Speech intelligibility measurement

A latent variable approach on transcription of utterances

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, & Psycholinguistics Research Centre
University of Antwerp, Antwerp, Belgium
E-mail: steven.gillis@uantwerpen.be

May 20, 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 [44, 72, 69, 34]. Because intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered [34], 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 [14].

The literature suggest two perspectives from which *speech intelligibility* can be assessed: the message and listener's perspective [5, 6]. 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 [5, 6]. On both instances, the children's utterances can be generated from reading at loud, contextualized utterances, or spontaneous speech tasks[1].

Moreover, perceptual studies can use multiple approaches to measure intelligibility, but they can be largely grouped into two: objective and subjective ratings [42]. In *objective rating* methods, listeners transcribe 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 words contained in an utterance [6]. In contrast, under *subjective rating* methods listeners directly infer the intelligibility score, by assessing the speech sample's quality through specific procedures, e.g. absolute holistic, analytic, or comparative judgments, among others.

It is easy to deduce that *objective rating* methods produce more valid[2] and reliable[3] scores than the *subjective rating* counterpart, and therefore, are used as an objective measure of intelligibility [6, 26].

Accompanying the previous, the literature supply a myriad of factors that are thought also contribute to the (under)development of intelligible spoken language [54, 8, 37, 27]. 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), gender, and additional disabilities (mental retardation, speech motor problems). 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, for each of the 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 in the Qualtrics environment [75]. Finally, the transcriptions were transformed into replicated entropy measures that served as our outcome variable.

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 [25]. More specifically, we model *speech intelligibility* as a latent variable that can be inferred from the replicate entropy measures. 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's estimates, leading us to incorrect inferences [50]. Finally, the method also provides a criterion on how consistent are the entropy replicates to measure speech intelligibility, i.e. a reliability score.

Second, we use Directed Acyclic Graph (DAG) [56, 18] 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 such causal assumptions [50, 77, 62].

[1] ordered on increasing level of ecological validity [30, 24]

[2] the extent to which scores are appropriate for their intended interpretation and use [47, 67].

[3] the extend to which a measure would give us the same result over and over again [67], 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 method.

2. Materials and Methods

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[4]. First, a sample of sixteen hearing-impaired children (ten boys, six girls) was selected based on the quality of their registered stimuli (utterances). 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 automated auditory brainstem response hearing tests for newborns, and receive 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 1, at the supplementary section A.1.

2.2. Stimuli

The stimuli consisted of the 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" [49] to a caregiver "who does not know the story".

The recordings were orthographically transcribed with the CLAN editor in CHAT format [48]. 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 [6]. The transcriptions were only used in the selection process of the stimuli for the experiment.

2.3. Experimental setup

The experiment was setup to perform a transcription task in the Qualtrics environment [75]. 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 entropy measures (H), which served as our outcome variable.

[4] similar to one outlined in Faes et al. [26]

The entropy of utterances (H) is a measure bounded in the continuum $[0, 1]$, and it was used as a quantification of (dis)agreement between listeners' transcriptions, where 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 ($H \to 1$) [6, 26]. The procedure followed to calculate the entropies is detailed in the supplementary section A.2.2.

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) [56, 18] 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*.

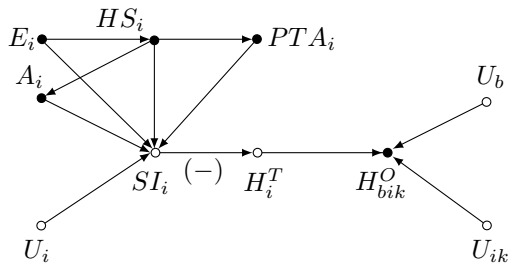

Figure 1: DAG: causal diagram describing the relationships among the analyzed variables

Figure 1 shows the *causal diagram* for our research hypothesis. In the figure, H_{bik}^O denote the *observed* entropy replicates nested within children and experimental blocks, 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 children's "true" entropy and speech intelligibility scores, respectively. In addition, A_i denotes the children's *hearing age*, E_i the etiology of the disease that led to the hearing impairment, HS_i the hearing status group, and PTA_i the post-implant pure tone average.

