Research Proposal:
Absolute versus comparative judgment

Jose Rivera

josemanuel.riveraespejo@uantwerpen.be

Steven Gillis

steven.gillis@uantwerpen.be

Sven De Maeyer

sven.demaeyer@uantwerpen.be

February 18, 2022

Abstract

High level description of the research.

Contents

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 [31, 62, 57, 22]. Because intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered [22], 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 there are two perspectives from which speech intelligibility (SI) can be assessed: the message and listener's perspective [6, 7]. The first, also known as acoustic studies, center its focus 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) [49]. Whereas the second, also known as perceptual studies, center its focus on making holistic assessments of the speech stimuli, e.g. measure their perceived quality [6, 7]. On both instances, the stimuli or representations (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 SI. However, they can be largely grouped into two: objective and subjective ratings (OR and SR, respectively) [26]. In OR, listeners transcribe children's utterances orthographically or phonetically, and use such information to construct an index of SI. In contrast, under SR, listeners directly produce the SI index using one or a combination of the following procedures: absolute holistic (HJ), analytic (AJ), or comparative (CJ) judgments, the last, also known as the relative holistic method.

It is easy to infer from the previous description, that under perceptual studies, OR methods are more valid[2] and reliable[3] than SR methods, and therefore as their name imply, are usually used as an objective measure of SI [7, 18]. However, given the demanding process in terms of number of listeners required, time, and ultimately cost entailed by OR methods, SR methods can be regarded as an efficient alternative, as long as we can ensure they provide equally valid and reliable SI measures.

Furthermore, within the SR methods, the literature evidence indicate that while HJ procedures are less time consuming than any other alternative [7], they suffer from a lack of intra- and inter-rater reliability[4] [41, 28, 26, 7]. Additionally, the literature inform us the procedure does not allow to assess subtle differences in the representations [7], while the scales derived from them are usually coarse, where children reach the higher levels fairly quickly [46], e.g. the Speech Intelligibility Rating (SIR) [15, 39].

In this context, CJ has received a growing attention, because it directly tackles the issues with the HJ procedures: it fosters reliable [60] and valid scores [10, 36], while the judges can focus only on the relevant aspects of the compared representations, i.e. the "just noticeable difference" [36]. Moreover, depending on the task, it provides a set of additional benefits, e.g. judges feel more comfortable using comparisons, which foster more accurate judgments [23],

[1] ordered on increasing level of ecological validity [21, 17]

[2] defined as the extent to which scores are appropriate for their intended interpretation and use [36, 54].

[3] the extend to which a measure would give us the same result over and over again [54], i.e. measure something, free from error, in a consistent way.

[4] the lack of *intra-rater reliability* happens when the listener rates the same speech recording (representation) after a time lapse, and does not arrive at exactly the same score. On the other hand, the lack of *inter-rater reliability* happens if two listeners, who independently rate the same representation, does not arrive to the same score [54].

it does not require a high level of expertise [36, 6], it encourages to tackle hard to operationlize or open-ended tasks [43, 44, 36], and the measurement of competencies [59], among others.

2 Research questions

Considering the previous, this proposal seeks to investigate CJ as a SR method. First, we want to investigate if CJ can be applied to the field of speech research. More specifically, we want to know if CJ can be used to assess children's SI. Second, we seek to prove *how valid, reliable and time efficient are the CJ methods to judge SI, compared to HJ*. More specifically, we seek to compare the HJ versus the dichotomous and ordinal versions of the CJ procedure (CJ-D and CJ-O, respectively).

3 Design

3.1 Judgments and transcriptions

3.1.1 Assumptions

On the one hand, HJ methods have their assumptions rooted in the Classical Test Theory (CTT), where an individual's observed score is composed of a ”true score”, and a random measurement error. Moreover, the true score is defined as the expected value of the score under an infinite number of independent test administrations[5].

On the other hand, CJ methods hinges on two principles: the law of comparative judgments [53], and the consensus of judges [36]. Under the former, the outcome of a comparison, i.e. a relation of preference, is determined by the perceived difference between the discriminal processes of pairs. A *discriminal process*, is the assumed physiological impact that a stimulus has on a listener. However, since this impact cannot be measured directly, we are forced to make some assumption about such process. The minimal assumption we can make is that the process' ordering on the psychological continuum, is the same as the stimulus' ordering that cause them. Moreover, as frequently observed in the field of psycho-physics, since the relationship between stimulus and its impact is not one-to-one, we are also forced to assume the impact has a dispersion/variability, called the *discriminal dispersion*[6]. Finally, the latter principle indicates the shared consensus across judges adds to the validity of the method [36]. This claim is supported by the fact that different listeners differ in the focus and broadness of their judgments [36], and that each representation is assessed by multiple judges, implying the final score is a reflection of the judges' collective expertise [44]. This only means that by combining various judgments, we come closer to the ”true” rankings of SI [35].

