Everything, altogether, all at once: Addressing data challenges when measuring speech intelligibility through entropy scores

Jose Manuel Rivera Espejo[1], **Sven De Maeyer**[1], **Steven Gillis**[2]

[1]Education sciences, University of Antwerp,
[2]Linguistics, University of Antwerp,

Corresponding author: Jose Manuel Rivera Espejo, JoseManuel.RiveraEspejo@uantwerpen.be

Abstract

Considering the imperative need to comprehensively address all data features when investigating unobservable and complex traits, this research aims to showcase the effectiveness of the Generalized Linear Latent and Mixed Model (GLLAMM) (Rabe-Hesketh et al., 2004a, 2004c, 2004b; Skrondal & Rabe-Hesketh, 2004) in handling entropy scores for investigating speech intelligibility theories. Utilizing transcriptions from spontaneous speech data originally collected by Boonen et al. (2021), the Bayesian Beta-proportion GLLAMM was employed to model the resulting entropy scores. The study compared the model's prediction accuracy with the Normal Linear Mixed Model (LMM) (Holmes et al., 2019) and investigated its capacity to estimate a latent intelligibility from manifest entropy scores. Additionally, it illustrated how the model can explore research theories concerning the impact of speaker-related factors on intelligibility. Results demonstrate the consistent superiority of the Beta-proportion GLLAMM over the Normal LMM in predicting the empirical phenomena. Furthermore, the model effectively quantified the latent potential intelligibility, enabling the ranking and comparison of individuals while accommodating for uncertainties. Lastly, it facilitated exploration of theories related to speaker-related factors and intelligibility. Despite these advantages, the introduction of these innovative statistical tools poses challenges for researchers seeking implementation. Nevertheless, the study suggests interesting future research directions, including power analysis, causal hypothesis formulation, and exploration of novel methods for intelligibility assessment. This study has implications for researchers and data analysts interested in quantitatively measuring and testing theories related to intricate, unobservable constructs, while emphasizing the accurate prediction of empirical phenomena.

1 Introduction

Intelligibility is at the core of successful, felicitous communication. Thus, being able to speak intelligibly is a major achievement in language acquisition and development. Moreover, intelligibility is considered to be the most practical index to assess competence in oral communication (R. D. Kent et al., 19943). Consequently, it serves as a key indicator for evaluating the effectiveness of various interventions like speech therapy or cochlear implantation (Chin et al., 2012).

The notion of speech intelligibility may appear deceptively simple, yet it is an intricate concept filled with inherent challenges in its assessment. Intelligibility refers to the extent to which a listener can accurately recover the elements in a speaker's acoustic signal, such as phonemes or words (Freeman et al., 2017; van Heuven, 2008; Whitehill & Chau, 2004). Furthermore, achieving intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered (Freeman et al., 2017). Hence, it is unsurprising that its accurate measurement faces challenges (R. Kent et al., 1989). These challenges arise from the interplay of its determinants, which encompass attributes of the communicative environment, such as background noise (Munro, 1998), speaker features like speaking rate (Munro & Derwing, 1998) or accent (Jenkins, 2000; Ockey et al., 2016), and listener characteristics like vocabulary proficiency or hearing ability (Varonis & Susan, 1985).

While several approaches have been proposed to assess intelligibility, they commonly rely on two types of speech samples: read-aloud or imitated, and spontaneous speech samples. Most studies favor read-aloud or imitated speech samples due to the substantial control they offer in selecting stimuli for intelligibility assessment. Additionally, these types of speech facilitate a direct and unambiguous comparison between a defined word target, produced by a speaker, and the listener's identification of it, as exemplified by multiple studies such as Castellanos et al. (2014), Chin et al.