Three main features can be emphasized from the figure. First, the children's speech intelligibility and "true" entropy scores are are drawn with open circles, indicating the scores are thought to be latent/unobservable variables [25] (see supplementary section A.3, on the appropriate interpretations of the scores). The figure also shows the scores are thought to be inferable from the (observed) entropy replicates. Finally, it shows the intelligibility score is inversely/negatively related to the "true" entropy, i.e. the lower the intelligibility the higher the entropy and vice versa, as expected from our the theory.

Second, the figure reflects the expected hierarchy of variability 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 [50]. 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. Therefore, we expect a hierarchy with children, replicates and block levels (U_i, U_{ik} and U_b, respectively).

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 [78, 57, 51, 12, 76, 55, 34]. Finally, we did not had any comparable expectation for the variability in the replicates, as this feature has not been investigated before.

Third, the figure shows the assumed relationship among the relevant variables [54, 8, 37, 27], and how these influence the children's intelligibility of speech. Furthermore, it also reveals that we assume the variables are independent, beyond the described relationships. Here follows a description of our causal hypothesis related to the relevant variables (see supplementary section A.4 for more details).

About ***hearing age*** and ***hearing status***, we expect the former to be responsible for the increase in the speech intelligibility of children. Several studies provide evidence that for NH children, intelligibility increases with chronological age [16, 17, 30, 31, 3, 10, 42], and similar evidence can be found for the

HI/CI children. Moreover, recent literature seem to suggest the effects are independent of the children's hearing status [6].

For the latter, we do not have a clear expectation about the intelligibility levels among the groups. Previous literature suggest that some HI/CI children catch up with their NH counterparts [73, 39, 9, 35, 11, 20, 74]. However, additional studies also seem to indicate the HI/CI children never reach similar levels than their NH counterparts [53, 12, 15, 36, 34, 23, 38].

Additionally, we expect ***pure tone average*** to have a small or null effect on speech intelligibility, as the evidence seem to suggest [6]. PTA is the child's subjective hearing sensitivity, aided or unaided, by their hearing apparatus.

Finally, we expect the ***Etiology*** of the disease that led to the hearing impairment to have a differential effect on speech intelligibility, within the HI/CI group. However, since the severity of the etiology cannot be easily ascertain, we cannot foresee the direction of such effects, e.g. genetic factors not necessarily lead to worse levels of speech intelligibility than factors related to infections.

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 [50, 77, 62].

2.5. Statistical analysis

Using the DAG described in previous section, we produced a full structural causal model though the use of probabilistic programming [43], that we later used to validate our models of analysis and code.

3. Results

3.1. Model selection and results

3.2. Speech intelligibility scale

3.3. Posterior predictive

3.4. Outlying observations

4. Discussion

talk about decision statements or thinking-at-loud tasks.. the listener provide a decision statement on why the selected stimulus sounded more intelligible

We relied on the DAG presented in section 2.4 only to define the analyses described 2.5. A better planning could be done by designing the experimental design according to our hypothesis described by the causal framework. but ....

5. 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.

6. Financial support

What is the financial support of the project

7. Conflicts of interest

The authors declare they have no conflict of interest.