3.1.2 Procedures

The HJ procedure consisted on two psycho-linguistic[7] stages: (1) select and mentally represent the stimulus' information, independent of other stimuli[8], and (2) rate the representation, ac-

[5] the National Council of Measurement in Education (NCME) Assessment Glossary: `https://www.ncme.org/resources/glossary`

[6] for a detailed explanation of the law, see Thurstone [53] and Verhavert [59] (p. 22-29)

[7] science concerned with human language production, comprehension, and acquisition [37].

[8] assumption that is not usually met due to anchoring bias (see section 3.1.3).

cording to a task. Therefore, under this procedure, listeners rate the stimulus' SI in an absolute manner, with an 100-point scale going from "very unintelligible" (0) to "very intelligible" (100) [7, 18].

In contrast, CJ is composed of three interrelated psycho-linguistic stages: (1) select and mentally represent the information of the pairs, (2) compare and weigh their relevant information, and (3) rate which representation is preferred, according to a task [56]. As a result, in CJ-D, the listeners rate a pair of stimuli in a dichotomous way, i.e. if stimulus A is more intelligible than B you observe a one in the outcome variable, and zero otherwise [9]. On the other hand, under CJ-O, the listeners rate both stimuli on a 5-point ordinal scale[9] where the outcome variable maps to the following preference relationships: $A >> B$, $A > B$, $A = B$, $A < B$, $A << B$, where $>>$, $>$, $<<$, $<$, and $=$ symbols indicate the level of preference and indifference between the pairs, respectively [55, 1].

3.1.3 Experimental settings

On both procedures, the experimental settings for the **judgment task** followed the next steps [6, 7]:

1. the listener take a seat in front of a computer screen, located at his(her) home, workplace, or the experimental laboratory.
2. the listener open Comproved[10] and select the rating task.
3. the listener read two set of instructions presented on the computer screen about:
    (a) how to perform the task, and
    (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 rate which stimulus sounded more intelligible by selecting the appropriate button, for CJ-D and CJ-O tasks, or a score from a slider on the computer screen, for the HJ task.
6. the listener provide a decision statement on why the selected stimulus sounded more intelligible.

Figure 1: Slider for the HJ task. Extracted from Boonen et al. [7].

Observational evidence indicate the HJ procedure might suffer from anchoring effects[11] or issues with the default option of the slider. About the former, the anchoring seem to happen

[9]evidence on the quality, reliability, and validity benefits of a 5-point scale can be found in Revilla et al. [47].

[10]software developed by the University of Antwerp designed to perform comparative judgments: `https://comproved.com/en/a`.

[11]a bias in decision that occurs when people anchor their decisions around a reference point, and adjust their choices relative to it [4, 30].

when listeners consider the previous assessment as a reference point for the next, effectively turning the task into a comparative rating, similar to CJ. About the latter, as the default setting for the slider is located on the far left for each new assessment (as seen in Figure 1), it is likely that such setting might impact the rating procedure[12]. Considering the previous, in order to minimize both issues, care is taken to randomize the display of stimuli within each listener. However, the researcher recognizes that a better approach to face the second problem would be to randomize the default setting of the slider, but this will not be applied nor investigated on the current research.

talk about decision statements or thinking-at-loud tasks.

Finally, for the **transcription task**, the followed steps were similar to the previous task. Although, at the fourth step, the listeners did not rated the stimulus but rather wrote their orthographic transcriptions, in a free text field in the Comproved environment.

3.2 Children

Thirty two (32) 7-year old children are selected using a large corpus of *spontaneously spoken speech*, collected by CLiPS over the last twenty years. The selection followed a two step procedure, similar to one outlined in Faes et al. [18]. First, a convenient sample of hearing impaired children is selected. Second, a matched sample of normal hearing children is selected.

For the first step, a convenient sample of 10 hearing impaired children with cochlear implant (HI/CI), and 10 hearing impaired children with hearing aids (HI/HA) is selected. The selection is based on the quality of their registered stimuli (utterances), as it is defined as in Section 3.3.

