3, we anticipated the ten utterances, originated from each of the thirty two children, were also observed within a group of transcribers (series) assigned to the observation, i.e. our data has a hierarchy with children, replicates and block levels (U_i, U_{ik} and U_b, respectively).

A figure visualising the structure of the data might help in explaining the data structure.

Considering the latter, we expect that if the experiment was 'set up right', the block random effects would explain a small amount of variability in the data, and its inclusion/exclusion in the model would not change the parameter estimates. Moreover, we expect a larger variability between children's *speech intelligibility*, at least larger than the block random effects. Several evidence suggest this is particularly true among HI/CI children [80, 60, 54, 11, 78, 58, 36]. Finally, we did not had any comparable expectation for the variability in the replicates, as this feature has not been investigated before.

What about residual variance? How is that captured in the DAG?

Third, the figure shows the assumed relationship (causal?) among the relevant variables [57, 6, 40, 28], and how these influence the children's intelligibility of speech. Furthermore, it also reveals we assume the variables are independent, beyond the described relationships. Here follows a description of our causal hypothesis for the relevant variables.

For ***hearing status*** (HS_i), it is clear that its inclusion directly corresponds with the main purpose of the current research endeavor. i.e. compare the *speech intelligibility* among NH and HI/CI children. Yet, we do not have a clear expectation about the levels among the groups. In terms of intelligibility, previous literature suggest that some HI/CI children manage to 'catch up' with their NH counterparts

[74, 42, 7, 37, 9, 21, 75], implying the two group attain similar levels. However, other studies also indicate HI/CI children never reach their NH analogues [56, 11, 14, 38, 36, 24, 41], therefore implying different levels of *intelligibility* among the groups.

I would expect more variability within the group of HI/CI children? (But I'm not a linguist of course :-))

For ***Hearing age*** (A_i), we expect the variable to be one of the main responsible for the increase in children's *speech intelligibility*, no mater the children's hearing status [5]. Several studies provide evidence that *intelligibility* increases with age [15, 16, 32, 33, 2, 8, 45], and the intuition of its effects can be easily asserted from Flexer [31]: "the more delayed the age of acquisition of a skill, the farther behind children are in the amount of cumulative practice they have had to perfect that skill. The same concept holds true for cumulative auditory practice".

In that sense, *Hearing age* is a composite variable constructed by combining the *chronological age* for the NH group, and the *device length of use* for the HI/CI group [27] (see table 3 in supplementary section A.1). The variable tries to approximate the amount of time a child has been actively hearing and developing his(her) language. However, no short of evidence has been presented in favor of using others surrogate measures, like *chronological age* [34, 42, 41] or *age at implantation* [57, 6, 9, 21]. We argue that the feasibility of using any other proxy measure, largely depends on the assumed reliability of the surrogate, to approximate the variable of interest. In that sense, although we recognize *hearing age* is not a 'perfect' proxy [27], we argue it is the most appropriate to test our hypothesis, based on the relevant literature and its assumed reliability to capture children's language development (although the latter has not been tested). Furthermore, we believe the variable serve two additional purposes: (i) control for sampling bias (expanded in supplementary section A.4), and (ii) de-confound the parameter estimates of *hearing status* [17].

Sounds like a strength of your study! :-) So I would explain it here instead of the appendix.

What is meant with 'de-confound'?

I think that this part can be left out

Lastly it is important to highlight that, for modeling purposes, using more than one of the aforementioned proxies in tandem is not recommended. It is apparent from the previous description, the three surrogate measures share high similarities in their data construction. This in turn, could cause problems in the modeling procedure, as including variables that provide 'similar information' might lead to a statistical problem known as multicollinearity, in which our estimates get biased and less precise [29], leading us to incorrect inferences and conclusions.

For ***pure tone average*** (PTA_i), we expect it to have a small or null effect on *speech intelligibility*, as the empirical evidence seem to suggest [5]. *Pure tone average* is the child's subjective hearing sensitivity, aided or unaided, by their hearing apparatus.

If pure tone average is unrelated to intelligibility, why did you put it in the model?

I don't understand this part

However, beyond the empirical evidence, the variable was included for two additional reasons. First, given that previous modeling efforts did not capture the full data hierarchy, it is possible that the effects of PTA on *speech intelligibility* has been largely overlooked. Although evidence seem to suggest the variable has no effect, it is also sensible to think that HI/CI children with severe hearing loss, as accounted by the variable, might develop their language at a slower rate. This is especially true, if we consider the signal provided by the cochlear implant is still degraded compared to normal hearing scenarios [22]. Lastly, the second reason for its inclusion lies in the possibility that the variable might be useful to de-confound the parameter estimates of *hearing status* [17], a purpose shared with the *hearing age* variable.

In the case of the ***Etiology*** of the disease that led to the hearing impairment (E_i), we expect it to have a differential effect on *speech intelligibility*, within the HI/CI group. However, since the severity of the etiology cannot be easily ascertain nor ordered, we cannot foresee the direction of such effects, i.e. genetic factors not necessarily lead to worse levels of language development and intelligibility, than factors related to infections.

The previous assumption does not have a correspondence with the empirical evidence, were the variable was deemed unimportant [5]. However, we argue that it is possible its effects have also been largely overlooked, due to the lack of statistical control on the data variability hierarchy, similar to the PTA case. Moreover, as with its predecessors, the variable might also be useful to de-confound the parameter estimates of *hearing status* [17], assuming our DAG is appropriate.

Finally, it is important to highlight the reason for the absence of other variables, deemed relevant by the literature, in our causal hypothesis.

It is uncommon to explain why you didn't include some variables. I think that (if you keep this part in the text) should dig deeper into the reason for their exclusion (frmo the DAG-perspective).

For the ***type of cochlear implantation***, i.e. bilateral or contralateral, the variable was not included because we did not expect it to be related to other variables in the DAG, i.e. the decision on receiving one or the other is solely based on the intelligibility outcome, no matter how the latter is measured. This in turn means that its inclusion/exclusion would not confound our estimates. Additionally, given that most of the children underwent through sequential bilateral implantation (eleven in total), we anticipated the effect of variable already permeates the HI/CI sample.

For the case of ***additional disabilities***, e.g. mental retardation or speech motor problems, given that the children did not reported any additional comorbidities, the variable was deemed unnecessary.

In the case of ***environmental factors***, such as communication modality, the variable was also deemed unnecessary, given that HI/CI children were raised orally using monolingual Dutch with a limited support of signs, a scenario similar to the NH group (see section 2.1).

Last but not least, ***gender*** was not included in our hypothesis, as no theoretical nor empirical evidence have been found on its effects [5].

As expected, it is possible that other unobserved confounding variables are not accounted by our assumptions, and therefore, our causal diagram. This is true for any type of social, behavioral and educational research. However, we argue that the additional transparency of our approach, and its ability to derive statistical procedures from causal assumptions, is its main strength [53, 79, 62].

2.5. Statistical analysis

This is the first time that you talk about the modeling part. I think a bit guidance for the reader might be helpful. (That you will use a Bayesian approach and will fit multiple models.)

Using the DAG, described in the previous section, we describe the algebraic formalism for our multiple Bayesian probabilistic models [46]. Each model targeted a different manner in which our research hypothesis could be investigated.

In general terms, every model was composed of two parts: a latent measurement model [26] and a structural equation model [44]. For the former, we represented the child's *speech intelligibility* as a latent variable, inversely related to the child's latent *true* entropy. Moreover, we modeled the latter as a variable that can be inferred from the observed entropy replicates in the following way:

$$H^O_{bik} \sim \text{BetaProp}\left(\bar{H}_{bi}, M_i\right) \tag{1}$$

$$\bar{H}_{bi} = \text{logit}^{-1}(a_b - SI_i) \tag{2}$$

$$H^T_i = \text{logit}^{-1}(-SI_i) \tag{3}$$

I got a bit lost here :-)

I think more explanation is necessary on this distribution and the reason for using it

where $\text{BetaProp}(\mu, \theta)$ defines the beta-proportion distribution with parameters μ and θ [30, 48]; with $\mu \in [0, 1]$ and $\theta > 0$. Furthermore, $\bar{H}_{bi}$ represented the average entropy measure nested within blocks and children ($b = 1, \ldots, 5$ and $i = 1, \ldots, 32$, respectively), while M_i denotes the 'sample size' of the distribution, possibly nested within children. Furthermore, H^O_{bik}, H^T_i and SI_i denotes the children's entropy replicates, *true* entropy and *speech intelligibility* score, respectively. Finally, a_b represents the block random effects, and $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$ the inverse-logit transformation.

From the previous algebraic structure we can notice. First, we modeled the average entropy of the replicates with $\mu = \bar{H}_{bi}$, which is a non-linear transformation of the block random effects and the children's *speech intelligibility*. Second, the children's *true* entropy H^T_i is inversely and non-linearly related to the *speech intelligibility* SI_i. Third and last, we captured the variability of the entropy replicates using $\theta = M_i$. See supplementary section A.5.1 for a detailed overview of the implications of using this approach.

For the second component we used a structural model [44]. In this portion, different iterations of our research hypothesis were proposed. Hereby we present the general representation, from which the others can be derived:

$$SI_i = a_i + \alpha + \alpha_{E[i],HS[i]} + \beta_{A,HS[i]}(A_i - \bar{A}) + \beta_{P,HS[i]} sPTA_i \tag{4}$$

where E_i and HS_i are defined as in the previous section, A_i is the *hearing age* in years, while $sPTA_i$ is the standardized version of the unaided post-implant pure tone average PTA_i. Moreover, a_i denoted the children's random intercepts, and α the fixed intercept. On the other hand, $\alpha_{E[i],HS[i]}$ denoted the *intelligibility* levels per etiology and hearing status group, while $\beta_{A,HS[i]}$ denoted the evolution of *intelligibility* per unit of *hearing age*, for each hearing status group. Finally, $\beta_{P,HS[i]}$ described the evolution of *intelligibility* per unit of the standardized pure tone average, for each hearing status group.

In the previous algebraic structure it is important to highlight, all parameters are estimated in the logit scale and centered at $sPTA_i = 0$ and $\bar{A} = 5$, the latter denoting the minimum *hearing age* in the sample. This is done to facilitate the interpretation of the parameters.

How are these models related to the DAG? Why these 13 models?

Thirteen statistical models were derived from the previous general description. Each model expressed one specific way in which our research hypothesis could be investigated. Table 1 provides an overview of the parameterization for the full set of proposed models. The models were characterized by two aspects: (i) the use of interactions and the selection of interacting variables, and (ii) the 'sample size' set to capture

the replicates variability. For the former, the options were: intercept only (model 1), and multivariate regression with no interactions (models 2 to 4), with age and hearing status interaction only (models 5, 8, and 11), with etiology and hearing status interaction only (models 6, 9, and 12), and a full interaction model comprising the previous two (models 7, 10 and 13). In a similar fashion, for the latter the options were: one fixed 'sample size' (models 1, 2, 5, 6, and 7), one estimated 'sample size' (models 3, 8, 9, and 10), and finally, one 'sample size' per child (models 4, 11, 12, and 13), which we dubbed *robust* models (see supplementary section A.5.2, where we expand about the need for *robust* models).