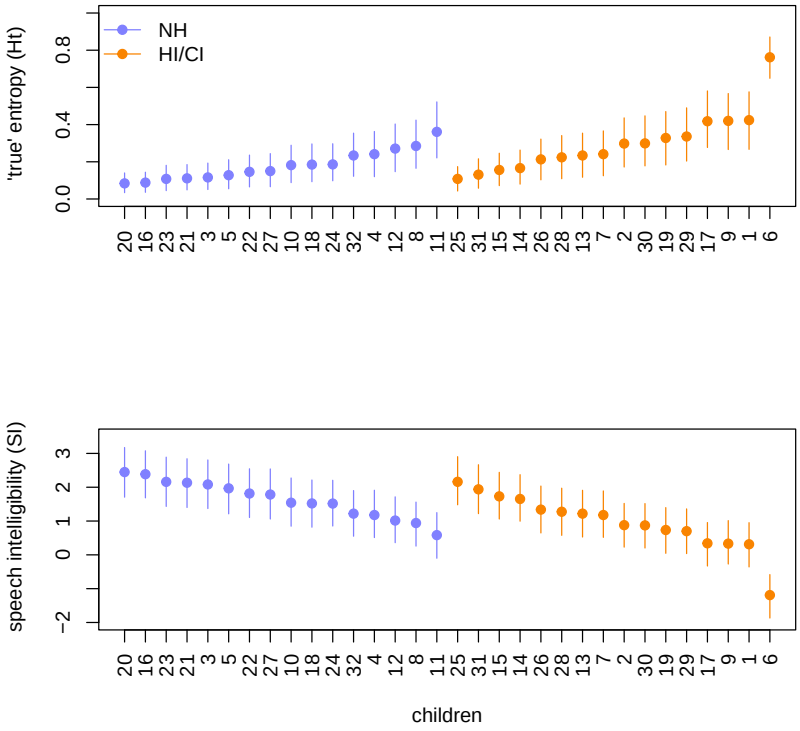

Figure 2: Posterior predictive: "true" entropy and speech intelligibility scales

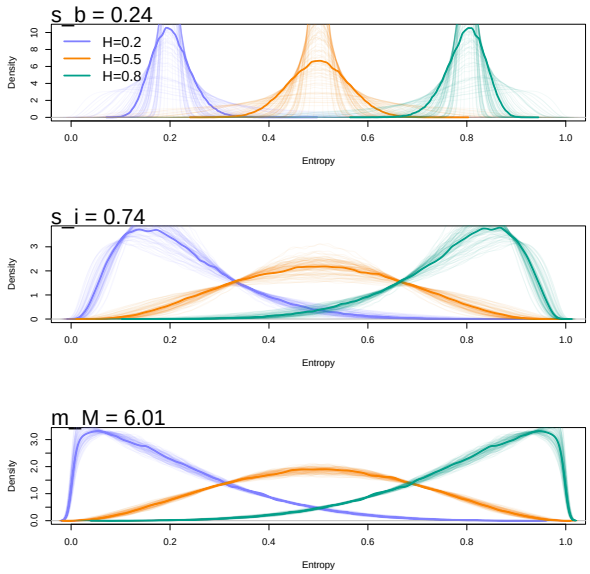

Figure 3: Posterior predictive: levels of variability

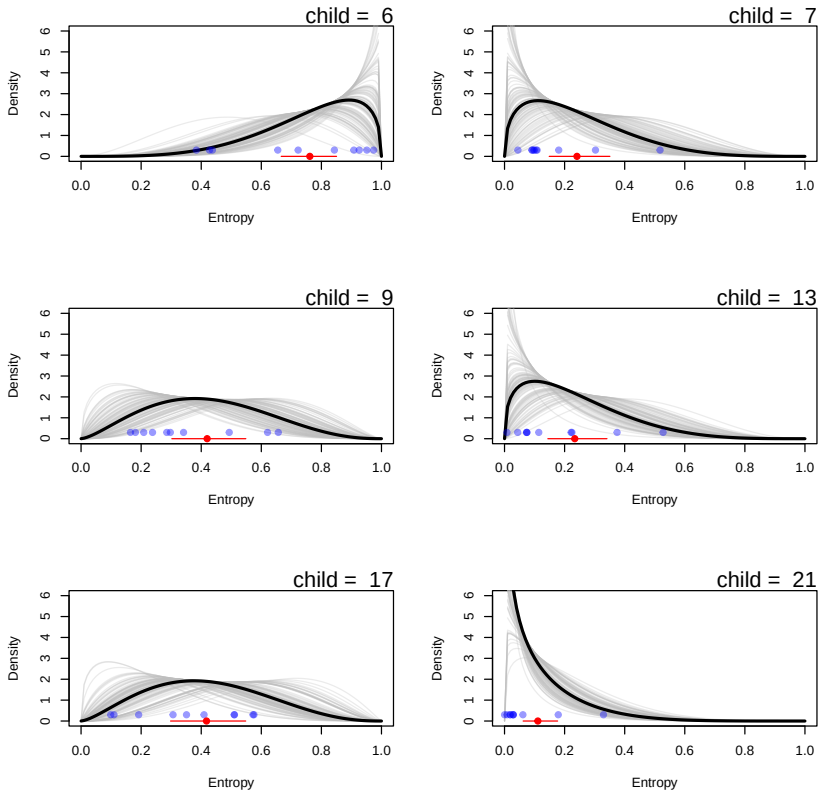

Figure 4: Posterior predictive: entropy replicates

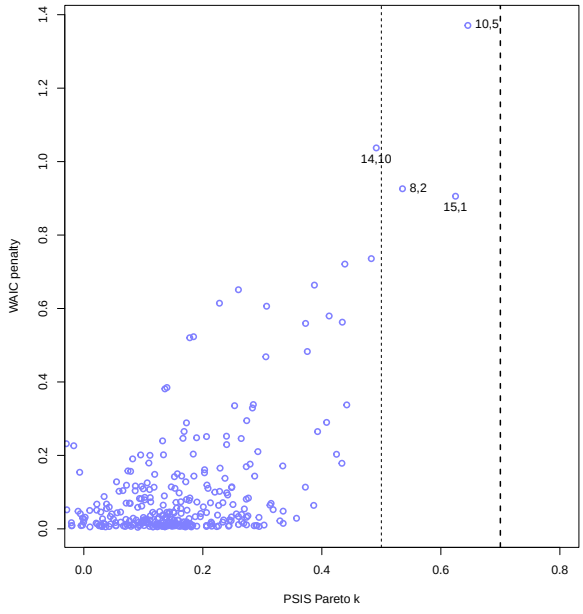

Figure 5: Outlying observations. Pairs (child, utterance) are reported for specific observations.

8. 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 1 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.

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 1: Characteristics of selected children.

A.2. Experiment details

A.2.1. Transcription task

The setting for the transcription task comprised the following steps [5, 6]:

1. the listener took a seat in front of a computer screen, located at the campus' computer laboratory.

2. the listener opened Qualtrics [75] 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*, e.g. the listeners were instructed to write one `X` to replace an unintelligible word, part of an utterance, or a complete utterance,
   b) *the aspects not to consider 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 environment.

A.2.2. Entropy calculation

The outcome from the transcription task was obtained following a two step procedure [6]. 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 2. This step was repeated for every one of the 6400 transcriptions. Lastly, we computed the entropy measure of the aligned transcriptions as in Shannon [64]:

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

where H is bounded in the continuum $[0, 1]$, 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 2: Alignment and entropy calculation. Extracted from Boonen et al. [6], 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 ($H \to 1$) [6, 26].

To exemplify relevant scenarios for the procedure, we generate the entropy for utterances 2, 4 and 5 present in Table 2. 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 of the transcriptions identify it with different words. Notice we consider the unidentifiable word [X] in the second transcription, as being different from the one in the last. This is done to avoid the artificial reduction of the entropy measure, as [X] values already indicate the lack of intelligibility of the word. Therefore, 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