(2012), Chin & Kuhns (2014), Freeman et al. (2017), Khwaileh & Flipsen (2010), and Montag et al. (2014). However, it has been demonstrated that these controlled speech samples exhibit limited efficacy in predicting intelligibility among hearing-impaired individuals (Cox et al., 1989; Ertmer, 2011). In contrast, spontaneous speech samples offer a more ecologically valid means for assessing intelligibility, as they resemble everyday informal speech more compared to read-aloud or imitated speech (Boonen et al., 2021). However, establishing a straightforward comparison between a predetermined word target and a listener's identification of it using spontaneous speech is no longer possible, since such a target is non-existent. Therefore, the link between a word target and a listener's identification of it can only be inferred indirectly (Flipsen, 2006; Lagerberg et al., 2014).

Yet, a metric of intelligibility can still be derived from transcriptions of spontaneous speech samples. In this approach, listeners transcribe orthographically multiple spontaneous speech samples produced by various speakers. These transcriptions are then aggregated into entropy scores, where lower scores indicate a higher degree of agreement among the listener's transcriptions and, consequently, higher intelligibility, while higher scores suggest lower intelligibility due to a lower degree of agreement in the transcriptions (Boonen et al., 2021; Faes et al., 2021). Notably, the aggregation procedure assumes that speech samples are considered "intelligible" if all listeners decode them in the same manner. These scores have been instrumental in examining differences in speakers' speech intelligibility, particularly between children with normal hearing and those with cochlear implants (Boonen et al., 2021).

However, despite their potential as a fine-grained metric of intelligibility, as proposed by Boonen et al. (2021), they exhibit a statistical complexity that cautions researchers against treating them as straightforward indices of intelligibility. This complexity emerges from the processes of data collection and transcription aggregation, endowing the scores with four distinctive features: boundedness, measurement error, clustering, and the possible presence of outliers and heteroscedasticity. Firstly, entropy scores are confined to the interval between zero and one, a phenomenon known as boundedness. Boundedness refers to the restriction of data values within specific bounds or intervals, beyond which they cannot occur (Lebl, 2022). Secondly, entropy scores are a manifestation of a speaker's intelligibility, with this intelligibility being the primary factor influencing the observed scores. This issue is commonly referred to as measurement error, signifying the disparity between the observed values of a variable, recorded under similar conditions, and some fixed *true value* which is not directly observable (Everitt & Skrondal, 2010). Thirdly, due to the repeated assessment of speakers through multiple speech samples, the scores exhibit clustering. Clustering occurs when outcomes stem from repeated measurements of the same individual, location, or time (McElreath, 2020). Lastly, driven by the specific small set of speakers and speech samples under scrutiny, these scores often display a potential for the presence of outliers and heteroscedasticity. Outliers are observations that markedly deviate from other sample data points in which they occur, while heteroscedasticity occurs when the outcome's variance depends on the values of another variable (Everitt & Skrondal, 2010).

Failure to collectively address these data features can result in numerous statistical challenges that might hamper the researcher's ability to investigate intelligibility. Notably, neglecting boundedness can, at best, lead to underfitting and, at worst, to misspecification. Underfitting occurs when statistical models fail to capture the underlying data patterns, potentially causing the generation of predictions outside the data range, thus hindering the model's ability to generalize when confronted with new data. Conversely, misspecification, marked by a poor representation of relevant aspects of the true data in the model's functional form, can lead to inconsistent and less precise parameter estimates (Everitt & Skrondal, 2010). Additionally, overlooking issues such as measurement error, clustering, outliers, or heteroscedasticity can lead to biased and less precise parameter estimates (McElreath, 2020), ultimately diminishing the statistical power of models and increasing the likelihood of committing type I or type II errors when addressing research inquiries.