Notice all proposals used block and children's random effects (a_b and a_i, respectively), as well as the fixed effect intercept (α). Moreover, the models with one fixed 'sample size' used a value of ten, corresponding to number of utterances per child (see supplementary section A.5.1 on its implications). Finally, all models contemplated the interaction of the standardized pure tone average and the hearing status. The latter was imposed, because the collection of data did not contemplated the measurement of the pure tone average for the NH group. Therefore, the only interpretable parameter is the one estimated for the HI/CI children.

I don't know how Bayesian the linguistic field is right now? But my prior is that Bayesian analysis is not so frequently used/ not known. ;-)
In that case, the two paragraphs below might raise some questions.

It is important to highlight that for all model implementations, the visual inspection of trace, trace-rank and autocorrelation plots was performed. Additionally, we also evaluated the Gelman-Rubin diagnostic and effective number of samples [39]. In all cases, the plots and statistics indicated the parameters achieved convergence, good mixing and lack of serial autocorrelation, all necessary requirements to be able to interpret the model estimates [53].

Additional details concerning the Bayesian implementation of the models, such as the representation of variability in a beta-proportion distribution, the parameters' priors and hyper-priors, the estimation procedure, the data pre-processing phase, and simulation studies are expanded in the supplementary section A.5.

Model	Name	Parameters						
		M_i	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$
1	Intercept only (fixed 'size')	10	a_b	a_i	α	–	–	–
2	No interaction (fixed 'size')	10	a_b	a_i	α	$\alpha_{HS[i]}$	β_A	$\beta_{P,HS[i]}$
3	No interaction (one 'size')	M	a_b	a_i	α	$\alpha_{HS[i]}$	β_A	$\beta_{P,HS[i]}$
4	No interaction (robust)	M_i	a_b	a_i	α	$\alpha_{HS[i]}$	β_A	$\beta_{P,HS[i]}$
5	Age interaction (fixed 'size')	10	a_b	a_i	α	$\alpha_{HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$
6	Etiology interaction (fixed 'size')	10	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	β_A	$\beta_{P,HS[i]}$
7	Full interaction (fixed 'size')	10	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$
8	Age interaction (one 'size')	M	a_b	a_i	α	$\alpha_{HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$
9	Etiology interaction (one 'size')	M	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	β_A	$\beta_{P,HS[i]}$
10	Full interaction (one 'size')	M	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$
11	Age interaction (robust)	M_i	a_b	a_i	α	$\alpha_{HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$
12	Etiology interaction (robust)	M_i	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	β_A	$\beta_{P,HS[i]}$
13	Full interaction (robust)	M_i	a_b	a_i	α	$\alpha_{E[i],HS[i]}$	$\beta_{A,HS[i]}$	$\beta_{P,HS[i]}$

Table 1: Proposed statistical models. Parameterizations.

3. Results

3.1. About the variability of the data

As expected, our data registered three levels of variability: between blocks and children variability, and finally, the entropy replicates variability.

Median or average of posterior distribution?

Evidence from the posterior estimates reveal the block random effects explained a small amount of variability in the data ($\sigma_b = 0.26$, top panel of figure 2), and its inclusion/exclusion in the model did not change the parameter estimates. The previous implied the experiment was correctly 'set up', as the series in which the utterances were transcribed did not explain a significant amount of variation, nor its exclusion biased the parameter estimates.

On the contrary, we observe a significantly larger variability between children. More precisely, the *speech intelligibility* variability between children was more than three times the block effects variability

0.74 vs 0.26?

($\sigma_i = 0.74$, middle panel of figure 2). The result corroborates previous evidence on the matter [80, 60, 54, 11, 78, 58, 36, 5].

I would move this to the discussion and stick here to describing the results

Lastly, for the entropy replicates variability, the posterior estimates reveal a reasonable finding: the amount of variability at the replicates level is even larger than the one observed at the children level ($M = 6$, bottom panel of figure 2). The latter implies there is significant error in measuring *speech intelligibility* using the entropy replicates, i.e. there is considerable measurement error [10].

The last two results have important consequences for our future inferences. What the results imply is that given the large amount of variability between children and the considerable presence of measurement error, the statistical models might have a harder time producing unequivocal inferences, in respect to the intelligibility levels of the *hearing status* groups, or other parameters of interest. This is particularly important, as simulation studies revealed this behavior was expected even with more reasonable levels of variability ($\sigma_i = 0.5$ and $M = 10$, see supplementary section A.5.5).

Difficult for a reader who didn't take the time to look at the results of the simulation studies in the supplementary section.

Why is represented by the colors?
The difference in scale of the y-axis makes it hard to compare the results across blocks, children and replicates

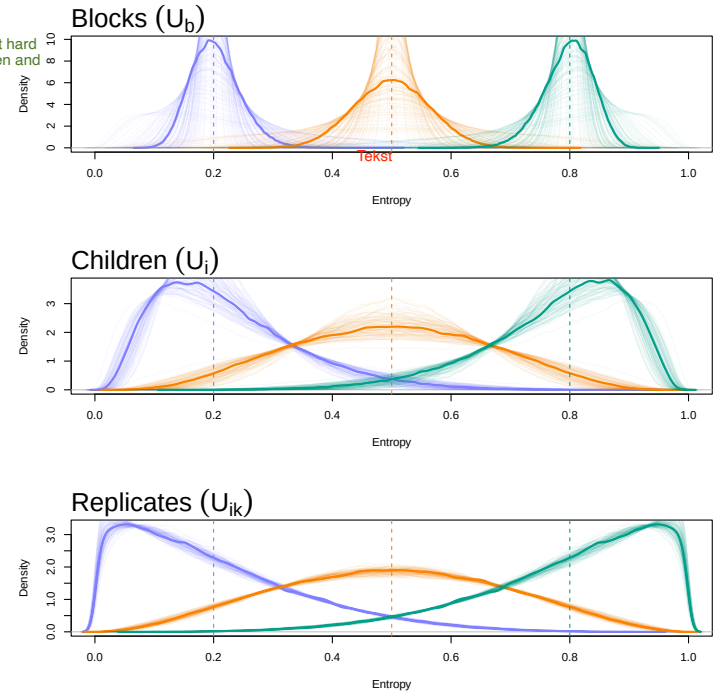

Figure 2: Model 10, posterior predictive: variability present in the data. Distributions are plotted on the entropy replicates scale, considering three different averages: $\mu = 0.2$, $\mu = 0.5$, and $\mu = 0.8$ (discontinuous lines). Thick solid lines represent the marginal distribution, thin solid lines depict 100 random posterior samples.

3.2. About our hypothesis

I think an explanation of the IT approach should be already provided in the methods section (when you explain the 13 models that will be fitted.)

The current research used the Information-Theoretic Approach [1, 12] for model selection and inference. The application of the approach required: (i) the expression of the research hypothesis into statistical models, (ii) the selection of the most plausible models, and (iii) to produce inferences based on one or multiple selected models. The first requirement of the approach is covered in sections 2.4 and 2.5, and expanded in supplementary section A.5. Here we use the results of the second requirement, detailed in the supplementary section A.5.6, to produce the final inferences.

I think step 2 should be explained here as well. (Selecting mode(ls) is kind of crucial for making inferences :))

As detailed in the supplementary section, the final conclusions of our research will be drawn from the comparisons of two models: (i) a no interaction model, with one estimated 'sample size' (model 3), and (ii) a full interaction model, with one estimated 'sample size' (model 10). The former is selected because it is the model with highest probabilistic support. The latter is considered because it encompasses the remaining highest supported models. Furthermore, no *robust* models are inspected, as we prefer a more parsimonious depiction of our hypothesis. Table 2 summarizes the parameters posterior estimates and contrasts of interest.

If so, why did you estimate these models?

Parameter	Mean	SD	CI 2.5%	CI 97.5%	HPDI 2.5%	HPDI 97.5%	n eff.	Rhat
Model 3: No interaction (one 'size')								
parameters:								
a	0.179	0.189	-0.197	0.554	-0.198	0.552	3'677.705	1.001
bP[2]	-0.117	0.166	-0.439	0.207	-0.440	0.202	1'738.855	1.000
bA	0.432	0.141	0.154	0.715	0.170	0.723	1'815.243	1.001
aHS[1]	0.284	0.235	-0.178	0.761	-0.203	0.728	2'719.833	1.000
aHS[2]	0.116	0.217	-0.303	0.537	-0.304	0.537	2'646.671	1.000
constrast:								
aHS[2]-aHS[1]	-0.168	0.246	-0.661	0.371	-0.650	0.324	n.a.	n.a.
Model 10: Full interaction (one 'size')								
parameters:								
a	0.217	0.179	-0.142	0.562	-0.118	0.580	3'902.629	0.999
bP[2]	-0.122	0.173	-0.457	0.220	-0.460	0.216	1'659.639	1.002
bAHS[1]	0.435	0.157	0.127	0.745	0.123	0.741	1'477.460	1.000
bAHS[2]	0.237	0.177	-0.119	0.596	-0.089	0.615	1'844.785	1.001
aEHS[1,1]	0.183	0.278	-0.358	0.726	-0.349	0.729	2'524.451	1.000
aEHS[2,2]	0.212	0.241	-0.259	0.684	-0.259	0.684	2'418.275	0.999
aEHS[3,2]	0.077	0.245	-0.402	0.547	-0.383	0.561	3'015.559	0.999
aEHS[4,2]	0.007	0.269	-0.522	0.530	-0.547	0.499	3'268.043	1.000
contrast:								
bAHS[2]-bAHS[1]	-0.197	0.208	-0.600	0.213	-0.590	0.223	n.a.	n.a.
aEHS[2,2]-aEHS[1,1]	0.029	0.352	-0.673	0.723	-0.714	0.682	n.a.	n.a.
aEHS[3,2]-aEHS[1,1]	-0.106	0.365	-0.823	0.607	-0.822	0.609	n.a.	n.a.
aEHS[4,2]-aEHS[1,1]	-0.176	0.380	-0.906	0.568	-0.928	0.535	n.a.	n.a.
aEHS[3,2]-aEHS[2,2]	-0.135	0.299	-0.719	0.453	-0.730	0.439	n.a.	n.a.
aEHS[4,2]-aEHS[2,2]	-0.205	0.344	-0.888	0.453	-0.902	0.430	n.a.	n.a.
aEHS[4,2]-aEHS[3,2]	-0.070	0.344	-0.744	0.612	-0.735	0.617	n.a.	n.a.