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 [44, 72, 69, 34]. More specifically, in the context of the transcription task, speech intelligibility can be inferred from the extent a set of transcribers can identify the words contained in an utterance [6].

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, by ensuring the transcription task was well understood and appropriately performed by the transcribers.

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. Causal framework details

In this section we make explicit some of the assumptions that guided our causal framework, and later, our statistical analysis.

For ***hearing age***, and its relevance in our research hypothesis, we describe how the variable is constructed and its inherent assumptions.

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 [26] (see Table 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* [32, 39, 38] or *age at implantation* [54, 8, 11, 20]. We argue that the feasibility of using any 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 [26], we argue is the most appropriate to test our hypothesis, based on the relevant literature review and its assumed reliability to capture children's language development (although the latter has not been tested). Moreover, the variable serve two additional purposes: (i) control for sampling bias (see section A.5), and (ii) de-confound the parameter estimates of *hearing status* [18].

Finally, it is important to highlight that using more than one of the aforementioned proxies in tandem for modeling purposes is not recommended. It is apparent from the previous description, the three surrogate measures share high similarities in their data construction. This in turn might cause problems in the modeling procedure, as including variables that provide "similar information" might lead to a problem known as multicollinearity, in which our estimates would be biased and less precise [28], leading us to wrong conclusions.

