Speech intelligibility measurement

A latent variable approach

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 19, 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 [21, 42, 39, 16]. Because intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered [16], 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 [8].

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 speech samples, e.g. their pitch, duration or stress (supra segmental characteristics), or the articulation of vowels and consonants (segmental characteristics) [33]. 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 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 [19]. 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 [5]. 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 infer that *objective rating* methods might produce more valid[2] and reliable[3] scores than the *subjective rating* counterpart, and therefore as their name imply, are usually used as an objective measure of intelligibility [5, 13].

Considering the previous, this paper investigates the speech intelligibility levels of normal hearing (NH) versus hearing impaired children with cochlear implants (HI/CI). For this purpose, redwe measured the entropy of representations coming from an spontaneous speech task, resulting from a transcription task.

Moreover, the paper make three specific contributions to the measurement and analysis of speech intelligibility.

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 as hearing-impaired 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 issue. Hence,

[1] ordered on increasing level of ecological validity [15, 12]

[2] the extent to which scores are appropriate for their intended interpretation and use [24, 37].

[3] the extend to which a measure would give us the same result over and over again [37], i.e. measure something, free from error, in a consistent way.

[4] similar to one outlined in Faes et al. [13]

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 as 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 2, at the supplementary section A.3.

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. The stimuli were documented when the child was telling a story cued by the picture book "Frog, where are you" [26] to a caregiver "who does not know the story".

The recordings were orthographically transcribed with the CLAN editor in CHAT format [25]. The transcriptions were only used in the selection process of the stimuli for the experiment. 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].

As a result, the data set consisted in a total of 320 stimuli presented to the listeners in a random order, based on the adaptive pairing algorithm [30] implemented in Qualtrics [43].

2.3. Experimental setup

The experiment was setup to perform a transcription task, where 100 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, resulting in 20 transcriptions per utterance. In total this amounted to 6400 transcriptions, i.e. 20 transcription times 320 utterances. The steps that comprised the task are detailed in section A.1.1.

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

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 ($H \to 1$) [5, 13]. The procedure followed to calculate the entropies is detailed in section A.1.2.

2.4. Causal framework

Where H_{ik} denoted the (observed) entropy replicates, H_i the (latent) "true" entropy, SI_i the (latent) speech intelligibility score (inversely related to H_i^T). Moreover, A_i denoted the "hearing" age (substracted the minimum age), E_i the etiology of disease that led to the hearing impairment, HS_i the hearing status and focus of our research, PTA_i the pure tone average (standardized). And Finally, B_b the block (will reduce $\sigma_{U_{ik}}$). The variables are assumed independent, beyond the described relationships, i.e.

$$\begin{aligned} P(\boldsymbol{U}) &= P(U_{ik}, U_i, U_A, U_E, U_{HS}, U_P, B_{bk}) \\ &= P(U_{ik})P(U_i)P(U_A)P(U_E)P(U_{HS})P(U_P)P(B_{bk}) \end{aligned}$$

2.5. Statistical analysis

Lastly, it is important to highlight that we composed a proxy measures of *hearing age*, i.e. amount of time a child has been actively hearing and developing his(her) language. First, for the NH group we used the *chronological age*. Finally, for the HI/CI group we use the *device's length of use* [13].

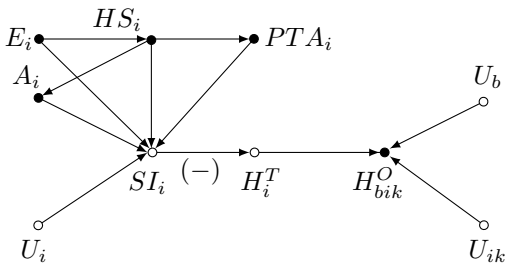

Figure 1: DAG: structural diagram describing the relationships among the analyzed variables