CI = compatibility interval
HDPI = highest posterior density interval
n eff. = effective number samples
Rhat = Gelman-Rubin diagnostic
n.a. = not available / not applicable
[1] = NH children, [2] = HI/CI children
[1,i] = NH children, [2,i] = genetic, [3,i] = CMV infection, [4,i] = unknown etiology

Table 2: Selected statistical models: results.

Before providing any parameter interpretation, it is important to highlight an statistical issue that will permeate all of our inferences. As it is detailed in section 3.1, with our current data size and given the large amount of variability registered at the children and replicates levels, the models are not able to produce unequivocal null hypothesis rejections for most of the parameters estimates, at a confidence level of 95%, i.e. the CI and HPDI will almost always include the zero value. This is particularly important, as rejecting the null hypothesis of certain parameters and contrast of interest provides unambiguous evidence for the current research hypothesis.

However, we consider that even when particular null hypothesis cannot be rejected, the models are still able show some preliminary evidence of effects and their direction. Given that parameter estimation under the Bayesian framework does not return parameters' point estimates, but rather the posterior distribution of possible values, we can still evaluate the amount of evidence towards a specific effect and its direction. In a complementary way, supplementary section A.5.5 reveal that our models are also able to affirm the estimate values with at least 60% power, depending on the estimates magnitudes, i.e. the models can correctly estimate values different from zero, when such alternative hypothesis is true.

In that sense, we will continue interpreting the parameters, providing evidence on the directionality of

the effects. Nevertheless, the reader should be cautious into understanding the CI and HPDI indicate that our data size might still not be enough, to ultimately define the direction of some effects with a certain level of confidence (usually 95%), and further observations at the children level might still be needed.

Considering the previous, for ***hearing status***, model three reveals that HI/CI children have a modest lower level of *speech intelligibility*, compared to their NH counterparts (`aHS[2]-aHS[1]`). Notice the CI and HPDI include the value of zero, and therefore, null hypothesis rejections at a 95% confidence level cannot be made. However, the posterior distribution of the contrast reveal that 76.8% of the estimates are unequivocally below zero (`aHS[2]-aHS[1]` < 0). In this sense, the result seem to mildly support previous evidence on the matter [56, 11, 14, 38, 36, 24, 41].

Nevertheless, model ten paints a more nuanced story considering the ***etiology*** of the disease. The contrasts of the model reveal that HI/CI children with genetic etiology manage to reach similar levels of *intelligibility* as NH children (`aEHS[2,2]-aEHS[1,1]`). However, the same cannot be said when the hearing impairment is caused either by CMV infection (`aEHS[3,2]-aEHS[1,1]`) or other unknown causes (`aEHS[4,2]-aEHS[1,1]`). The posterior distribution of the aforementioned contrasts show the estimates are distinctly below zero with a probability of 62.4% and 68%, respectively.

About the former result, we theorize a intuitive explanation. Since hearing loss is a brain issue, rather than ear issue [31], as children with genetic causes grow developing their language using the hearing apparatus, they grow constructing effective auditory neural pathways that benefit from the new hearing input from the start, even when that input is slightly degraded [22]. However, the same cannot be said for other etiology status. What is hinted is that, a child with other etiology might need to 're-wire' his(her) auditory neural connections in order to fully take advantage of the new hearing signal. This is particularly important, if we consider these children are trying to achieve this 're-wiring' process, during an important maturing stage of the brain cortex (first twelve months) [31].

In relation to ***hearing age***, model three reveals what we initially expected: the factor is one of the main determinants for the increase in children's *speech intelligibility*, showing statistically significant moderate effects (`bA`) [18, 64]. However, model ten offers preliminary evidence contrary to what was previously found [5], indicating there is a difference in the evolution of *intelligibility* between HI/CI and NH children (`bAHS[2]-bAHS[1]`). The posterior distribution of the contrast show 83.1% of the estimates are unequivocally below zero. We intuit the latter result have two possible complementary explanations, either: (i) the children develop their language differently at different stages of their (*hearing*) age [31], or (ii) given that the hearing signal provided by the apparatus is degraded [22], HI/CI children might take longer to achieve similar levels of *intelligibility* than their NH counterparts.

Finally, for ***pure tone average***, we see both models support our initial hypothesis: HI/CI children with severe hearing loss, as accounted by the variable, develop their language at a slower rate than their NH counterparts (`bP[2]`). Notice again that 95% confidence null hypothesis rejections cannot be made. However, the posterior distribution of the parameter reveal the estimates are unequivocally below zero with approximately 76% probability (`bP[2]` < 0). Therefore, from perspective of effect sizes [18, 64], the estimates can be considered small in magnitude, but present nonetheless.

More interestingly still is that this results couple rather well with the second explanation for the effects of *hearing age*, i.e. the more degraded the signal a child receives, either because of the apparatus or the severity of their hearing-impairment, the slower the *speech intelligibility* evolves.

3.3. The speech intelligibility scale

As previously stated, one of the main benefits of the current methodology is that we can 'construct' an *intelligibility* score of the sampled children, which in turn, allow us to test different hypothesis and even make individual comparisons at the appropriate level. See supplementary section A.3 for the appropriate interpretation of the scores.

Figure 3 depicts the estimated *true* entropy and *speech intelligibility* scores per child. Three points can be highlighted from the figure. First, from the horizontal discontinuous lines, we can clearly assess that on average NH children have a modest higher level of intelligibility (lower entropy) than the HI/CI counterparts. However, given the children's variability, we cannot reject the hypothesis that HI/CI have similar levels than the NH children, as one can notice from the overlapping shaded areas of both groups.

Second, the figure reveals there is some inherited uncertainty on the score estimates, resulting from using entropy replicates as a tool for measuring *speech intelligibility*. The latter can be observed from the vertical lines, describing the highest posterior density intervals for the *intelligibility* of each child. We

Tekst

argue this last trait is one of the main strengths of the method, as it allow us to not only make individual comparisons, but also consider the inherited (un)certainty of the comparison.

Third and final, the non-linear relationship between *speech intelligibility* and the *true* entropy scale is clearly noticed in the HPD intervals, where lowest/highest entropy values are paired with narrower intervals. Clear examples of this can be observed at the top panel of figure 3, where children 20 and 25 show narrower intervals at the lowest part of the scale, while child 6 shows the analogous at the highest end of the same. The previous behavior is expected under non-linear statistical models. Nevertheless, this particular trait has an intuitive interpretation under our specific application, i.e. we can be more certain of the children's *true* entropy values at the extreme of the hypothetical construct (either zero or one), rather than in the middle part of the scale.

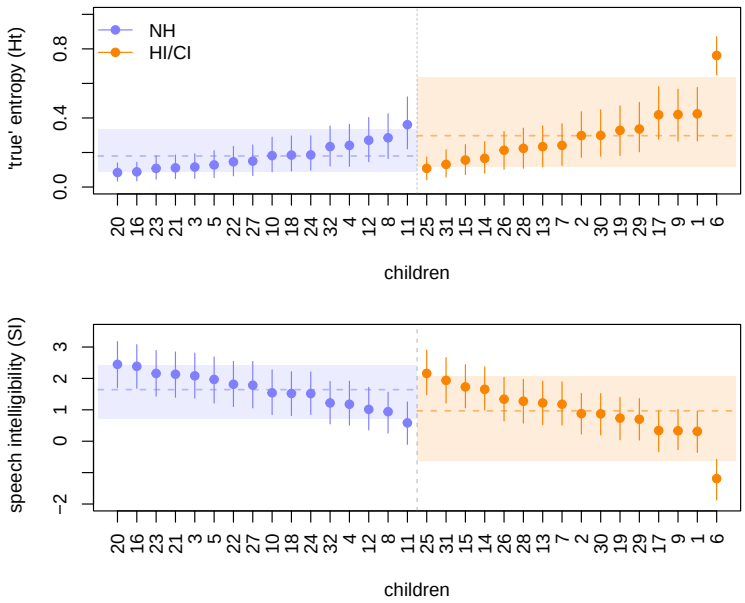

Figure 3: Model 10, posterior predictive: *true* entropy and *speech intelligibility* scales. Colored circles represent mean values per *hearing status* group. Thin vertical continuous lines depict 95% highest posterior density intervals (HDPI). Thin horizontal discontinuous lines represent the marginal average for the *hearing status* group. Shaded area represent the marginal averages 95% confidence interval.

3.4. Posterior predictive

It's a bit strange that this is reported after the actual results?
(Imagine that these checks would be bad. You should first adjust and refit your models)

The posterior predictive results give us with a tool to assess how good our models were to reconstruct the observed data. Figure 4 shows our full interaction model managed to 'recreate' our data, while also provided the (un)certainty around the entropy replicates. The latter was shown in two ways: (i) through the point estimates and HPD intervals for the *true* entropy scale (red points and horizontal lines), and (ii) though the expected estimated marginal and sampled distributions for the entropy replicates (thick and thin solid distribution lines, respectively).

3.5. Influential observations

This part is a bit unexpected for me because it doesn't relate to your research goals and isn't mentioned in the methods section.

Finally, implementing the statistical models under the Bayesian framework provide us with one last advantage: we were able to identify influential observations that might sway our estimates, under our model framework [53]. We argue this method is preferable, as the identification of influential observations through preliminary or univariate procedures might lead to erroneous exclusion of information, ultimately damaging our research inferences. See supplementary section A.5.2 for more details.

Figure 5 provides an overview of the identified high-leverage observations (pair child, utterance). Model ten reveals some observations for children 8, 10, and 15 can be deemed prominent. The first two were NH children, while the last was an HI/CI analogue.

move this to appendix

After a careful inspection, the only similitude between these observations was that all of them registered the lowest replicate values observed in the sample (0.007 for child 8, and zero for the remaining two). Therefore, it makes sense our model found these points 'surprising', as they are mostly out of the expected range for the outcome. The latter is particularly true for child 15, as it is not expected that an HI/CI child have traits of perfect *intelligibility*, as the replicate seem to indicate.

However, notice additionally that even with their presence, no further evidence was found in favor of *robust* models over other more parsimonious iterations of our research hypothesis (see supplementary section A.5.6). As a result, none of these observation were excluded from the analysis, but they were identified and can be further investigated.