Regarding ***hearing status***, it is clear that its inclusion directly corresponds with the main purpose of the current research endeavor. i.e. compare the levels of speech intelligibility among NH and HI/CI children.

In the case of ***pure tone average***, the variable was included for two reasons. First, given that previous modeling efforts did not capture the full hierarchy of variability, it is possible that the effects of PTA on speech intelligibility has been largely overlooked. As one can infer, it might be 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 the signal in normal hearing scenarios [21]. Finally, as with *hearing age*, the variable might be useful to de-confound the parameter estimates of *hearing status* [18].

For ***etiology***, it is possible that its effects on speech intelligibility has also been largely overlooked, due to the lack of control on the full hierarchy of variability, similar to the *pure tone average* case. Moreover, as with its predecessors, the variable might also be useful to de-confound the parameter estimates of *hearing status* [18], assuming our DAG is appropriate.

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

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 it is measured. This in turn means that its inclusion/exclusion would not confound our estimates. Additionally, given that most of the children underwent though sequential bilateral implantation (eleven in total), we anticipated the effect of variable already permeates the sample, therefore, if we wanted to investigate its effect, a larger sample size would be required.

For the case of ***additional disabilities***, e.g. mental retardation or speech motor problems, there was no need to consider it, as no additional comorbidities were reported.

In the case of ***environmental factors***, such as communication modality, the current sample of 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 [6].

A.5. 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 [18, 19], 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.6. Model details

A.6.1. Definition

Previous research already used hierarchical models with the replicated entropy measures as outcomes [6, 26]. Hierarchical models are powerful to control for heterogeneity in the data, and also to avoid pre-aggregating procedures that could be pernicious for a proper statistical inference [50].

These claims are easier to understand using a though experiment within our research. Consider we have two children with the same mean entropy, but the second child shows more variability across the 10 utterances than the first. It is clear that the average entropy measure informs about the child's average SI, indicating that both children have similar level. However, the entropy's heterogeneity across the 10 utterances also informs about the child's SI, as a higher variability imply transcribers agreed less about the second child's intelligibility.

The intuition derived from the previous though experiment is similar to the one presented in Boonen et al. [6], and it is what justify our use of a hierarchical model. More specifically, we will use a Hierarchical (Mixed) Beta Regression model [29], for which we argue, its implementation is rather trivial under the bayesian framework, and we present it in the following lines.

First, figure **??** depicts the DAG representation of the model. For the measurement error part, section **??** reveals the (observed) entropy replicates H^O_{ik} can represent multiple realizations of a child's *true* entropy H^T_i, measured with error e_i. As a result, we can say the k'th entropy measure is nested within the i'th child, where $k = 1, \ldots, K$, $i = 1, \ldots, I$, $K = 10$ utterances, and $I = 32$ children.

Second, for the hypothesis part, we can say the child's *true* entropy H^T_i is inversely explained by the child's speech intelligibility index SI_i, and in turn, the latter by a set of covariates. Notice from Figure **??**, we propose two sets of models. The model in panel (a) use hearing status (HS_i) and hearing age (A_i) as covariates. The use of hearing status is justified as we are interested in comparing SI among groups, defined by the children's hearing characteristics (NH, HI/CI, and HI/HA). On the other hand, we expect hearing age[5] and its interaction with hearing status, to also have an effect on the SI index, as previous evidence have shown the speech of HI children gradually approximate that of NH children [7].

Notice the model depicted in panel (a) is interested on (what we can call) *total effects*, i.e. the effects of the hearing characteristics, not independent from the effects of the hearing apparatus (cochlear implant or hearing aid). This is important to understand for two reasons. Since a hearing apparatus is fitted onto a child depending on aspects such as the locus and severity of his(her) hearing impairment [45]: (1) such specific children's characteristics could confound the (beneficial) effects of using specific hearing apparatuses, while (2) because children are selected from a convenient sample, not representative of their respective populations (see section **??**), the need to control for such characteristics is paramount, if we seek to obtain effects that can generalize better and beyond our sample[6].