In the realm of computational statistics and data analysis, several models have been developed to address some of these data features individually and, at times, collectively. For instance, Ferrari & Cribari-Neto (2004) and Simas et al. (2010) initially introduced and expanded beta regression models to handle outcomes constrained within the unit interval. Subsequently, Figueroa-Zúñiga et al. (2013) extended these models to address data clustering. Over time, beta regression models have evolved to accommodate clustering and measurement errors in covariates, as demonstrated by Carrasco et al. (2012) and Figueroa-Zúñiga et al. (2018). Furthermore, robust versions of these models have been proposed to account for other statistical data issues, such as outliers and heteroscedasticity, as seen in Bayes et al. (2012) and Figueroa-Zúñiga et al. (2021). Robust models are a general class of statistical procedures designed to reduce the sensitivity of the parameter estimates to mild or moderate departures of the data from the model's assumptions (Everitt & Skrondal, 2010). Ultimately, the work of Rabe-Hesketh and colleagues introduced the Generalized Linear Latent and Mixed Model (GLLAMM) (Rabe-Hesketh et al., 2004a, 2004c, 2004b; Skrondal & Rabe-Hesketh, 2004), a unified framework that can simultaneously tackle with all of the aforementioned data features.

All of these models have found moderate adoption in various fields, including speech communication (Boonen et al., 2021), psychology (Unlu & Aktas, 2017), cognition (Lopes et al., 2023; Verkuilen & Smithson, 2013), education (Pereira et al., 2020), health care [Ghosh (2019); Kangmennaang_et_al_2023], chemistry (de Brito Trindade et al., 2021), and policy analysis (Choi, 2023; Dieteren et al., 2023; Zhang et al., 2023). Specifically, in the domain of speech communication, Boonen et al. (2021) addressed data clustering within the context of intelligibility research. Conversely, de Brito Trindade et al. (2021) and Kangmennaang et al. (2023) concentrated on tackling non-normal bounded data with measurement error in covariates, within the context of chemical reactions and health care access, respectively. Remarkably, despite these individual efforts, there is, to the best of the authors' knowledge, no study comprehensively addressing all of these data features in a principled way while also transparently and systematically documenting the Bayesian estimation of the resulting statistical models.

1.1 Research questions

Considering the imperative need to comprehensively address all data features when investigating unobservable and complex traits, this investigation aims to demonstrate the efficacy of the Generalized Linear Latent and Mixed Model (GLLAMM) in handling entropy scores features when exploring research theories concerning speech intelligibility. To achieve this objective, the study will reexamine data originating from transcriptions of spontaneous speech samples, initially collected by Boonen et al. (2021). Subsequently, this data will be aggregated into entropy scores and subjected to modeling through the Bayesian Beta-proportion GLLAMM.

To address the primary objective, the study poses three key research questions. First, given the importance of accurate predictions in developing useful practical models and testing research theories (Shmueli & Koppius, 2011), *Research Question 1 (RQ1)* evaluates whether the Beta-proportion GLLAMM yields more accurate predictions than the widely used Normal Linear Mixed Model (LMM) (Holmes et al., 2019). Second, acknowledging that intelligibility is an unobservable, intricate concept and a key indicator of oral communication competence (R. D. Kent et al., 19943), *Research Question 2 (RQ2)* investigates how the proposed model can estimate speakers' latent intelligibility from manifest entropy scores. Thirdly, recognizing that research involves developing and comparing theories, *Research Question*

3 (RQ3) illustrates how these research theories can be examined within the model's framework. Specifically, RQ3 assesses the influence of speaker-related factors on the newly estimated latent intelligibility.

The findings of this study will equip researchers investigating speech intelligibility using entropy scores, or those grappling with similar data challenges, with a statistical tool that improves upon existing research models. The tool will provide an assessment of the predictability of empirical phenomena, along with the capability to develop a quantitative measure for the latent variable of interest. The latter, in turn, will facilitate the appropriate comparison of existing theories related to the latent variable, and even the development of new ones.

2 Methods

2.1 Data

The data comprised the transcriptions of spontaneous speech samples originally collected by Boonen et al. (2021). The data is not publicly available due to privacy restrictions. Nonetheless, the data can be provided by the corresponding author upon reasonable request.