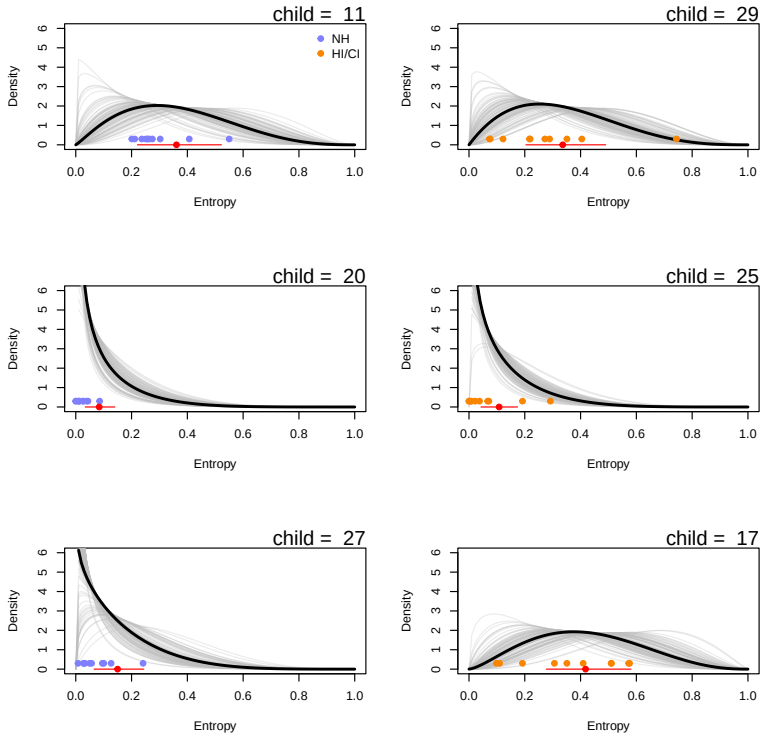

Figure 4: Model 10, posterior predictive: entropy replicates, *true* entropy, and distributions. Colored points represent the entropy replicates per *hearing status* group. Red points with lines represent the mean *true* entropy with 95% highest posterior density interval (HPDI). Thick solid line represents the marginal distribution, thin solid lines depicts 100 random posterior samples for the distributions.

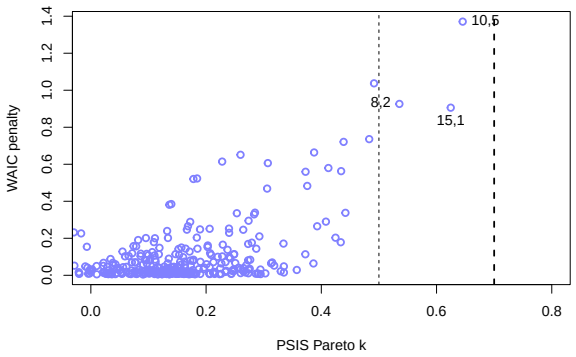

Figure 5: Model 10, influential observations. Pairs child, utterance are reported for specific observations. Thin and thick discontinuous lines indicate the minimum and extreme recommended threshold of identification (0.5 and 0.7, respectively).

4. Discussion

Here you give a nice overview of your research goal (and its advantages/ strengths over the usual approach used)

As stated in previous paragraphs, the main focus of the present research was to investigate the speech intelligibility levels of normal hearing (NH) versus hearing-impaired children with cochlear implants (HI/CI). For that purpose, we set up an experiment where ten utterances, from thirty two children, were transcribed and transformed into entropy measures, which served as our outcome variable.

Rather than use the common multilevel approach, the current research proposed a statistical method novel to the field of linguistics. First, we modeled the *speech intelligibility* as a latent variable that can be inferred from the entropy replicates [26, 44]. The previous not only allow us to 'construct' an *intelligibility* score, but also to control for all the variability assumed present in our data. We argue the former was important as it allowed us to make comparisons at appropriate levels, while the latter was important to not 'manufacture' false confidence in our parameter estimates, ultimately leading us to incorrect inferences [53]. Second, we used Directed Acyclic Graphs (DAG) [59, 17] to describe our causal and non-causal hypothesis about the *intelligibility* of speech, and determine which variables were relevant for its development. Finally, we implemented the method under the Bayesian framework, where we provided the steps and code for its implementation.

These are two other methodological improvements of your research?

I think it would help if you framed it like this from the start of the article.

The statistical results highlighted two specific points. First, and most important, considering the large variability registered at the children and replicates levels, we were not able to produce unequivocal null hypothesis rejections, at a 95% confidence level. As a result, in some way, we were impeded to clearly delimit the magnitude and direction of some effects. Furthermore, simulation studies revealed the need for larger sample sizes in order to produce such decisive results.

These are only reported on the supplementary sections.

Nevertheless, as a second point, we have shown that by using the posterior distribution of the estimates, we were able to provide 'preliminary' evidence on some of our causal hypothesis. On the one hand, we found there was mild evidence towards a difference in *intelligibility* between *hearing status* groups, in favor of NH children. However, the evidence also revealed the difference was mostly observed between NH and HI/CI children with etiologies other than genetic. In that sense, we further hypothesized that since hearing loss is a brain issue, rather than ear issue [31], these children might be 're-wiring' their auditory neural pathways during an important maturing stage of the brain cortex (first twelve months) [31], and that is what causes the delay in language development. It is fair to say then, the researchers believe it would be interesting to further inspect this hypothesis, with the support of Magnetic Resonance Imaging (MRI).

For a reader this might not be so obvious (because the CI and HDPI reported in the table are similar to a 95CI in an frequentist table).

On the other hand, our analysis clearly identified *hearing age* as one of the main determinant in the evolution of *speech intelligibility*. However, the story produced by our models was more nuanced, as they also indicated there might be a different evolution in *intelligibility* between NH and HI/CI children. The last result gave evidence contrary to what was previously found [5]. Furthermore, as in the case of *hearing status*, we hypothesized this outcome have two possible complementary explanations, either: (i) the children develop their language differently, at different stages of their (*hearing*) age [31], or (ii) given that the hearing signal provided by the apparatus is degraded [22], HI/CI children might take longer to achieve similar levels of *intelligibility* than their NH counterparts. Therefore, the researchers believe it would be interesting to assess the latter hypothesis, and for that purpose, further collection of data would be required, i.e we would need to identify the hearing signal degradation for the cochlear implants.

Furthermore, our analysis provided 'mild' evidence that HI/CI children with more severe hearing loss, as accounted by the *pure tone average*, develop their language at slower rate than their NH analogues. We further noticed, the previous result couples rather well with the second hypothesis proposed to explain the effects of *hearing age*, i.e. the more degraded the hearing signal, either because of the apparatus or the severity of the hearing-impairment, the slower the *intelligibility* evolves.

As perspectives for future research and further critiques, two finals points can be highlighted. First, the procedure for calculating the entropy replicates weighted equally any differences in the transcriptions. This implied the entropy values were equally increased by fairly small deviances, e.g. kikker '*frog*' versus kikkers '*frogs*', than with considerable ones, e.g. mismatches as in jongen '*boy*' versus hond '*dog*'. In that sense, the researchers also share the believe of Boonen et al. [5], that a further refinement of the entropy calculation can be made, in which the linguistic distance of the transcriptions can be considered.

Second and last, although we used a DAG to state our causal and non-causal hypothesis, it is clear, the benefits of the tool shine when it is also used to plan the data collection. As stated in the document, the tool not only helps to make our hypothesis more transparent, but also allow us to derive statistical procedures from such assumptions. Scenarios like the lack of power to test specific effects, and the need for larger samples could have been easily foresaw during this part of the experimental planning [53, 35].

5. Acknowledgments

A special thanks to Richard McElreath, for timely response and help. Further gratitude is expressed to Tine van Daal, Marijn Gijsen, Lies Appels, and Silvana Romero for giving feedback to preliminary versions of this document.

6. Author contributions

All authors contributed to the development of the causal hypothesis. Jose Rivera performed the statistical analysis, Sven de Maeyer supervised the production of the documents and statistical results, and Steven Gillis collected the data.

7. Financial support

The project was financed by the Flemish Government through the University Research Fund (BOF).

8. Conflicts of interest

The authors declare they have no conflict of interest.

9. Research transparency and reproducibility

The model simulation procedures and testing that support the findings of this study are openly available at `https://github.com/jriveraespejo/PhD_UA_paper1`.

Due to the privacy and confidentiality of subjects, the data set in which the model was implemented cannot be put online.

A. Supplementary

A.1. Children characteristics

Table 3 shows the detailed information of the sampled children. The referred table includes the variable used for the matching procedure, i.e. chronological age, while also additional variables thought to be relevant for our hypothesis. No other variables are included as no known additional comorbidities, beside their hearing impairment, are suspected. Additionally, notice the pure tone average (PTA) data was not collected for the NH children.

Child	Gender	Chronological age (y;m)	Device length of use (y;m)	Hearing age (y;m)	Etiology	PTA (dB.)	
						unaided	aided
	HI/CI children						
1	female	05;07	05;00	05;00	Genetic	120	19
2	male	06;04	05;09	05;09	CMV	106	23
3	male	06;07	05;10	05;10	Genetic	114	35
4	female	06;10	06;00	06;00	Unknown	120	20
5	female	07;00	06;03	06;03	CMV	115	25
6	male	07;00	05;08	05;08	Genetic	93	32
7	female	07;00	06;08	06;08	Genetic	117	17
8	female	07;00	05;05	05;05	Unknown	112	42
9	male	07;00	05;05	05;05	CMV	120	15
10	female	07;01	05;11	05;11	Genetic	120	35
11	male	07;01	05;07	05;07	Genetic	113	42
12	male	07;02	06;05	06;05	Genetic	120	37
13	male	07;08	06;10	06;10	CMV	114	27
14	male	07;09	06;02	06;02	CMV	120	35
15	male	08;07	07;10	07;10	CMV	120	33
16	male	08;08	09;09	09;09	Genetic	95	27
	NH children						
17	female	06;05	n.a.	06;05	n.a.	n.a.	n.a.
18	female	06;06	n.a.	06;06	n.a.	n.a.	n.a.
19	female	06;07	n.a.	06;07	n.a.	n.a.	n.a.
20	female	06;09	n.a.	06;09	n.a.	n.a.	n.a.
21	female	06;09	n.a.	06;09	n.a.	n.a.	n.a.
22	male	06;09	n.a.	06;09	n.a.	n.a.	n.a.
23	male	06;09	n.a.	06;09	n.a.	n.a.	n.a.
24	male	06;10	n.a.	06;10	n.a.	n.a.	n.a.
25	female	07;01	n.a.	07;01	n.a.	n.a.	n.a.
26	male	07;01	n.a.	07;01	n.a.	n.a.	n.a.
27	male	07;04	n.a.	07;04	n.a.	n.a.	n.a.
28	female	07;08	n.a.	07;08	n.a.	n.a.	n.a.
29	male	07;08	n.a.	07;08	n.a.	n.a.	n.a.
30	female	07;09	n.a.	07;09	n.a.	n.a.	n.a.
31	female	08;00	n.a.	08;00	n.a.	n.a.	n.a.
32	female	08;01	n.a.	08;01	n.a.	n.a.	n.a.

(y;m) = (years;months)

n.a. = not applicable / not available

Table 3: Characteristics of selected children.