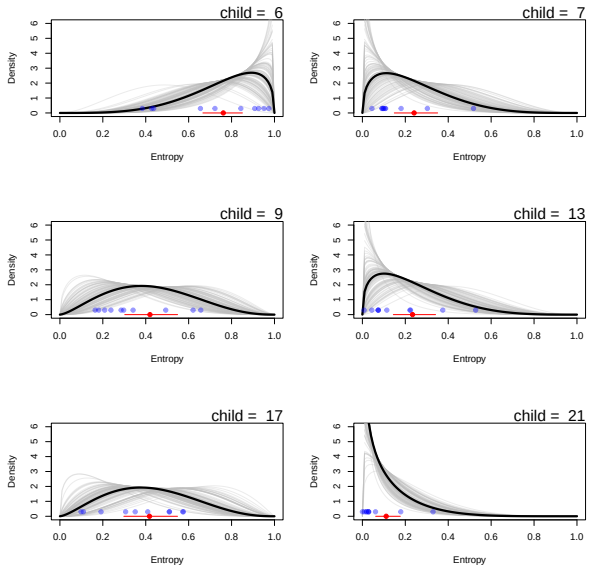

Figure 2: Posterior predictive: entropy replicates

3. Results

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

5. Author contributions

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.

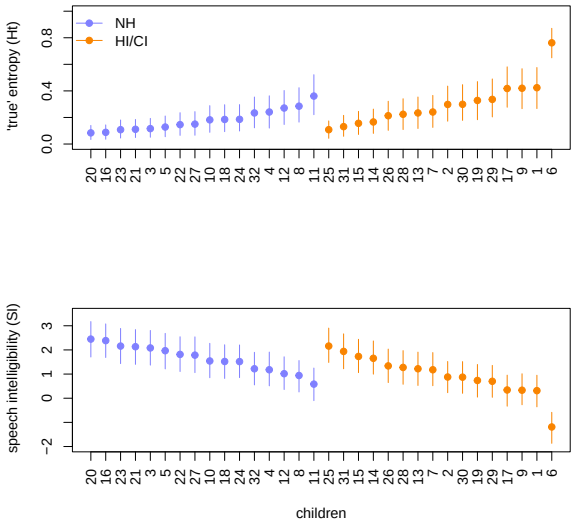

Figure 3: Posterior predictive: “true" entropy and intelligibility scales

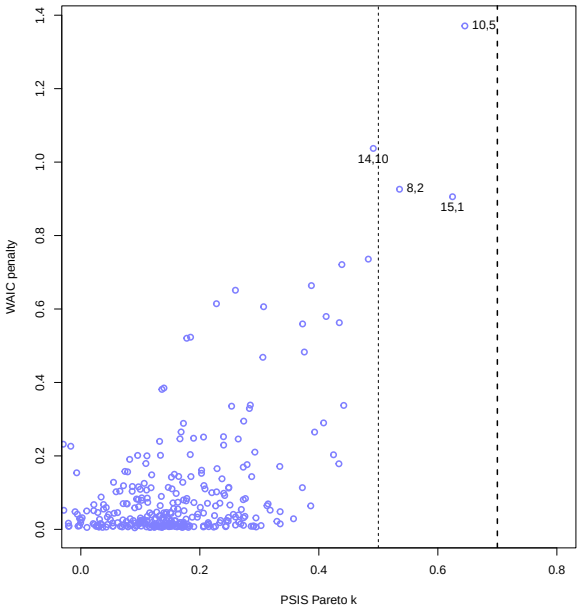

Figure 4: Outlying observations

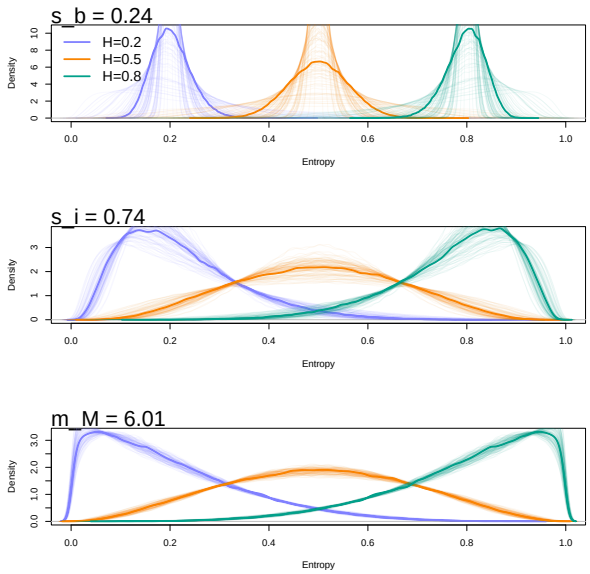

Figure 5: Posterior predictive: levels of variability

8. Research transparency and reproducibility

The model's 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. Experiment details