2.1.1 Speakers

Boonen et al. (2021) selected 32 speakers, comprising 16 normal hearing children (NH) and 16 hearing-impaired children with cochlear implants (HI/CI). At the time of the collection of the speech samples, the NH group were between 68 and 104 months old ($M = 86.3$, $SD = 9.0$), while HI/CI group were between 78 and 98 months old ($M = 86.3$, $SD = 6.7$).

2.1.2 Speech samples

Boonen and colleagues selected speech samples from a large corpus of children's spontaneously spoken speech recordings. These recordings were obtained as the children narrated a story prompted by the picture book "Frog, Where Are You?" (Mayer, 1969) to a caregiver 'unfamiliar with the story'. Before recording, the children were allowed to skim over the booklet and examine pictures. Prior to the selection process, the recordings were orthographically transcribed using the CHAT format in the CLAN editor (MacWhinney, 2020). These transcriptions were exclusively used in the curation of appropriate speech samples. To ensure the quality of the selection, Boonen and colleagues excluded sentences containing syntactically ill-formed or incomplete statements, with background noise, crosstalk, long hesitations, revisions, or non-words. Finally, ten speech samples were randomly chosen for each of the 32 selected speakers. Each of these samples comprised a single sentence with a length of three to eleven words ($M = 7.1$, $SD = 1.1$). The process resulted in a total of 320 selected sentences collectively comprising $2,263$ words.

2.1.3 Listeners

Boonen and colleagues recruited 105 students from the University of Antwerp. All participants were native speakers of Belgian Dutch and reported no history of hearing difficulties or prior exposure to the speech of hearing-impaired speakers.

2.1.4 Transcription task and entropy scores

The 320 speech samples and 105 listeners were randomly assigned to five blocks, with each block consisting of approximately 21 listeners who transcribed 64 sentences presented in random order. This resulted in a total of $47,514$ transcribed words from the original $2,263$ words present in the speech samples. These orthographic transcriptions were automatically aligned with a python script (Boonen et al., 2021), at the sentence level in a column-like grid structure like the one presented in Table 1. This alignment process was repeated for each sentence within each speaker and block, and the output was manually checked and adjusted (if needed) in order to appropriately align the words. For more details on the random assignment and alignment procedures refer to Boonen et al. (2021).

Next, this study aggregated the aligned transcriptions by listener yielding $2,2634$ entropy scores, one score per word. The entropy scores were calculated following Shannon's formula (1948):

$$H_{wsib} = \frac{\sum_{k=1}^{K} p_k \cdot log_2(p_k)}{log_2(J)} \tag{1}$$

where H_{wsib} denotes the entropy scores confined to an interval between zero and one, with w defining the word index, s the sentence index, i the speaker index, and b the block index. Moreover, K describes the number of different word types within transcriptions, and J defines the total number of word transcriptions. Notice that by design, the total number of word transcriptions J corresponds with the number of listeners per block, i.e., 21 listeners. Lastly, $p_k = \sum_{j=1}^{J} 1(T_{jk})/J$ denotes the proportion of word types within transcriptions, with $1(T_{jk})$ describing an indicator function that takes the value of one when the word type k is present in the transcription j. See Section 6.1 for a calculation example of entropy scores.

These entropy scores served as the outcome variable, capturing agreement or disagreement among listeners' word transcriptions. Lower scores indicated a higher degree of agreement between transcriptions and therefore higher intelligibility, while higher scores indicated lower intelligibility, due to a lower degree of agreement in the transcriptions (Boonen et al., 2021; Faes et al., 2021). Furthermore, no score is excluded from the modeling process using univariate procedures, rather, the identification of highly influential observations is performed within the context of the proposed models, as recommended by McElreath (2020).

Table 1: Hypothetical alignment of word transcriptions and entropy scores. Note: Extracted from Boonen et al. (2021), and slightly modified for illustrative purposes. Entropy scores are calculated the first sentence, produced by the first speaker assigned to the first block, and transcribed by five listeners ($s = 1, i = 1, b = 1, J = 5$). Transcriptions are in Dutch with English translation. *[B]* represent a blank space, and *[X]* an unidentifiable speech.