For the second step, 12 normal hearing children (NH) are matched on gender, age, and regional background, to the groups selected in the previous step. The matching is done through a Manual or Propensity Score Matching (PSM) procedure, explain the appropriate procedure.

Finally, the researcher considers important to highlight two relevant points from the children's selection process. First, while the matching procedure for the NH group uses the child's *age* (at recording), the method cannot use the same variable for the other two groups. This is due to the fact that *age* is merely used as a proxy, for the amount of time a child has been developing his(her) language. In that sense, more appropriate variables to use under the HI/CI and HI/HA groups would be e.g. the *device length of use*, which approximates the "hearing age" of such children, or their *vocabulary size*, which resembles their "lexical age" [18]. For this research, we consider the *device length of use* as the simplest one to implement. Second, due to the nature of the sample selection procedure, we cannot ensure the HI/CI and HI/HA, nor the NH group, are representative of their respective populations. Therefore, inferences beyond this particular set of children must be taken with care.

3.3 Stimuli

The stimuli consisted of the children's utterances (sentences of similar length) recovered from previously mentioned CLiPS corpus. More specifically, we use a portion of the corpus that consisted of 10 utterances recordings, for each of the 32 selected children. The stimuli were documented when the child was telling a story cued by the picture book "Frog, where are you" [38] to a caregiver "who does not know the story". The quality of the stimuli was ensured by

[12] compelling evidence on how default settings impact several decision process can be found in Kahneman [30] and Johnson and Goldstein [29].

selecting utterances with no syntactically ill-formed or incomplete sentences, any background noise, cross-talk, long hesitations, revisions or non-words [7].

As a result, the data set consisted in a total of 320 utterances[13] presented to the listeners in a random order, based on the adaptive pairing algorithm [44] implemented in Comproved[14].

Similar designs were used by Boonen et al. [6] and Faes et al. [18]. However, in the former case the number of samples were low, while in the latter, the design was unbalanced and not conducive to appropriate inferences from the pairwise comparisons.

3.4 Comparisons / assessments

In terms the number of comparison per representation (stimulus) required for CJ, Verhavert [59] provided compelling evidence that between 17 and 20 comparisons were enough to achieved a reliable score, measured by the Scale Separation Reliability ($SSR = 0.80$). The current research uses the higher end of such values (20).

On the other hand, based on [source] only 5 assessments per representation were required under the HJ method, to achieve what?. Therefore, we use the same number of assessments under HJ[15].

3.5 Judges and transcribers

The generation of the ratings required the participation of 180 judges (listeners). The judges were students from the Toegepaste Taalkunde bachelor or from the Taal- en Letterkunde master's degree. On both cases, the judges participated in the procedure as part of their course credit. Since we expected the CJ tasks to be 4-times more demanding, in terms of time and effort, than the HJ task, we decided to allocate 4-times more judges to such task. Table 1 describes the judge allocation, the total number of judgments, and the number of judgments per judge.

	Method	Number Utterances	Number (per stimuli) assessments	comparisons	Total judgments	Number judges	Judgments per judge
1	CJ-D	320	n.a.	20	6400	80	80
2	CJ-O	320	n.a.	20	6400	80	80
3	HJ	320	5	n.a.	1600	20	80

n.a.= not applicable

Table 1: Design to rate 320 stimuli per SR method.

[13] under the Design of Experiments (DoE) literature, we would say we have 32 experimental units with 10 replicate runs each, making a total of 320 experimental runs. As it is defined in Lawson [34], an experimental unit is the item under study upon which something is changed, while a replicate run is the experiment conducted with the same factor settings, but using different experimental units.

[14] evidence suggest that the number of comparisons and pairing algorithm impacts the reliability, validity and efficiency of the procedure [11, 12, 36, 60]. However, this is not investigated in the current research.

[15] under DoE literature, this implies we will have 20 and 5 duplicates for each replicate run, under the CJ and HJ procedures, respectively. As defined in Lawson [34], duplicates are repeated measurements of the same experimental unit from one run, where it is possible the measured dependent variable vary among duplicates due to measurement error.

On the other hand, for the transcription task, 100 transcribers participated in the experiment. The participants and stimuli were divided into five groups, where each group of 20 students $(100/5)$ transcribed 64 stimuli on their series $(320/5)$, resulting in 20 transcriptions per utterance $(64 \times 100/320)$. In total we registered 6400 transcriptions[16].