Considering the previous, we propose the model depicted in panel (b), where we control for the possible confounding variables etiology (E_i), as a proxy of locus, and unaided PTA (PTA_i), as a proxy for hearing impairment severity. In that sense, the model would estimate (what we can call) the *direct effects* of the hearing apparatus, independent of the children's characteristics.

Lastly, we proceed to use probabilistic programming to declare the algebraic structure of our models. Given the panel (a) model is nested within the panel (b) model, we declare only the model structure for the latter:

$$H^O_{bik} \sim \text{BetaProp}\left(P_{bi}, M_{ik}\right) \tag{2}$$

$$P_{bi} = \alpha_b + H^T_i \tag{3}$$

$$H^T_i = \text{logit}^{-1}(-SI_i) \tag{4}$$

$$SI_i = a_i + \alpha + \alpha_{E[i],HS[i]} + \beta_{A,HS[i]}(A_i - \bar{A}) + \beta_P PTA_i \tag{5}$$

$$\tag{6}$$

[5] see section **??** to know how the variable is defined.
[6] follow the *notes* folder, to see a graphical though experiment.

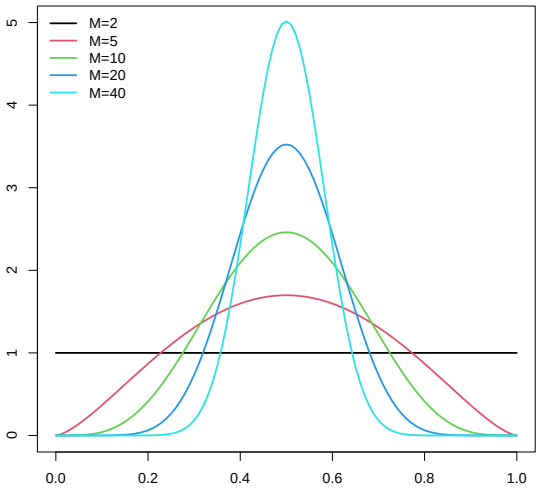

Figure 6: Variability in a Beta-Proportional distribution.

where $\text{logit}(x) = \log[x/(1-x)]$, and $\text{logit}^{-1}(x) = \exp(x)/(1+\exp(x))$. Additionally, a $\text{BetaProp}(\mu, \theta)$ distribution is equal to a $\text{Beta}(\alpha, \beta)$ distribution, with $\alpha = \mu\theta$, $\beta = (1-\mu)\theta$. For our purposes, $\mu = H_i^T$ and $\theta = M_i$, the latter denoting the "sample size" of the distribution. Moreover, a_i denote the children's random effects, α the fixed effects' intercept, $\alpha_{HS[i]}$ and $\beta_{A,HS[i]}$ the intercept and slope of "hearing age" per hearing status group, $\alpha_{E[i]}$ the intercept per etiology group, and β_P the slope for the standardized PTA levels.

Three important things need to be noticed from the previous algebraic structure. First, all the parameters are estimated in the logit scale and centered at $PTA_i = 0$ and $\bar{A}$, which denotes the minimum hearing age in the sample. Second, instead of a latent measurement error U_{ik}, we use the latent "sample size" parameter M_{ik} to model the heterogeneity/variability of the duplicate entropies. This effectively works as a measurement error model for the replicates, as the parameter defines the shape of the distribution. Third, if we do not consider etiology and PTA values in equation (4), we obtain the panel (a) model.

A.6.2. Priors

$$
\begin{aligned}
M_i &\sim \text{LN}(\mu_M, \sigma_M) && (7)\\
a_i &\sim \text{N}(\mu_a, \sigma_a) && (8)\\
\alpha &\sim \text{N}(0, 0.3) && (9)\\
\alpha_{HS[i]} &\sim \text{N}(0, 0.3) && (10)\\
\beta_{A,HS[i]} &\sim \text{N}(0, 0.3) && (11)\\
\alpha_{E[i]} &\sim \text{N}(0, 0.5) && (12)\\
\beta_P &\sim \text{N}(0, 0.3) && (13)\\
\mu_M &\sim \text{N}(0, 5) && (14)\\
\sigma_M &\sim \text{Exp}(1) && (15)\\
\mu_a &\sim \text{N}(0, 0.5) && (16)\\
\sigma_a &\sim \text{Exp}(1) && (17)\\
& && (18)
\end{aligned}
$$

Third, we use mildly informative priors to state our uncertainty regarding the direction and magnitude of the effects[7].