Transcription Number	Words 1	2	3	4	5
1	de	jongen	ziet	een	kikker
	the	boy	sees	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

2.2 Statistical models

This section articulates the probabilistic formalism of both the Normal LMM and the proposed Beta-proportion GLLAMM. Subsequently, it details the set of fitted models and the estimation procedure, along with the criteria employed to assess the quality of the Bayesian inference results. Lastly, the section outlines the methodology employed for model comparison.

The selection of the Bayesian approach was based on three key properties. Firstly, empirical evidence from prior research demonstrates that Bayesian methods outperform frequentist methods, particularly in handling complex and over-parameterized models (Baker, 1998; Kim & Cohen, 1999). This superiority is evident when dealing with complex models, like the proposed GLLAMM, that are challenging to program or are not viable under frequentist methods (Depaoli, 2014). Secondly, the approach allows for the incorporation of prior information, ensuring that certain parameters are confined within specified boundaries. This helps mitigate non-convergence or improper parameter estimation issues commonly observed in complex models under frequentist methods (Martin & McDonald, 1975; Seaman & Stamey, 2011). In this study, for example, this property was leveraged to incorporate information about the variances of random effects and constrain them to be positive. Lastly, the Bayesian approach demonstrates proficiency in handling relatively small sample sizes (Baldwin & Fellingham, 2013; Depaoli, 2014; Lambert et al., 2006). In this case, despite the study dealing with 2,263 entropy scores, these were derived from a modest sample size of 32 speakers, from whom the inferences are drawn. Consequently, reliance on the asymptotic properties of frequentist methods may not be warranted in this context, underscoring the pertinence of this property to the current study.

2.2.1 Normal LMM

The general mathematical formalism of the Normal LMM posits that the likelihood of the (manifest) entropy scores follow a normal distribution, i.e.

$$H_{wsib} \sim \text{Normal}\left(\mu_{sib}, \sigma_i\right) \tag{2}$$

where μ_{sib} represents the average entropy at the word-level and σ_i denotes the standard deviation of the average entropy at the word-level, varying for each speaker. Given the clustered nature of the data, μ_{sib} is defined by the linear combination of individual characteristics and several random effects:

$$\mu_{sib} = \alpha + \alpha_{HS[i]} + \beta_{A,HS[i]}(A_i - \bar{A}) + u_{si} + e_i + a_b \tag{3}$$

where HS_i and A_i denote the hearing status and chronological age of speaker i, respectively. Additionally, α denote the general intercept, $\alpha_{HS[i]}$ represents the average entropy for each hearing status group, and $\beta_{A,HS[i]}$ denotes the evolution of the average entropy per unit of chronological age A_i for each hearing status group. Furthermore, u_{si} denotes the sentence-speaker random effects measuring the unexplained entropy variability within sentences for each speaker, e_i denotes the speaker random effects describing the unexplained entropy variability between speakers, and a_b denotes the block random effects assessing the unexplained variability between experimental blocks.

Several notable features of the Normal LLM can be discerned from the equations. Firstly, Equation 2 indicates that the variability of the average entropy at the word level can differ for each speaker, enhancing the model's *robustness* to mild or moderate data departures from the normal distribution assumption, such as in the presence of heteroscedasticity or outliers. Secondly, Equation 3 reveals that the model assumes no transformation is applied to the relationship between the average entropy and the linear combination of characteristics. This is commonly known as a direct link function. Moreover, Equation 3 indicates that chronological age is centered around the minimum chronological age in the sample $\bar{A}$. The *centering* procedure prevents the interpretation of parameters outside the range of chronological ages available in the data (Everitt & Skrondal, 2010). Lastly, the equation implies the model considers separate intercept and separate slope of age for each hearing status group, i.e., NH and HI/CI speakers, $\alpha_{HS[1]}$ and $\alpha_{HS[2]}$, respectively.