4 Statistical analysis

4.1 Data

4.1.1 Outcomes

On the one hand, the outcome for the **judgment task** was obtained following the procedure outlined in sections 3.1.2 and 3.1.3, with the total number of judgments per procedure detailed in Table 1.

On the other hand, the outcome from the **transcription task** was obtained following a two step procedure [7]. 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[16] (see Section 3.5). Lastly, we computed the entropy measure of the aligned transcriptions as in Shannon [50]:

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

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.

Entropy was used as an objective measure of SI, i.e. a quantification of (dis)agreement between listeners' transcriptions. Utterances yielding a high degree of agreement between transcribers were considered highly intelligible, and therefore registered a lower entropy $(H \to 0)$. In contrast, utterances yielding a low degree of agreement were considered as exhibiting low intelligibility, and therefore registered a higher entropy $(H \to 1)$ [7, 18].

Using Table 2, we exemplify the entropy calculation for utterances 2, 4 and 5, which represent relevant scenarios for the procedure. Notice that every calculation considers 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:

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

[16]under DoE literature, the design corresponds to 32 experimental units with 10 replicates each, making a total of 320 experimental runs. Moreover, we register 20 duplicates (transcriptions) for each run, making a total of 6400 transcriptions.

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

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

Table 2: Example of five aligned transcriptions and its corresponding entropy calculations. Extracted from Boonen et al. [7], and slightly modified with illustrative purposes.

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:

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

4.1.2 Covariates

The characteristics of the selected children is detailed in Table 3 from Appendix A.1. The table includes all the variables used for the matching procedure in Section 3.2, and additionally, shows the child's etiology, i.e. the cause of their hearing impairment, and their post-implant pure tone average (PTA), i.e. the child's subjective hearing sensitivity, aided and unaided by their hearing apparatus. No other variables are included, as no known additional comorbidities, beside their hearing impairment, is suspected.

From the table, describe summaries from the table.

4.1.3 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. [6] decided to eliminate "outlying" observations based on misfit analysis [36], while van Daal [56] and Boonen et al. [7] 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 [43, 44]. Moreover, they have been proven to be less sensitive, as they are calculated with a low number of judgments per representation [43].

On the other hand, for the case of univariate outlier analysis, we argue that outlying observations are interesting cases to analyze [40], and usually they cannot be identified properly outside the context of a full model [40], 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 4.4), the researcher would rather make the model robust against their influence, playing on the strengths of the bayesian framework, than to eliminate the observations.

4.2 Model estimation

The models proposed in sections 4.5 and 4.4 will be estimated under the Bayesian framework[17]. More specifically, we will use the No-U-Turn Hamiltonian Monte Carlo algorithm (No-U-Turn HMC) [5, 16, 25, 42]. `Stan` [52] will be the software package that will provide us with the No-U-Turn HMC machinery, while `R` [45] and its integration packages [51], the software that will allow us to analyze its outputs.

4.3 Model selection

Following the successful and comprehensive analysis in van Daal [56] and Lesterhuis [36], the current research will also use the Information-Theoretic Approach (ITA) [3, 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 [27]. 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

[17] see Rivera [48] (p. 11-13, 15-27) for a detailed description of its benefits and shortcomings.

variable relationships can provide valid causal inferences [24, 40]. 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) [61], and the Pareto-smoothed importance sampling cross-validation (PSIS) [58][18]. Two reasons justify our decision. First, both criteria allow us to embrace the full flexibility and information of our bayesian implementation (outlined in Section 4.4). Last, and more important, both criteria provide us with the best approximations for the out-of-sample (cross-validated) deviance [40]. The deviance is the best approximation for the Kullback-Liebler (KL) divergence [33], i.e. a measure of how far a model is from describing the *true* distribution of our data. McElreath [40] 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[19].

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

4.4 Models and hypothesis

Considering the objectives outlined in Section 2, we will evaluate seven models that serve different, but interrelated, objectives:

(1) a hierarchical beta model for entropy

(2) a measurement model for:

   (a) absolute holistic judgment (HJ)

   (b) dichotomous comparative judgment (CJ-D)

   (c) ordinal comparative judgment (CJ-O)

(3) a full integrating model for:

   (a) models (1) and (2a)

   (b) models (1) and (2b)

   (c) models (1) and (2c)

4.4.1 Hierarchical beta model

Previous research already used hierarchical models with the replicated entropy measures as outcomes [7, 18]. 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 [40].