A.2. Experiment details

A.2.1. Transcription task

The setting for the transcription task comprised the following steps [4, 5]:

1. the listener took a seat in front of a computer screen, located at the campus' computer laboratory.
2. the listener opened Qualtrics [77] and select the transcription task.
3. the listener read two set of instructions presented on the computer screen about:
   a) *how to perform the task*,
   b) *the aspects not considered for the task*.
4. the listener hear the stimuli through high quality headphones, set at a comfortable volume.
5. the listener wrote the orthographic transcriptions of the utterances, in a free text field in the software environment.

A.2.2. Entropy calculation

The outcome from the transcription task was obtained following a two step procedure [5]. First, we aligned the participant's orthographic transcriptions, at the utterance level, in a column-like grid structure similar to the one presented in Table 4. This step was repeated for every one of the 6400 transcriptions. Lastly, we computed the entropy measure of the aligned transcriptions as in Shannon [65]:

$$H = H(\boldsymbol{p}) = \frac{-\sum_{i=1}^{n} p_i \cdot \log_2(p_i)}{\log_2(N)} \tag{5}$$

where H is bounded in the unit interval $[0, 1]$, n denotes the number of word occurrences within each utterance, p_i the probability of such word occurrence, and N the total number of aligned transcriptions per utterance.

Transcription number	Utterance				
	1	2	3	4	5
1	de	jongen	ziet	een	kikker
	the	boy	see	a	frog
2	de	jongen	ziet	de	[X]
	the	boy	sees	the	[X]
3	de	jongen	zag	[B]	kokkin
	the	boy	saw	[B]	cook
4	de	jongen	zag	geen	kikkers
	the	boy	saw	no	frogs
5	de	hond	zoekt	een	[X]
	the	dog	searches	a	[X]
Entropy	0	0.3109	0.6555	0.8277	1

[B] = blank space, [X] = unidentifiable word

Table 4: Alignment and entropy calculation. Extracted from Boonen et al. [5], and slightly modified with illustrative purposes.

Entropy was used as a quantification of (dis)agreement between listeners' transcriptions, i.e. utterances yielding a high degree of agreement between transcribers were considered highly intelligible, and therefore registered a lower entropy ($H \to 0$). In contrast, utterances yielding a low degree of agreement were considered as exhibiting low intelligibility, and therefore registered a higher entropy ($H \to 1$) [5, 27].

To exemplify relevant scenarios for the procedure, we generate the entropy for utterances 2, 4 and 5 present in Table 4. To make the example easy to calculate, we assume our data consisted only of five transcriptions in total ($N = 5$).

For the second utterance, we observe that four transcriptions identify it with the word *jongen*, while the last with the word *hond*. Therefore, we registered two word occurrences ($n = 2$), with probabilities $\boldsymbol{p} = (p_1, p_2) = (4/5, 1/5)$, and entropy measure equal to:

$$\begin{aligned} H &= \frac{-\sum_{i=1}^{2} p_i \cdot \log_2(p_i)}{\log_2(5)} \\ &= \frac{-\left[0.8 \log_2(0.8) + 0.2 \log_2(0.2)\right]}{\log_2(5)} \\ &\approx 0.3109 \end{aligned}$$

For the fourth utterance, we observe that two transcriptions identify it with the word *een*, one with *de*, one with *geen*, and one with a blank space [B]. Notice the blank space was not expected in such position, therefore, it was considered as a different word occurrence. As a result, the scenario had four word occurrences ($n = 4$), with probabilities $\boldsymbol{p} = (p_1, p_2, p_3, p_4) = (2/5, 1/5, 1/5, 1/5)$, and entropy measure equal to:

$$\begin{aligned} H &= \frac{-\sum_{i=1}^{4} p_i \cdot \log_2(p_i)}{\log_2(5)} \\ &= \frac{-\left[0.4 \log_2(0.4) + 3 \cdot 0.2 \log_2(0.2)\right]}{\log_2(5)} \\ &\approx 0.8277 \end{aligned}$$

Finally, for the fifth utterance, we observe that all transcriptions identify it with different words. Notice that when a transcriber did not managed to identify (part of) the complete utterance, (s)he was instructed to write [X] to replace it. However, if more than one transcriber used [X] for an unidentifiable word, each was considered as being different from one another. The latter is done to avoid the artificial reduction of the entropy measure, as [X] values already indicate the lack of intelligibility of the word. Therefore, for the fifth utterance we registered five word occurrences ($n = 5$), with probabilities $\boldsymbol{p} = (p_1, \ldots, p_5) = (1/5, \ldots, 1/5)$, and entropy measure equal to:

$$\begin{aligned} H &= \frac{-\sum_{i=1}^{5} p_i \cdot \log_2(p_i)}{\log_2(5)} \\ &= \frac{-5 \cdot 0.2 \log_2(0.2)}{\log_2(5)} \\ &= 1 \end{aligned}$$

A.3. About speech intelligibility

As described in the introduction, intelligible speech can be defined as the extent to which the elements in an speaker's acoustic signal, e.g. phonemes or words, can be correctly recovered by a listener [47, 73, 70, 36]. More specifically, in the context of the transcription task, speech intelligibility can be inferred from the extent a set of transcribers can identify the word contained in an utterance [5].

Therefore in this paper, through the implementation of our proposed model, *speech intelligibility* is interpreted as a latent trait of individuals (hypothetical construct), which underlies the probability of observing a set of entropy replicates; that in turns, describes the ability of transcribers to identify the words in an utterance. Henceforth, statements such *'speech intelligibility is influenced by'* can be read as *'the probability of observing a set of entropy replicates for each individual in the sample is influenced by'*. Similar interpretation can be extended to the (latent) *true* entropy measures.

Despite this practical approach, we emphasize we did our best to ensure the construct validity of our study: (i) by ensuring the transcription task was well understood, and appropriately performed by the transcribers, and (ii) by using children's utterances coming from spontaneous speech, a task with the highest level of ecological validity [32, 25].

We then expect speech intelligibility, as measured by our model, to reflect the (general) *intelligibility* of speech possessed by individuals, but do not deal with general epistemological considerations on the connection between the two.

A.4. Sampling bias

As it happens in most observational, and some experimental studies, ours can also be a potential victim of sampling bias. While stratifying on the selection variables can help to balance the samples, and even 'correct' the estimates [17, 20], as we do here by controlling for *hearing age*. Given the sample's selection and matching procedures, we cannot ensure the HI/CI nor the NH groups are representative of their respective populations.

Nevertheless, we argue that by controlling for other relevant confounders, the qualitative results presented in this study holds. However, we cannot discard the presence of unobservable variables that could bias our results, and in that sense, inferences beyond this particular set of children must be taken with care.

A.5. Model details

A.5.1. Variability in the beta-proportion distribution

Figure 6 shows the implications of different 'sample sizes' (M_i) on the dispersion of the beta-proportion distribution [48]. The panels show different average entropies: the middle panel assumes $\mu = 0.5$, the left $\mu = 0.2$, and the right $\mu = 0.8$ (as shown by the discontinuous lines).

In all three panels we notice two prevalent pattern: (i) the higher the 'sample size', the less dispersed is the distribution, and (ii) as expected from non-linear models, the behavior of the dispersion depends on the location of the distribution, i.e. their average value.

Concerning the first patterns, we expect that if the posterior estimates for M_i reach lower values, it would imply the entropy replicates H^O_{bik} had high dispersion. In contrast, higher M_i estimates would imply a lower dispersion in the replicates. In this particular case, a good comparison point is the number of utterances/replicates. If $M_i < 10$, it would imply that our effective 'sample size' (measuring points) is less than the actual number of replicates, and therefore, we would need more replicates to provide an accurate measure of the (latent) *true* entropy. In contrast, $M_i > 10$ would imply the opposite.

As a final point, it is important to highlight that as the approach uses the 'sample size' parameter to model the replicates' heterogeneity, we are effectively estimating a measurement error model [10] on entropy values.

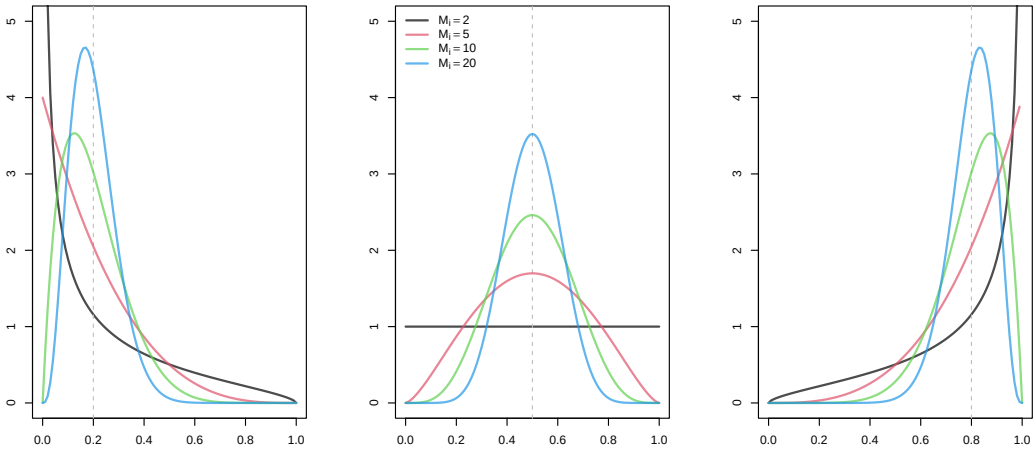

Figure 6: Variability in a beta-proportional distribution. Discontinuous lines describe the average value for the distribution (μ), solid lines describe the distribution assuming different $\theta = M_i$.

A.5.2. Data pre-processing

Besides the exclusion of corrupted observations, e.g. no available transcription, no other information was excluded before the modeling process. This decision departs from what has been done in previous research

[4, 69, 5]. The reason is that we believe the identification of influential observations, through preliminary or univariate procedures, might lead to erroneous exclusion of information, ultimately biasing our results. We believe the identification of *outliers* should not be done outside the context of a full modeling effort [53], as what can behave as an *outlier* based on a univariate analysis, might behave as expected under the appropriate model.

Considering the previous, instead of eliminating information based on preliminary analysis, we proposed a set of models that can be qualified as *robust* against influential observations. Considering the flexibility of the Bayesian framework, the proposal of such models was a trivial task (see Table 1).