2.2.2 Beta-proportion GLLAMM

The general mathematical formalism of the proposed Beta-proportion GLLAMM comprises four components: a response model likelihood, a linear predictor, a link function, and a structural model. The likelihood of response model posits the entropy scores follow a Beta-proportion distribution,

$$H_{wsib} \sim \text{BetaProp}\left(\mu_{ib}, M_i\right) \tag{4}$$

whereμ_{ib} denotes the average entropy at the word-level and M_i signifies the *dispersion* of the average entropy at the word-level, varying for each speaker. Additionally, μ_{ib} is defined as,

$$\mu_{ib} = \text{logit}^{-1}[a_b - SI_i] \tag{5}$$

where $\text{logit}^{-1}(x) = exp(x)/(1 + exp(x))$ is the inverse-logit link function, a_b denotes the block random effects, and SI_i describes the speaker's latent *potential intelligibility*. Conversely, the structural equation model relates the speakers' latent potential intelligibility to the individual characteristics:

$$SI_i = \alpha + \alpha_{HS[i]} + \beta_{A,HS[i]}(A_i - \bar{A}) + e_i + u_i \tag{6}$$

where α defines the general intercept, $\alpha_{HS[i]}$ denotes the potential intelligibility for different hearing status groups, and $\beta_{A,HS[i]}$ indicates the evolution of potential intelligibility per unit of chronological age for each hearing status group. Furthermore, e_i represents speakers block effects, describing unexplained potential intelligibility variability between speakers, and $u_i = \sum_{s=1}^{S} u_{si}/S$ denotes sentence random effects, assessing the average unexplained potential intelligibility variability among sentences within each speaker, with S denoting the total number of sentences per speaker.

Several features are evident in this probabilistic representation. Firstly, akin to the Normal LMM, Equation 4 reveals that the *dispersion* of average entropy at the word level can differ for each speaker. This enhances the model's robustness to mild or moderate data departures from the beta-proportion distribution assumption. Secondly, in contrast with the Normal LMM, Equation 5 shows the potential intelligibility of a speakers has a negative non-linear relationship with the entropy scores, explicitly highlighting the inverse relationship between intelligibility and entropy. This feature also maps the unbounded linear predictor to the bounded limits of the entropy scores. Thirdly, in contrast with the Normal LMM, Equation 6 demonstrates that the structural parameters are interpretable in terms of the latent potential intelligibility scores, where the scale of the latent trait is set by the general intercept α, as required in latent variable models (Depaoli, 2021). Furthermore, the equation implies the model also considers separate intercept and separate slope of age for each hearing status group, i.e., NH and HI/CI speakers ($\alpha_{HS[1]}$ and $\alpha_{HS[2]}$, respectively). Additionally, Equation 6 indicates that chronological age is *centered* around the minimum chronological age in the sample $\bar{A}$. Lastly, the same equation assumes the intelligibility scores have two sources of unexplained variability: e_i and u_i. The former represents inherent differences in potential intelligibility among different speakers, while the latter assumes that different sentences measure potential intelligibility differently due to variations in word difficulties and their interplay within the sentence.

2.2.3 Prior distributions

Bayesian procedures require the incorporation of priors. Priors are probability distributions summarizing the information about known or assumed parameters prior to observing any empirical data (Everitt & Skrondal, 2010). Upon observing empirical data, these priors undergo updating to posterior distributions following Bayes' rule (Jeffreys, 1998). In cases requiring greater modeling flexibility, a more refined representation of the parameters' priors can be defined in terms of hyperparameters and hyperpriors. *Hyperparameters* refer to parameters indexing a family of possible prior distributions for the original parameter, while *hyperpriors* are prior distributions for such hyperparameters (Everitt & Skrondal, 2010).

This study establishes priors and hyperpriors for the parameters of both the Normal LMM and the Beta-proportion GLLAMM using prior predictive simulations. This procedure entails the semi-independent simulation of parameters, which are subsequently transformed into simulated data values according to the models' specifications. This procedure aims to establish meaningful priors and comprehend their implications within the context of the model before incorporating any information derived from empirical data (McElreath, 2020). For reader inspection, the prior predictive simulations are provided in the accompanying digital walk-through document (refer to Section 2.2.7 Open Science Statement).