[18] van Daal [56] used the Akaike's Information Criterion (AIC) [2] with similar purposes.

[19] see McElreath [40] (p. 202-211) for the intuition and detailed derivation of the argument.

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. [7], and it is what justify our use of a hierarchical model. More specifically, we will use a Hierarchical (Mixed) Beta Regression model [19], for which we argue, its implementation is rather trivial under the bayesian framework, and we present it in the following lines.

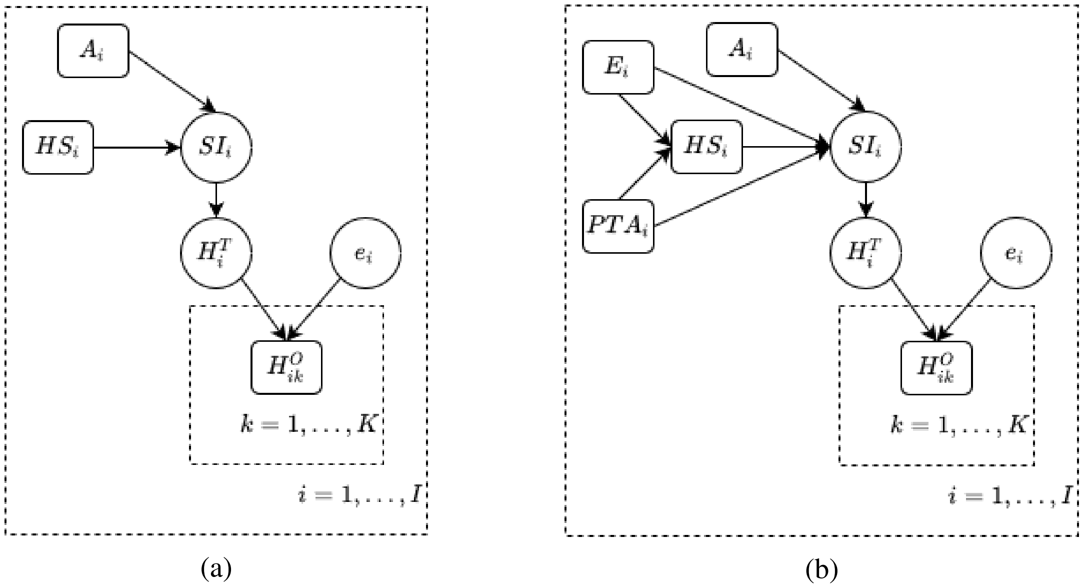

Figure 2: DAG for the hierarchical beta regression model for entropy. (a) *total effects* of hearing status, (b) *direct effects* of hearing apparatus. Circles represent latent variables, squares observed values or covariates, and large discontinuous squares the nesting within specific units.

First, figure 2 depicts the DAG representation of the model. For the measurement error part, section 4.1.1 reveals the (observed) entropy replicates H_{ik}^O can represent multiple realizations of a child's *true* entropy H_i^T, measured with error e_i. As a result, we can say the k'th entropy measure is nested within the i'th child, where $k = 1, \dots, K$, $i = 1, \dots, I$, $K = 10$ utterances, and $I = 32$ children.

Second, for the hypothesis part, we can say the child's *true* entropy H_i^T 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 2, 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[20] 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 [8].

[20]see section 3.2 to know how the variable is defined.

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 [32]: (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 3.2), the need to control for such characteristics is paramount, if we seek to obtain effects that can generalize better and beyond our sample[21].

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:

$$\begin{aligned}
&\text{Likelihood:} \\
&H^O_{ik} \sim \text{BetaProp}\left(H^T_i, M_i\right) && (2) \\
&\text{Transformed parameters:} \\
&H^T_i = \text{logit}^{-1}(-SI_i) && (3) \\
&\text{Linear predictor:} \\
&SI_i = a_i + \alpha + \alpha_{HS[i]} + \beta_{A,HS[i]}(A_i - \bar{A}) + \alpha_{E[i]} + \beta_P PTA_i && (4) \\
&\text{Priors:} \\
&M_i \sim \text{LN}(\mu_M, \sigma_M) && (5) \\
&a_i \sim \text{N}(\mu_a, \sigma_a) && (6) \\
&\alpha \sim \text{N}(0, 0.5) && (7) \\
&\alpha_{HS[i]} \sim \text{N}(0, 0.5) && (8) \\
&\beta_{A,HS[i]} \sim \text{N}(0, 0.3) && (9) \\
&\alpha_{E[i]} \sim \text{N}(0, 0.5) && (10) \\
&\beta_P \sim \text{N}(0, 0.3) && (11) \\
&\text{Hyper-priors:} \\
&\mu_M \sim \text{N}(0, 5) && (12) \\
&\sigma_M \sim \text{Exp}(1) && (13) \\
&\mu_a \sim \text{N}(0, 0.5) && (14) \\
&\sigma_a \sim \text{Exp}(1) && (15) \\
& && (16)
\end{aligned}$$