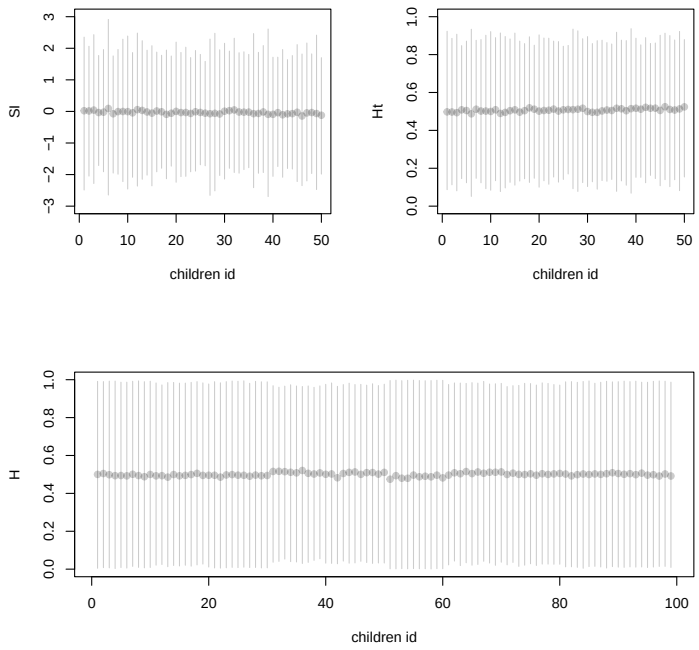

Figure 7: Prior distribution implications: speech intelligibility, *true* entropy and observed entropy scales. Circles represent mean values, lines depict the 95% compatibility intervals.

A.5.3. Priors and hyper-priors

For all models, the selection of priors and hyper-priors was done through prior predictive simulation [53]. The selected priors were considered mildly informative and regularizing.

Figure 7 shows the implication of our assumptions on the three outcome scales of interest: the *speech intelligibility*, the *true* entropy, and the *observed* entropy replicates. Notice, no undesirable assumption has crept in any of the scales. Consequently, the estimates are free to visit a wide range of results, while also establishing a low probability of reaching impossible outcomes.

Hereby follows a description of the priors for the parameters already defined in section 2.5:

$$M_i \sim \text{Log-Normal}(\mu_M, \sigma_M) \tag{6}$$
$$a_b \sim \text{Normal}(\mu_b, \sigma_b) \tag{7}$$
$$a_i \sim \text{Normal}(\mu_i, \sigma_i) \tag{8}$$
$$\alpha \sim \text{Normal}(0, 0.2) \tag{9}$$
$$\alpha_{E[i],HS[i]} \sim \text{Normal}(0, 0.3) \tag{10}$$
$$\beta_{A,HS[i]} \sim \text{Normal}(0, 0.3) \tag{11}$$
$$\beta_{P,HS[i]} \sim \text{Normal}(0, 0.3) \tag{12}$$

while the hyper-priors were defined as follows:

$$\mu_M \sim \text{Normal}(0, 0.5) \tag{13}$$
$$\sigma_M \sim \text{Exponential}(1) \tag{14}$$
$$\mu_b \sim \text{Normal}(0, 0.2) \tag{15}$$
$$\sigma_b \sim \text{Exponential}(1) \tag{16}$$
$$\mu_i \sim \text{Normal}(0, 0.2) \tag{17}$$
$$\sigma_i \sim \text{Exponential}(1) \tag{18}$$

where μ_M represent the average 'sample size' and σ_M the variability of such average. Moreover, μ_b and μ_i represent the average block and children random effect, respectively. Finally, in a similar manner, σ_b and σ_i represent the variability present at the blocks and children, respectively (see section 3.1).

A.5.4. Estimation procedure

The proposed models in Table 1 were estimated under the Bayesian framework. More specifically, we used the No-U-Turn Hamiltonian Monte Carlo algorithm (No-U-Turn HMC) [3, 23, 43, 55] implemented in `Stan` [67]. We used four chains of 2000 iterations each, from which the first 1000 were discarded as a warm-up required by the software, leaving 1000 effective iterations.

Additionally, all procedures were performed in `R` 4.1.2 [61] and its integration packages [66], in a 64-bit system with a MINGW 32-bit cross-compiler (`x86_64-w64-mingw32/x64`).

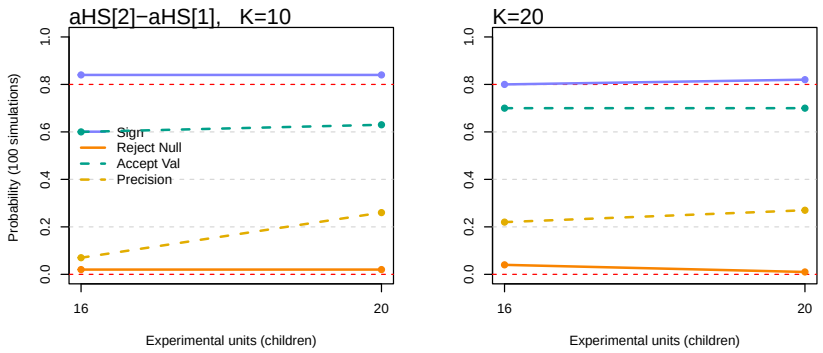

Figure 8: Group contrasts: power on small hearing group differences (≈ 0.15)

A.5.5. Simulation

Prior to fitting real data, all the proposed models were tested on simulated data. The simulation was done with two purposes: (i) to test the models' ability to recover the 'real' parameter values [35], and (ii) to test the power of our data size and assumed hypothesis [48]. On both instances, we used the *equivalent prior sampling* method [76], using an idealized data informed by the appropriate children population [19].

The results of our simulations highlight two key points. First, regarding the models ability to recover the 'real' estimate values, the probabilistic implementations work as intended, i.e. the statistical models manage to recover the 'real' values, if enough sample size is available.

Second, regarding the power of our data to affirm parameter estimates, reject null hypothesis, or provide enough precision for our inferences, three additional points can be made. On the one hand, assuming a small difference among the *hearing status* groups (approximately 0.15), sixteen children per group are enough to provide at least 60% certainty that our estimates affirm the predicted value, i.e. our model can correctly estimate values different from zero, when such alternative hypothesis is true (see green discontinuous line in figure 8). Moreover, the previous power threshold surpasses 80% if we assume medium group differences (approximately 0.4, not shown). Furthermore, in the case of other parameters of interest, even if the assumptions involve small effects (e.g. $bA = 0.15$ or $bP = 0.1$), our results always show the specific assumed value is tenable with power above 80%.

In contrast, the power analysis also reveal that with sufficient amount of variability at the children and replicates levels ($\sigma_i = 0.5$ and $M_i = 10$), the models do not have enough power to produce unequivocal null hypothesis rejections (see orange continuous lines in figure 8), i.e. the compatibility and highest posterior density intervals contain the value of zero, in almost all simulations. The latter implies that with reasonable expectations about the variability between children and the presence of measurement error (multiple replicates to measure one latent value), the statistical models require a data size significantly larger than sixteen children per group. Further model testing indicate that such unequivocal inferences can be accomplished, if we have a sample size similar to the full population size (70 children per group).

Lastly, and complementary to the previous result, the statistical precision of our estimates do not achieve power levels above the 80% threshold with our current data size (see yellow discontinuous line in figure 8). The latter implies that with sixteen children per group, the standard error of our estimates will not allow us to produce definite conclusions, e.g. we will not be certain if the two *hearing status* groups are not different because they are truly not statistically different or because the sample size is not enough to produce unequivocal inferences. This result is also related to the assumptions regarding the variability between children and the presence of measurement error. Similar to the previous case, sufficient levels of precision power can be reached, if we have a sample size similar to the full population size.

For more details on the assumed effects and the simulation code, refer to the GitHub repository: `https://github.com/jriveraespejo/PhD_UA_paper1`

A.5.6. Model selection

As stated in previous sections, the current research used the Information-Theoretic Approach [1, 12] for model selection. The application of the approach required: (i) the expression of the research hypothesis into statistical models, (ii) the selection of the most plausible models, and (iii) to produce inferences based on one or multiple selected models. Since the first step is covered in sections 2.4 and 2.5, and expanded in supplementary section A.5, here we proceed with the intermediate step.

We used the widely applicable information criterion (WAIC) [72], and the Pareto-smoothed importance sampling cross-validation (PSIS) [71] as the criteria to select among competing models. Two reasons justify our decision. First, both criteria allow us to use all the information of our Bayesian models, i.e. the posterior distribution of the parameters. Last, and more important, both criteria provide us with the best approximations for the out-of-sample (cross-validated) deviance [53]. The deviance is the best approximation for the Kullback-Liebler (KL) divergence [49], i.e. a measure of how far a model is from describing the *true* distribution of our data. In that sense, by comparing the deviance between competing models, we can measure which model is the farthest from '*perfect predictive accuracy*' for our data [53].

Table 5 and 6 report the criteria for all the proposed models, sorted by the appropriate statistic. Three patterns are visible from the tables. First, we notice that around 63% of the evidence (WE) support two types of models: (a) no interaction and one estimated 'sample size', and (b) three (similar) interaction models with one estimated 'sample size'. Furthermore, 35% of the evidence is condensed in the diverse types of *robust* models, while the remaining 2% is negligible dispersed in the remaining ones.

Second, comparing the differences in WAIC/PSIS from the 'best' model (dWAIC and dPSIS, respectively), we arrive at the same conclusion as the previous paragraph, i.e. the model with no interaction and one estimated 'sample size' is the best model, followed by the interaction models with one 'sample size', and a bit further away, by the *robust* models. One important point to highlight is that, even when there seem to be a hierarchy of fit among the models, effectively, all of these are indistinguishable from each other, as it can be inferred from their precision estimate (dSE). However, we can also observe, a more precise support for models with any type of interaction and one estimated 'sample size', than for the *robust* models.

Third, we observe a greater over-fitting penalty (pWAIC and pPSIS, respectively) for models with a fixed 'sample size' than any other model. This is quite particular, as one would expect that a model with less parameters would over-fit less the data. However in this case, it implies the complexity of the data requires more parameters, especially considering there might be some highly influential observations that might sway the estimates of the less parameterized models (see section 3.5).

more explicit here about influential analysis

Model	Name	WAIC	SE	dWAIC	dSE	pWAIC	WE
3	No interaction (one ‘size’)	-621.0	42.99	0.0	–	31.1	0.18
10	Full interaction (one ‘size’)	-621.0	42.98	0.1	0.78	31.2	0.18
9	Etiology interaction (one ‘size’)	-620.7	43.07	0.4	0.48	31.2	0.15
8	Age interaction (one ‘size’)	-620.2	42.91	0.8	0.76	32.1	0.12
11	Age interaction (robust)	-620.1	42.70	0.9	2.21	34.2	0.12
12	Etiology interaction (robust)	-619.9	42.77	1.1	2.06	34.4	0.10
4	No interaction (robust)	-619.2	42.81	1.8	2.18	34.8	0.07
13	Full interaction (robust)	-618.8	42.65	2.2	2.17	35.0	0.06
5	Age interaction (fixed ‘size’)	-578.2	52.51	42.8	16.85	50.7	0.00
6	Etiology interaction (fixed ‘size’)	-578.2	52.63	42.8	16.95	50.7	0.00
2	No interaction (fixed ‘size’)	-577.6	52.59	43.5	16.97	51.0	0.00
1	Intercept only (fixed ‘size’)	-576.8	52.98	44.3	17.40	51.8	0.00
7	Full interaction (fixed ‘size’)	-575.8	52.58	45.3	16.96	52.1	0.00

SE	WAIC approximate standard error
dWAIC	difference in WAIC from the best model
dSE	standard error for the difference in WAIC from the best model
pWAIC	WAIC over fitting penalty
WE	weight of evidence

Table 5: Fit of statistical models: WAIC.