2.2.3.1 Normal LMM

For the parameters of the Normal LMM, non-informative priors and hyperpriors are established to align with analogous model assumptions in frequentist methods. A *non-informative* prior reflects the distributional commitment of a parameter to a wide range of values within a specific parameter space (Everitt & Skrondal, 2010). The specified priors are as follows:

$$
\begin{aligned}
r_S &\sim \text{Exponential}\,(2) \\
\sigma_i &\sim \text{Exponential}\,(r_S) \\
m_i &\sim \text{Normal}\,(0, 0.05) \\
s_i &\sim \text{Exponential}\,(2) \\
e_i &\sim \text{Normal}\,(m_i, s_i) \\
m_b &\sim \text{Normal}\,(0, 0.05) \\
s_b &\sim \text{Exponential}\,(2) \\
a_b &\sim \text{Normal}\,(m_b, s_b) \\
\alpha &\sim \text{Normal}\,(0, 0.05) \\
\alpha_{HS[i]} &\sim \text{Normal}\,(0, 0.2) \\
\beta_{A,HS[i]} &\sim \text{Normal}\,(0, 0.1)
\end{aligned}
\tag{7}
$$

2.2.3.2 Beta-proportion GLLAMM

For the parameters of the Beta-proportion GLLAMM, weakly informative priors and hyperpriors are established. *Weakly informative priors* reflect the distributional commitment of a parameter to a weakly constraint range of values within a realistic parameter space (McElreath, 2020). The specified priors are as follows:

$$
\begin{aligned}
r_M &\sim \text{Exponential}\,(2) \\
M_i &\sim \text{Exponential}\,(r_M) \\
m_i &\sim \text{Normal}\,(0, 0.05) \\
s_i &\sim \text{Exponential}\,(2) \\
e_i &\sim \text{Normal}\,(m_i, s_i) \\
m_b &\sim \text{Normal}\,(0, 0.05) \\
s_b &\sim \text{Exponential}\,(2) \\
a_b &\sim \text{Normal}\,(m_b, s_b) \\
\alpha &\sim \text{Normal}\,(0, 0.05) \\
\alpha_{HS[i]} &\sim \text{Normal}\,(0, 0.3) \\
\beta_{A,HS[i]} &\sim \text{Normal}\,(0, 0.1)
\end{aligned}
\tag{8}
$$

2.2.4 Fitted models

This study evaluates the comparative predictive capabilities of both the Normal LMM and the Beta-proportion GLLAMM (RQ1) while simultaneously examining various formulations regarding how speaker-related factors influence intelligibility (RQ3). In this context, the predictive capabilities of the models are intricately connected to these formulations. As a result, the study requires fitting 12 different models, each representing a specific manner to investigate one or both research questions. The models comprised six versions of both the Normal LMM and the Beta-proportion GLLAMM. The differences among the models hinged on (1) whether they addressed data clustering in conjunction with measurement error, denoted as the model type, (2) the assumed distribution for the entropy scores, which aimed to handle boundedness, (3) whether the model incorporates a robust feature to address mild or moderate departures of the data from distributional assumptions, and (4) the inclusion or exclusion of speaker-related factors in the models. A detailed overview of the fitted models is available in Table 2.

Table 2: Fitted models.