[21] follow the *notes* folder, to see a graphical though experiment.

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.

Four 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 e_i, we use the latent "sample size" parameter M_i to model the heterogeneity/variability of the duplicate entropies. This effectively works as a measurement error model for the duplicates, as the parameter defines the shape of the distribution. Third, we use mildly informative priors to state our uncertainty regarding the direction and magnitude of the effects[22]. Fourth, if we do not consider etiology and PTA values in equation (4), we obtain the panel (a) model.

4.4.2 Measurement models

(in process)

4.4.3 Integration models

(in process)

4.5 Model evaluation

As described in Section 4.4, all seven models serve different but interrelated objectives.

4.5.1 Objective ranking

As expected, model (1) allow us to construct the *most objective* measure of a child's SI ranking. Moreover, using such model we will be able to test some research hypothesis of our interest.

4.5.2 Reliability

The two remaining sets of models allow us to assess reliability in different ways. Under the second set of models (2a, 2b, 2c), we will be able to assess two forms of the inter-rater reliability: (a) the judges's reliability (JSR), and (b) the scale reliability, also known as the Scale Separation Reliability (SSR) [11].

Following Fisher [20] and Wright [63] the measures were defined as follows:

$$JSR = \frac{G_J^2}{(1+G_J^2)}, \quad G_J = \frac{\sigma_\alpha}{\sigma_J} \tag{17}$$

$$SSR = \frac{G_\alpha^2}{(1+G_\alpha^2)}, \quad G_\alpha = \frac{\sigma_\alpha}{\sum_{i=1}^{C} se_{\alpha,i}} \tag{18}$$

(in process)

[22] see Rivera [48] (p. 18-19) for an intuition on prior elicitation.

4.5.3 validity

the direction, magnitude, and quality of inference of the same research hypothesis outlined for Model (1), in the newly estimated models.

(in process)

4.5.4 time efficiency

(in process)

4.5.5 statistical efficiency

(in process)

A Appendix: Tables

A.1 Children characteristics

Child	Hearing Status	Gender	Regional background	Age (y;m)	Device use (y;m)	Etiology	PTA (dB.)	
							unaided	aided
1	NH	male				genetic		
2	HI/CI	female				CMV infection		
3	HI/HA					unknown		
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								

(y;m) = (years;months)
NH = normal hearing,
HI/CI = hearing impaired / cochlear implant,
HI/HA = hearing impaired / hearing aid

Table 3: Characteristics of selected children.

Bibliography

[1] Agresti, A. [1992]. Analysis of ordinal paired comparison data, *Journal of the Royal Statistical Society* **41**(2): 287–297.
doi: https://doi.org/10.2307/2347562.

[2] 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.

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

[4] Baddeley, M. [2017]. *Behavioural Economics: A Very Short Introduction*, Oxford University Press.

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

[6] 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.

[7] 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.

[8] 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.

[9] Bradley, R. and Terry, M. [1952]. Rank analysis of incomplete block designs: I. the method of paired comparisons, *Biometrika* **39**(3-4): 324–345.
doi: https://doi.org/10.2307/2334029.

[10] Bramley, T. [2008]. Paired comparison methods, *in* P. Newton, J. Baird, H. Goldsteing, H. Patrick and P. Tymms (eds), *Techniques for monitoring the comparability of examination standards*, GOV.UK., pp. 246—-300.
url: https://www.gov.uk/government/publications/techniques-for-monitoring-the-comparability-of-examination-standards.

[11] Bramley, T. [2015]. Investigating the reliability of adaptive comparative judgment. Cambridge Assessment Research Report.
url: http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf.