A.1.1. 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 [43] 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.1.2. 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 1. This step was repeated for every one of the 6400 transcriptions. Lastly, we computed the entropy measure of the aligned transcriptions as in Shannon [34]:

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

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 1: 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, 13].

To exemplify relevant scenarios for the procedure, we generate the entropy for utterances 2, 4 and 5 in Table 1. 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.2. 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 [9, 10], 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.3. Children characteristics

Table 2 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, is suspected.

As it was mentioned in previous paragraphs, *hearing age* is a composite measure, that tries to approximate the amount of time a child has been actively hearing and developing his(her) language. The

variable is constructed combining the *chronological age* for the NH group, and the *device length of use* for the HI/CI group [13].

Additionally, the table reports the child's etiology and their post-implant pure tone average (PTA). The etiology shows the cause of the children's hearing impairment, while PTA reports the child's subjective hearing sensitivity, aided and unaided by their hearing apparatus.

Child	Gender	Chronological age (y;m)	Device length of use (y;m)	Hearing age (y;m)	Etiology	PTA (dB.) unaided	PTA (dB.) 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 2: Characteristics of selected children.

A.4. 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 [21, 42, 39, 16]. 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 [5].

Therefore in this paper, through the implementation of our proposed model, *speech intelligibility* is interpreted as a latent trait of individuals, 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'*.

Despite this practical approach, we want to 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 unobserved) intelligibility of speech possessed by individuals, but do not deal with general epistemological considerations on the connection between the two.

A.5. DAG: factors influencing Intelligibility

A.6. Model details

A.6.1. Definition

Previous research already used hierarchical models with the replicated entropy measures as outcomes [5, 13]. 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 [27].

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. [5], and it is what justify our use of a hierarchical model. More specifically, we will use a Hierarchical (Mixed) Beta Regression model [14], 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 [6].

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 [22]: (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

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

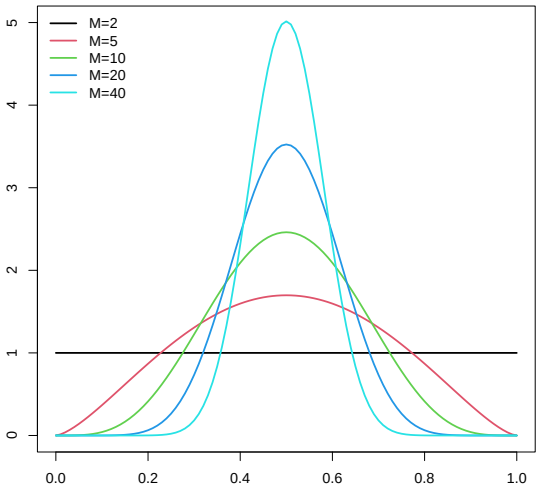

Figure 6: Variability in a Beta-Proportional distribution.

the latter:

$$H^O_{bik} \sim \text{BetaProp}\left(P_{bi}, M_{ik}\right) \tag{2}$$
$$P_{bi} = \alpha_b + H^T_i \tag{3}$$
$$H^T_i = \text{logit}^{-1}(-SI_i) \tag{4}$$
$$SI_i = a_i + \alpha + \alpha_{E[i],HS[i]} + \beta_{A,HS[i]}(A_i - \bar{A}) + \beta_P PTA_i \tag{5}$$
$$\tag{6}$$

where $\text{logit}(x) = \log\left[x/(1-x)\right]$, 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, $\mu = H^T_i$ 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.