Model	Name	PSIS	SE	dPSIS	dSE	pPSIS	WE
3	No interaction (one ‘size’)	-619.8	42.98	0.0	–	31.8	0.22
10	Full interaction (one ‘size’)	-619.7	42.97	0.0	0.82	31.8	0.22
9	Etiology interaction (one ‘size’)	-619.2	43.05	0.5	0.51	32.0	0.17
8	Age interaction (one ‘size’)	-619.0	42.89	0.7	0.81	31.8	0.15
11	Age interaction (robust)	-617.8	42.70	2.0	2.19	35.4	0.08
12	Etiology interaction (robust)	-617.7	42.77	2.1	2.07	35.5	0.08
4	No interaction (robust)	-616.5	42.81	3.3	2.19	36.1	0.04
13	Full interaction (robust)	-616.3	42.64	3.4	2.13	36.2	0.04
6	Etiology interaction (fixed ‘size’)	-576.0	52.63	43.7	17.00	51.7	0.00
5	Age interaction (fixed ‘size’)	-575.9	52.54	43.9	16.94	51.9	0.00
2	No interaction (fixed ‘size’)	-575.3	52.59	44.4	17.03	52.1	0.00
1	Intercept only (fixed ‘size’)	-574.6	52.97	45.1	17.44	52.9	0.00
7	Full interaction (fixed ‘size’)	-573.8	52.61	46.0	17.01	53.1	0.00

SE	WAIC approximate standard error
dPSIS	difference in PSIS from the best model
dSE	standard error for the difference in PSIS from the best model
pPSIS	PSIS over fitting penalty
WE	weight of evidence

Table 6: Fit of statistical models: PSIS.

Bibliography

[1] Anderson, D. [2008]. *Model Based Inference in the Life Sciences: A Primer on Evidence*, Springer.

[2] Baudonck, N., Buekers, R., Gillebert, S. and Van Lierde, K. [2008]. Speech intelligibility of flemish children as judged by their parents, *Folia Phoniatrica et Logopaedica* **61**(5): 288–295.
doi: https://doi.org/10.1159/000235994.

[3] Betancourt, M. and Girolami, M. [2012]. Hamiltonian monte carlo for hierarchical models.
url: https://arxiv.org/abs/1312.0906v1.

[4] Boonen, N., Kloots, H. and Gillis, S. [2020]. Rating the overall speech quality of hearing-impaired children by means of comparative judgements, *Journal of Communication Disorders* **83**: 1675–1687.
doi: https://doi.org/10.1016/j.jcomdis.2019.105969.

[5] Boonen, N., Kloots, H., Nurzia, P. and Gillis, S. [2021]. Spontaneous speech intelligibility: early cochlear implanted children versus their normally hearing peers at seven years of age, *Journal of Child Language* pp. 1–26.
doi: https://doi.org/10.1017/S0305000921000714.

[6] Boons, T., Brokx, J., Dhooge, I., Frijns, J., Peeraer, L., Vermeulen, A., Wouters, J. and van Wieringen, A. [2012]. Predictors of spoken language development following pediatric cochlear implantation, *Ear and Hearing* **33**(5): 617–639.
doi: https://doi.org/10.1097/AUD.0b013e3182503e47.

[7] Boons, T., De Raeve, T., Langereis, M., Peeraer, L., Wouters, L. and van Wieringen, A. [2013]. Expressive vocabulary, morphology, syntax and narrative skills in profoundly deaf children after early cochlear implantation, *Research in Developmental Disabilities* **34**(6): 2008–2022.
doi: https://doi.org/10.1016/j.ridd.2013.03.003.
url: https://www.sciencedirect.com/science/article/pii/S0891422213001078.

[8] Bowen, C. [2011]. Table1: Intelligibility.
url: http://www.speech-language-therapy.com.

[9] Bruijnzeel, H., Ziylan, F., Stegeman, I., V., T. and Grolman, W. [2016]. A systematic review to define the speech and language benefit of early (<12 months) pediatric cochlear implantation, *Audiol Neurotol* **21**: 113–126.
doi: https://doi.org/10.1159/000443363.

[10] Carroll, J. [2006]. *Measurement error in nonlinear models: a modern perspective*, Chapman and Hall/CRC.
doi: https://doi.org/10.1201/9781420010138.

[11] Castellanos, I., Kronenberger, W., Beer, J., Henning, S., Colson, B. and Pisoni, D. [2014]. Preschool speech intelligibility and vocabulary skills predict long-term speech and language outcomes following cochlear implantation in early childhood, *Cochlear Implants International* **15**(4): 200–210.
doi: https://doi.org/10.1179/1754762813Y.0000000043.

[12] Chamberlain, T. [1965]. The method of multiple working hypotheses, *Science* **148**(3671): 754–759.
url: https://www.jstor.org/stable/1716334.

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

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

[15] Chin, S. and Tsai, P. [2001]. Speech intelligibility of children with cochlear implants and children with normal hearing: A preliminary report. Progress report. Indiana University, Bloomington, Indiana.

[16] Chin, S., Tsai, P. and Gao, S. [2003]. Connected speech intelligibility of children with cochlear implants and children with normal hearing, *American journal of speech-language pathology* **12**(4): 440–451.
doi: https://doi.org/10.1044/1058-0360(2003/090).
url: https://pubs.asha.org/doi/10.1044/1058-0360

[17] Cinelli, C., Forney, A. and Pearl, J. [2022]. A crash course in good and bad controls, *SSRN* .
doi: http://dx.doi.org/10.2139/ssrn.3689437.
url: https://ssrn.com/abstract=3689437.

[18] Cohen, J. [1988]. *Statistical power analysis for the behavioral sciences*, Routledge.

[19] De Raeve, L. [2016]. Cochlear implants in belgium: Prevalence in paediatric and adult cochlear implantation, *European Annals of Otorhinolaryngology, Head and Neck Diseases* **133**: S57–S60. 12th European Symposium on Pediatric Cochlear Implant (ESPCI 2015).
doi: https://doi.org/10.1016/j.anorl.2016.04.018.
url: https://www.sciencedirect.com/science/article/pii/S1879729616300813.

[20] Deffner, D., Rohrer, J. and McElreath, R. [2022]. A causal framework for cross-cultural generalizability, *Advances in Methods and Practices in Psychological Science* . (in press).

[21] Dettman, S., Dowell, R., Choo, D., Arnott, W., Abrahams, Y., Davis, A., Dornan, D., Leigh, J., Constantinescu, G., Cowan, R. and Briggs, R. [2016]. Long-term communication outcomes for children receiving cochlear implants younger than 12 months, *Otology & Neurotology* **37**(2): e82–e95.
doi: https://doi.org/10.1097/MAO.0000000000000915.

[22] Drennan, W. and Rubinstein, J. [2008]. Music perception in cochlear implant users and its relationship with psychophysical capabilities, *Journal of Rehabilitation Research and Development* **45**: 779–790.
doi: https://doi.org/10.1682/JRRD.2007.08.0118.

[23] Duane, S., Kennedy, A., Pendleton, B. and Roweth, D. [1987]. Hybrid monte carlo, *Physics Letters B* **195**(2): 216–222.
doi: https://doi.org/10.1016/0370-2693(87)91197-X.
url: https://www.sciencedirect.com/science/article/pii/037026938791197X.

[24] Duchesne, L. and Marschark, M. [2019]. Effects of age at cochlear implantation on vocabulary and grammar: A review of the evidence, *American Journal of Speech-Language Pathology* **28**(4): 1673–1691.
doi: 10.1044/2019_AJSLP-18-0161.
url: https://pubs.asha.org/doi/abs/10.1044/2019$_A JSLP - 18 - 0161$.

[25] Ertmer, D. [2011]. Assessing speech intelligibility in children with hearing loss: Toward revitalizing a valuable clinical tool, *Language, Speech, and Hearing Services in Schools* **42**(1): 52–58.
doi: https://doi.org/10.1044/0161-1461(2010/09-0081).

[26] Everitt, B. [1984]. *An Introduction to Latent Variable Models*, Monographs on Statistics and Applied Probability, Springer Dordrecht.
doi: https://doi.org/10.1007/978-94-009-5564-6.

[27] Faes, J., De Maeyer, S. and Gillis, S. [2021]. Speech intelligibility of children with an auditory brainstem implant: a triple-case study, pp. 1–50. (submitted).

[28] Fagan, M., Eisenberg, L. and Johnson, K. [2020]. Investigating early pre-implant predictors of language and cognitive development in children with cochlear implants, *in* M. Marschark and H. Knoors (eds), *Oxford handbook of deaf studies in learning and cognition*, Oxford University Press, pp. 46–95.
doi: https://doi.org/10.1093/oxfordhb/9780190054045.013.3.

[29] Farrar, D. and Glauber, R. [1967]. Multicollinearity in regression analysis: The problem revisited, *Review of Economics and Statistics* **49**(1): 92–107.
doi: https://doi.org/10.2307/1937887.
url: https://www.jstor.org/stable/1937887.

[30] Figueroa-Zúñiga, J., Arellano-Valle, R. and Ferrari, S. [2013]. Mixed beta regression, *Computational Statistics Data Analysis* **61**: 137–147.
doi: https://doi.org/10.1016/j.csda.2012.12.002.

[31] Flexer, C. [2011]. Cochlear implants and neuroplasticity: linking auditory exposure and practice, *Cochlear Implants International* **12**(sup1): S19–S21.
doi: https://doi.org/10.1179/146701011X13001035752255.

[32] Flipsen, P. [2006]. Measuring the intelligibility of conversational speech in children, *Clinical Linguistics & Phonetics* **20**(4): 303–312.
doi: https://doi.org/10.1080/02699200400024863.

[33] Flipsen, P. [2008]. Intelligibility of spontaneous conversational speech produced by children with cochlear implants: A review, *International Journal of Pediatric Otorhinolaryngology* **72**(5): 559–564.
doi: https://doi.org/10.1016/j.ijporl.2008.01.026.
url: https://www.sciencedirect.com/science/article/pii/S0165587608000645.

[34] Flipsen, P. and Colvard, L. [2006]. Intelligibility of conversational speech produced by children with cochlear implants, *Journal of Communication Disorders* **39**(2): 93–108.
doi: https://doi.org/10.1016/j.jcomdis.2005.11.001.
url: https://www.sciencedirect.com/science/article/pii/S0021992405000614.