[7] see Rivera [61] (p. 18-19) for an intuition on prior elicitation.

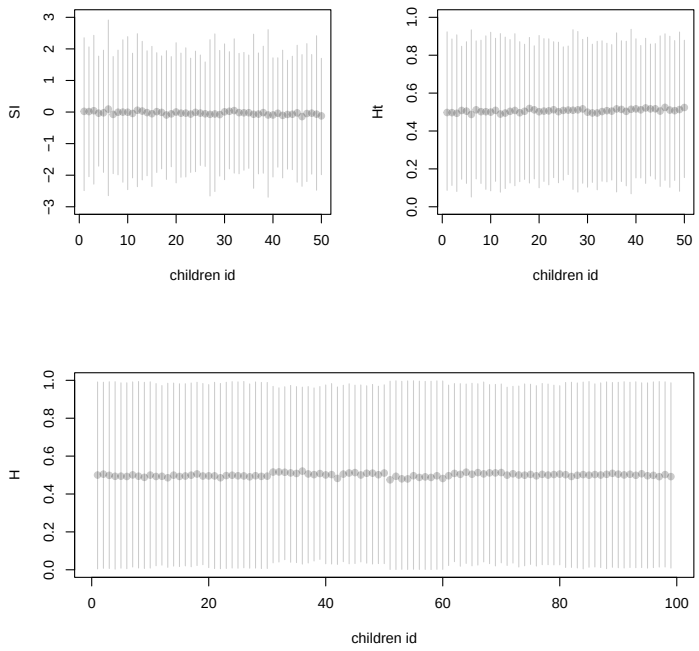

Figure 7: Prior distribution implications: speech intelligibility, “true” entropy and observed entropy scales.

A.6.3. Estimation

The models proposed in sections **??** and **??** will be estimated under the Bayesian framework[8]. More specifically, we will use the No-U-Turn Hamiltonian Monte Carlo algorithm (No-U-Turn HMC) [4, 22, 41, 52]. `Stan` [66] will be the software package that will provide us with the No-U-Turn HMC machinery, while `R` [60] and its integration packages [65], the software that will allow us to analyze its outputs.

A.6.4. Pre-processing

Besides the exclusion of corrupted observations, e.g. no available rating, no other experimental run nor duplicate was eliminated before the modeling process. This decision departs from what it is observed in previous research, e.g. Boonen et al. [5] decided to eliminate "outlying" observations based on misfit analysis [47], while van Daal [68] and Boonen et al. [6] did the same based on univariate outlier analysis.

For the case of misfit analysis, we argue that such procedures cannot be used without caution. The literature points out that in the context of CJ, these statistics are always relative, i.e. they depend on other stimulus and judges included in the assessment [58, 59]. Moreover, they have been proven to be less sensitive, as they are calculated with a low number of judgments per representation [58].

On the other hand, for the case of univariate outlier analysis, we argue that outlying observations are interesting cases to analyze [50], and usually they cannot be identified properly outside the context of a full model [50], i.e. what can behave as an outlier based on a univariate analysis, can behave as expected under the appropriate model.

Considering the previous, if we still manage to identify outlying observations within the context of the proposed models (see Section **??**), the researcher would rather make the model robust against their influence, playing on the strengths of the bayesian framework, than to eliminate the observations.

A.6.5. Simulation

Preliminary to the data collection, we simulated data in silico to test the models and inform data collection procedure. The simulation code is available in the GitHub repository. [33]

[8] see Rivera [61] (p. 11-13, 15-27) for a detailed description of its benefits and shortcomings.