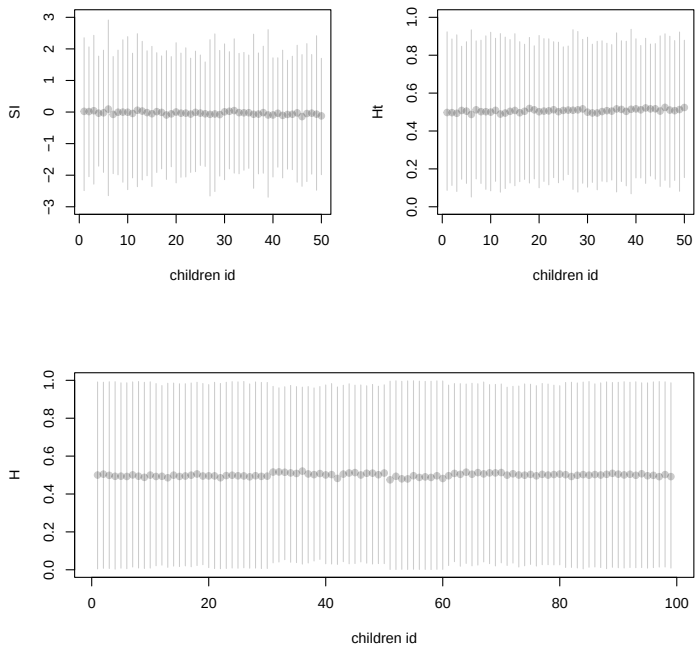

Figure 7: Prior distribution implications. Speech intelligibility, "true" entropy and observed entropy scales.

A.6.2. Priors

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

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

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) [3, 11, 18, 28]. `Stan` [36] will be the software package that will provide us with the No-U-Turn HMC machinery, while `R` [31] and its integration packages [35], the software that will allow us to analyze its outputs.

[7] see Rivera [32] (p. 18-19) for an intuition on prior elicitation.
[8] see Rivera [32] (p. 11-13, 15-27) for a detailed description of its benefits and shortcomings.

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. [4] decided to eliminate "outlying" observations based on misfit analysis [24], while van Daal [38] and Boonen et al. [5] 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 [29, 30]. Moreover, they have been proven to be less sensitive, as they are calculated with a low number of judgments per representation [29].

On the other hand, for the case of univariate outlier analysis, we argue that outlying observations are interesting cases to analyze [27], and usually they cannot be identified properly outside the context of a full model [27], 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. Several functional correlation between age and knowledge have been simulated, and the model used in the analysis - which includes age as a ordinal categorical predictor of knowledge with monotonically increasing effect - has been able to recover the different shapes. Causal effect of activities, family composition and schooling have been simulated and tested.

The simulated data have been used -albeit in a previous version- to estimate the minimum number of interviewees necessary to recover the parameter values. If individuals were to name a maximum of 300 items in the freelist, 50 interviewees would have been suffcient to obtain reliable estimates of the parameters. Given that data collection in vivo is much less regular and less controllable than in silico, we roughly doubled the number of interviewees and that of questions.

A.6.6. Model selection

Following the successful and comprehensive analysis in van Daal [38] and Lesterhuis [24], the current research will also use the Information-Theoretic Approach (ITA) [2, 7] 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 [20]. 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 [17, 27]. 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) [41], and the Pareto-smoothed importance sampling cross-validation (PSIS) [40][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 [27]. The deviance is the best approximation for the Kullback-Liebler (KL) divergence [23], i.e. a measure of how far a model is from describing the *true* distribution of our data. McElreath [27] 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).

[9] van Daal [38] used the Akaikes Information Criterion (AIC) [1] with similar purposes.

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.

[10]see McElreath [27] (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] 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] 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.

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

[8] 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.

[9] 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.

[10] 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).

[11] 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.

[12] 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).

[13] 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).

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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**: 1675–1687.
doi: https://doi.org/10.1044/2020$_J SLHR - 20 - 00008$.

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

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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/.

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

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

[34] 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.

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

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

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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

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