[12] Bramley, T. and Vitello, S. [2018]. The effect of adaptivity on the reliability coefficient in adaptive comparative judgement, *Assessment in Education: Principles, Policy and Practice* **26**(1): 43–58.
doi: https://doi.org/10.1080/0969594X.2017.1418734.

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

[16] 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.

[17] 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).

[18] 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).

[19] 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.

[20] Fisher, W. [1992]. Reliability, separation, strata statistics. Rasch Measurement Transactions.
url: https://rasch.org/rmt/rmt63i.htm.

[21] 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.

[22] 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.

[23] Gill, T. and Bramley, T. [2013]. How accurate are examiners' holistic judgements of script quality?, *Assessment in Education: Principles, Policy and Practice* **20**: 308—-324.
doi: https://doi.org/10.1080/0969594X.2013.779229.

[24] 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.

[25] 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.

[26] 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$_JSLHR - 20 - 00008$.

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

[28] Johannisson, T., Lohmander, A. and Persson, C. [2014]. Assessing intelligibility by single words, sentences and spontaneous speech: A methodological study of speech production of 10-year-olds, *Logopedics Phoniatrics Vocology* **39**: 159–168.
doi: https://doi.org/10.3109/14015439.2013.820487.

[29] Johnson, E. and Goldstein, D. [2003]. Do defaults save lives?, *Science* **302**(5649): 1338–1339.
doi: https://doi.org/10.1126/science.1091721.
url: https://www.science.org/doi/abs/10.1126/science.1091721.

[30] Kahneman, D. [2013]. *Thinking Fast and Slow*, Farrar, Straus and Giroux.

[31] 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.

[32] 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.

[33] 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.

[34] Lawson, J. [2018]. *Design and Analysis of Experiments with R*, Chapman and Hall/CRC.

[35] Lee, M., Steyvers, M. and Miller, B. [2014]. A cognitive model for aggregating people's rankings, *Plos One* **9**: 1–9.
doi: https://doi.org/10.1371/journal.pone.0096431.

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

[37] Levelt, W. [1993]. *Speaking: From Intention to Articulation*, The MIT Press.
doi: https://doi.org/10.7551/mitpress/6393.001.0001.

[38] 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.

[39] McDaniel, D. and Cox, R. [1992]. Evaluation of the speech intelligibility rating (sir) test for hearing aid comparisons, *Journal of Speech and Hearing Research* **35**(3): 686–693.
doi: https://doi.org/10.1044/jshr.3503.686.

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

[41] McLeod, S., Harrison, L. and McCormack, J. [2012]. The intelligibility in context scale: Validity and reliability of a subjective rating measure, *Journal of Speech, Language and Hearing Research* **55**: 648–656.
doi: https://doi.org/10.1044/1092-4388(2011/10-0130).

[42] 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.

[43] 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.

[44] 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.

[45] 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/.

[46] Raeve, L. [2010]. A longitudinal study on auditory perception and speech intelligibility in deaf children implanted younger than 18 months in comparison to those implanted at later ages, *Otology and Neurotology* **31**(8): 1261–1267.
doi: https://doi.org/10.1097/MAO.0b013e3181f1cde3.

[47] Revilla, M., Saris, W. and Krosnick, J. [2014]. Choosing the number of categories in agree–disagree scales, *Sociological Methods and Research* **43**(1): 73–97.
doi: https://doi.org/10.1177/0049124113509605.

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

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

[50] 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.

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

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

[53] Thurstone, L. [1927]. A law of comparative judgment, *Psychological Review* **34**(4): 482–499.
doi: https://doi.org/10.1037/h0070288.

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

[55] Tutz, G. [1986]. Bradley-terry-luce model with an ordered response, *Journal of Mathematical Psychology* **30**(3): 306–316.
doi: https://doi.org/10.1016/0022-2496(86)90034-9.

[56] 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.

[57] 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.

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

[59] Verhavert, S. [2018]. *Beyond a Mere Rank Order: The Method, the Reliability and the Efficiency of Comparative Judgment*, PhD thesis, University of Antwerp.

[60] Verhavert, S., Bouwer, R., Donche, V. and De Maeyer, S. [2019]. A meta-analysis on the reliability of comparative judgement, *Assessment in Education: Principles, Policy and Practice* **26**(5): 541–562.
doi: https://doi.org/10.1080/0969594X.2019.1602027.

[61] 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.

[62] 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.

[63] Wright, B. [1996]. Reliability and separation. Rasch Measurement Transactions.
url: https://rasch.org/rmt/rmt63i.htm.