A.6.6. Model selection

Following the successful and comprehensive analysis in van Daal [68] and Lesterhuis [47], the current research will also use the Information-Theoretic Approach (ITA) [2, 13] for the selection of competing models. The approach considers three steps: (1) state our hypothesis into statistical models, (2) select among competing models, and (3) make inferences based on one or multiple models.

First, for the translation of our working hypotheses into statistical models, we will use Directed Acyclic Graphs (DAG) and probabilistic programming [43]. A DAG is the simplest representation of a Graphical Causal Model (GCM), a heuristic model that contains information not purely statistical, but unlike a detailed statistical model, it allow us to deduce which variable relationships can provide valid causal inferences [40, 50]. In summary, a DAG is a reasonable way to state our hypothesis, and make our assumption more transparent. However, abide by the *no-free lunch* rule, the causal inferences produced under the DAG will only be valid if the assumed DAG is correct. In contrast, the probabilistic programming will serve as the algebraic formalist to define our statistical models.

Second, to select among competing models, we will use the Widely Applicable Information Criterion (WAIC) [71], and the Pareto-smoothed importance sampling cross-validation (PSIS) [70][9]. Two reasons justify our decision. First, both criteria allow us to embrace the full flexibility and information of our bayesian implementation (outlined in Section **??**). Last, and more important, both criteria provide us with the best approximations for the out-of-sample (cross-validated) deviance [50]. The deviance is the best approximation for the Kullback-Liebler (KL) divergence [46], i.e. a measure of how far a model is from describing the *true* distribution of our data. McElreath [50] points out that is a rather benign characteristic of the model's selection procedure that we do not need the KL divergence's absolute value, as the *true* distribution of our data is not available (otherwise, we would not need a statistical model). But rather, using the difference in deviance between competing models, we can measure which model is the farthest from *perfect (predictive) accuracy* for our data[10].

Finally, considering the evidence provided by the previous step, we proceed to make inferences based on one or multiple models.

[9] van Daal [68] used the Akaikes Information Criterion (AIC) [1] with similar purposes.
[10] see McElreath [50] (p. 202-211) for the intuition and detailed derivation of the argument.

Bibliography

[1] Akaike, H. [1974]. A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6): 716–723.
doi: https://doi.org/10.1109/TAC.1974.1100705.

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

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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

[18] 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.

[19] 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).

[20] 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.

[21] 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.

[22] 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.

[23] 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$_A JSLP - 18 - 0161$.

[24] 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).

[25] 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.

[26] 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).

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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$_JSLHR-H-14-0173$.

[37] 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.

[38] 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.

[39] 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.

[40] Hernán, M. and Robins, J. [2020]. *Causal Inference: What If*, 1 edn, Chapman and Hall/CRC.
url: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book.

[41] 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.

[42] 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$_JSLHR-20-00008$.
url :*https* : *//pubs.asha.org/doi/abs/*10.1044/2020$_JSLHR-20-00008$.

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

[44] 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.

[45] Korver, A., Smith, R., Van Camp, G., Schleiss, M., Bitner-Glindzicz, M., Lustig, L., Usami, S. and Boudewyns, A. [2017]. Congenital hearing loss, *Nature Reviews Disease Primers* **3**(16094): 278–289.
doi: https://doi.org/10.1038/nrdp.2016.94.

[46] 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.

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

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

[49] 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.

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

[51] 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$_JSLHR-H-14-0190$.

[52] 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.

[53] 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).

[54] 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.

[55] 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.

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

[57] 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).

[58] Pollitt, A. [2012a]. Comparative judgement for assessment, *International Journal of Technology and Design Education* **22**: 157–170.
doi: https://doi.org/10.1007/s10798-011-9189-x.

[59] Pollitt, A. [2012b]. The method of adaptive comparative judgement, *Assessment in Education: Principles, Policy and Practice* **19**: 281–300.
doi: https://doi.org/10.1080/0969594X.2012.665354.

[60] 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/.

[61] Rivera, J. [2021]. *Generalized Linear Latent and Mixed Models: method, estimation procedures, advantages, and applications to educational policy.*, PhD thesis, KU Leuven.

[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] 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.

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

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

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

[68] 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.

[69] 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.

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

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

[76] 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.

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

[78] 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.