[35] Fogarty, L., Madeleine, A., Holding, T., Powell, A. and Kandler, A. [2022]. Ten simple rules for principled simulation modelling, *PLOS Computational Biology* **18**(3): 1–8.
doi: https://doi.org/10.1371/journal.pcbi.1009917.

[36] Freeman, V., Pisoni, D., Kronenberger, W. and Castellanos, I. [2017]. Speech intelligibility and psychosocial functioning in deaf children and teens with cochlear implants, *Journal of Deaf Studies and Deaf Education* **22**(3): 278–289.
doi: https://doi.org/10.1093/deafed/enx001.

[37] Geers, A. and Nicholas, J. [2013]. Enduring advantages of early cochlear implantation for spoken language development, *Journal of speech, language, and hearing research* **56**(2): 643–655.
doi: https://doi.org/10.1044/1092-4388(2012/11-0347.

[38] Geers, A., Nicholas, J., Tobey, E. and Davidson, L. [2016]. Persistent language delay versus late language emergence in children with early cochlear implantation, *Journal of Speech, Language, and Hearing Research* **59**(1): 155–170.
doi: 10.1044/2015_JSLHR-H-14-0173.
url: https://pubs.asha.org/doi/abs/10.1044/2015$_JSLHR-H-14-0173$.

[39] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. [2014]. *Bayesian Data Analysis*, Texts in Statistical Science, third edn, Chapman and Hall/CRC.

[40] Gillis, S. [2018]. Speech and language in congenitally deaf children with a cochlear implant, *in* E. Dattner and D. Ravid (eds), *Handbook of Communication Disorders: Theoretical, Empirical, and Applied Linguistic Perspectives*, De Gruyter Mouton, chapter 37, pp. 765–792.
doi: https://doi.org/10.1515/9781614514909-038.

[41] Grandon, B., Martinez, M., Samson, A. and Vilain, A. [2020]. Long-term effects of cochlear implantation on the intelligibility of speech in french-speaking children, *Journal of Child Language* **47**(4): 881892.
doi: https://doi.org/10.1017/S0305000919000837.

[42] Habib, M., Waltzman, S., Tajudeen, B. and Svirsky, M. [2010]. Speech production intelligibility of early implanted pediatric cochlear implant users, *International Journal of Pediatric Otorhinolaryngology* **74**(8): 855–859.
doi: https://doi.org/10.1016/j.ijporl.2010.04.009.
url: https://www.sciencedirect.com/science/article/pii/S0165587610002004.

[43] Hoffman, M. and Gelman, A. [2014]. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, *Journal of Machine Learning Research* **15**: 1593–1623.
url: https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf.

[44] Hoyle, R. e. [2014]. *Handbook of Structural Equation Modeling*, Guilford Press.

[45] Hustad, K., Mahr, T., Natzke, P. and Rathouz, P. [2020]. Development of speech intelligibility between 30 and 47 months in typically developing children: A cross-sectional study of growth, *Journal of Speech, Language, and Hearing Research* **63**(6): 1675–1687.
doi: https://doi.org/10.1044/2020$_{J}SLHR-20-00008$.
url :$https://pubs.asha.org/doi/abs/10.1044/2020_{J}SLHR-20-00008$.

[46] Jaynes, E. [2003]. *Probability Theory: The Logic of Science*, Cambridge University Press.

[47] Kent, R., Weismer, G., Kent, J. and Rosenbek, J. [1989]. Toward phonetic intelligibility testing in dysarthria, *Journal of Speech and Hearing Disorders* **54**(4): 482–499.
doi: https://doi.org/10.1044/jshd.5404.482.

[48] Kruschke, D. [2015]. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, Elsevier.
url: https://www.sciencedirect.com/book/9780124058880/doing-bayesian-data-analysis.

[49] Kullback, S. and Leibler, R. [1951]. On information and sufficiency, *The Annals of Mathematical Statistics* **22**(1): 79–86.
url: http://www.jstor.org/stable/2236703.

[50] Lesterhuis, M. [2018]. *The validity of comparative judgement for assessing text quality: An assessors perspective*, PhD thesis, University of Antwerp.

[51] MacWhinney, B. [2020]. *The CHILDES Project: Tools for Analyzing Talk*, Lawrence Erlbaum Associates. 3rd Edition.
doi: https://doi.org/10.21415/3mhn-0z89.

[52] Mayer, M. [1969]. *Frog, where are You?*, Boy, a Dog, and a Frog, Dial Books for Young Readers.
url: https://books.google.be/books?id=Asi5KQAACAAJ.

[53] McElreath, R. [2020]. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*, Chapman and Hall/CRC.

[54] Montag, J., AuBuchon, A., Pisoni, D. and Kronenberger, W. [2014]. Speech intelligibility in deaf children after long-term cochlear implant use, *Journal of Speech, Language, and Hearing Research* **57**(6): 2332–2343.
doi: https://doi.org/10.1044/2014_JSLHR-H-14-0190.
url: https://pubs.asha.org/doi/abs/10.1044/2014$_{J}SLHR-H-14-0190$.

[55] Neal, R. [2012]. Mcmc using hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. Jones and X. Meng (eds), *Handbook of Markov Chain Monte Carlo*, Chapman Hall/CRC Press, chapter 5, pp. 113–162.
url: https://arxiv.org/abs/1206.1901.

[56] Nicholas, J. and Geers, A. [2007]. Will they catch up? the role of age at cochlear implantation in the spoken language development of children with severe to profound hearing loss, *Journal of speech, language, and hearing research* **50**(4): 1048–1062.
doi: https://doi.org/10.1044/1092-4388(2007/073).

[57] Niparko, J., Tobey, E., Thal, D., Eisenberg, L., Wang, N., Quittner, A. and Fink, N. [2010]. Spoken Language Development in Children Following Cochlear Implantation, *JAMA* **303**(15): 1498–1506.
doi: https://doi.org/10.1001/jama.2010.451.

[58] Nittrouer, S., Caldwell-Tarr, A., Moberly, A. and Lowenstein, J. [2014]. Perceptual weighting strategies of children with cochlear implants and normal hearing, *Journal of Communication Disorders* **52**: 111–133.
doi: https://doi.org/10.1016/j.jcomdis.2014.09.003.
url: https://www.sciencedirect.com/science/article/pii/S0021992414000768.

[59] Pearl, J. [2009]. *Causality: Models, Reasoning and Inference*, Cambrige University Press.

[60] Peng, S., Spencer, L. and Tomblin, J. [2004]. Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience, *Journal of speech, language, and hearing research* **47**(6): 1227–1236.
doi: https://doi.org/10.1044/1092-4388(2004/092).

[61] R Core Team [2015]. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
url: http://www.R-project.org/.

[62] Rohrer, J., Schmukle, S. and McElreath, R. [2021]. The only thing that can stop bad causal inference is good causal inference, *PsyArXiv* .
doi: https://doi.org/10.31234/osf.io/mz5jx.

[63] Rowe, B. and Levine, D. [2018]. *A Concise Introduction to Linguistics*, Routledge.

[64] Sawilowsky, S. [2009]. New effect size rules of thumb, *Journal of Modern Applied Statistical Methods* **8**(2).
doi: https://doi.org/10.22237/jmasm/1257035100.
url: http://digitalcommons.wayne.edu/jmasm/vol8/iss2/26.

[65] Shannon, C. [1948]. A mathematical theory of communication, *The Bell System Technical Journal* **27**(3): 379–423.
doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[66] Stan Development Team [2020]. RStan: the R interface to Stan. R package version 2.21.2.
url: http://mc-stan.org/.

[67] Stan Development Team. [2021]. *Stan Modeling Language Users Guide and Reference Manual, version 2.26*, Vienna, Austria.
url: https://mc-stan.org.

[68] Trochim, W. [2022]. The research methods knowledge base.
url: https://conjointly.com/kb/.

[69] van Daal, T. [2020]. *Making a choice is not easy?!: Unravelling the task difficulty of comparative judgement to assess student work*, PhD thesis, University of Antwerp.

[70] van Heuven, V. [2008]. Making sense of strange sounds: (mutual) intelligibility of related language varieties. a review, *International Journal of Humanities and Arts Computing* **2**(1-2): 39–62.
doi: https://doi.org/10.3366/E1753854809000305.

[71] Vehtari, A., Simpson, D., Gelman, A., Yao, Y. and Gabry, J. [2021]. Pareto smoothed importance sampling.
url: https://arxiv.org/abs/1507.02646.

[72] Watanabe, S. [2013]. A widely applicable bayesian information criterion, *Journal of Machine Learning Research 14* **14**: 867–897.
url: https://dl.acm.org/doi/10.5555/2567709.2502609.

[73] Whitehill, T. and Chau, C. [2004]. Single-word intelligibility in speakers with repaired cleft palate, *Clinical Linguistics and Phonetics* **18**: 341–355.
doi: https://doi.org/10.1080/02699200410001663344.

[74] Wie, O. B. [2010]. Language development in children after receiving bilateral cochlear implants between 5 and 18 months, *International Journal of Pediatric Otorhinolaryngology* **74**(11): 1258–1266.
doi: https://doi.org/10.1016/j.ijporl.2010.07.026.
url: https://www.sciencedirect.com/science/article/pii/S0165587610003708.

[75] Wie, O., Torkildsen, J., Schauber, S., Busch, T. and Litovsky, R. [2020]. Long-term language development in children with early simultaneous bilateral cochlear implants, *Ear and Hearing* **41**(5): 1294–1305.
doi: https://doi.org/10.1097/AUD.0000000000000851.

[76] Winkler, R. [1967]. The assessment of prior distributions in bayesian analysis, *Journal of the American Statistical Association* **62**(319): 776–800.
doi: https://doi.org/10.1080/01621459.1967.10500894.
url: ttps://www.tandfonline.com/doi/abs/10.1080/01621459.1967.10500894.

[77] Wright, B. [2005]. Qualtrics. (Version December 2018).
url: www.qualtrics.com.

[78] Yanbay, E., Hickson, L., Scarinci, N., Constantinescu, G. and Dettman, S. [2014]. Language outcomes for children with cochlear implants enrolled in different communication programs, *Cochlear Implants International* **15**(3): 121–135.
doi: https://doi.org/10.1179/1754762813Y.0000000062.

[79] Yarkoni, T. [2020]. The generalizability crisis, *The Behavioral and brain sciences* **45**(e1).
doi: https://doi.org/10.1017/S0140525X20001685.

[80] Young, G. and Killen, D. [2002]. Receptive and expressive language skills of children with five years of experience using a cochlear implant, *Annals of Otology, Rhinology & Laryngology* **111**(9): 802–810.
doi: https://doi.org/10.1177/000348940211100908.