Model	Model type	Entropy distribution	Robust feature	Fixed effects $\beta_{HS[i]}$	β_A	$\beta_{A,HS[i]}$
1	LMM	Normal	No	No	No	No
2	LMM	Normal	No	Yes	Yes	No
3	LMM	Normal	No	Yes	No	Yes
4	LMM	Normal	Yes	No	No	No
5	LMM	Normal	Yes	Yes	Yes	No
6	LMM	Normal	Yes	Yes	No	Yes
7	GLLAMM	Beta-prop.	No	No	No	No
8	GLLAMM	Beta-prop.	No	Yes	Yes	No
9	GLLAMM	Beta-prop.	No	Yes	No	Yes
10	GLLAMM	Beta-prop.	Yes	No	No	No
11	GLLAMM	Beta-prop.	Yes	Yes	Yes	No
12	GLLAMM	Beta-prop.	Yes	Yes	No	Yes

2.2.5 Estimation and chain quality

The models were estimated using `R` version 4.2.2 (R Core Team, 2015) and `Stan` version 2.26.1 (Stan Development Team., 2021). Four Markov chains were implemented for each parameter, each with distinct starting values. Each chain underwent $4,000$ iterations, where the first $2,000$ serving as a warm-up phase and the remaining $2,000$ were considered samples from the posterior distribution. Verification of stationarity, convergence, and mixing for the parameter chains involved graphical analysis and diagnostic statistics. Graphical analysis utilized trace, trace-rank, and autocorrelation plots (ACF). Diagnostic statistics included the *potential scale reduction factor statistics* $\widehat{\mathrm{R}}$ with a cut-off value of 1.05 (Vehtari et al., 2021). Furthermore, to confirm whether the parameters posterior distributions were generated with a sufficient number of uncorrelated sampling points, each posterior distribution density plot was inspected along with their effective sample size statistics n_{eff} (Gelman et al., 2014).

In general, both graphical analysis and diagnostic statistics indicated that all chains exhibited low to moderate autocorrelation, explored the parameter space in a seemingly random manner, and converged to a constant mean and variance in their post-warm-up phase. Moreover, the density plots and statistics collectively confirmed that all posterior distributions are unimodal distributions with values centered around a mean, generated with a satisfactory number of uncorrelated sampling points, making substantive sense compared to the models' prior beliefs. The trace, trace-rank, ACF, and distribution density plots, along with $\widehat{\mathrm{R}}$ and n_{eff} statistics, are provided in the accompanying digital walk-through document for reader inspection (refer to Section 2.2.7 Open Science Statement).

2.2.6 Model comparison

The study compares the fitted models using three criteria: the deviance information criterion (`DIC`) by Spiegelhalter et al. (2002), the widely applicable information criterion (`WAIC`) by Watanabe (2013), and the Pareto Smoothing Importance Sampling criterion (`PSIS`) by Vehtari et al. (2017). These criteria score models in terms of deviations from perfect predictive accuracy, with smaller values indicating less deviation (McElreath, 2020). Specifically, `DIC` measures in-sample deviations, while `WAIC` and `PSIS` offer an approximate measure of out-of-sample deviations. Deviations from perfect predictive accuracy serve as the closest estimate for the Kullback-Leibler divergence (Kullback & Leibler, 1951), which measures the degree to which a model accurately represents the true distribution of the data. Moreover, `WAIC` and `PSIS` are considered full Bayesian criteria as they incorporate all the information encompassed in the parameter's posterior distribution. This effectively integrates and reports the inherent uncertainty in the predictive accuracy estimates. Predictive accuracy aside, `PSIS` offers an additional advantage in identifying highly influential data points. To achieve this, the criterion uses a built-in warning system that flags observations that make out-of-sample predictions unreliable. The key intuition is that observations that are relatively unlikely, according to the model, exert more influence and render predictions more unreliable than those relatively expected (McElreath, 2020).

2.2.7 Open Science Statement

In an effort to improve the transparency and replicability of the analysis, this study provides access to an online walk-through. The digital document contains all the code and materials utilized in the study. Furthermore, the walk-through meticulously follows the When-to-Worry-and-How-to-Avoid-the-Misuse-of-Bayesian-Statistics checklist (WAMBS checklist) developed by Depaoli & van de Schoot (2017). This checklist outlines the ten crucial points that need careful scrutiny when employing Bayesian inference procedures. The digital walk-through is available at the following URL: `https://jriveraespejo.github.io/paper1_manuscript/`

3 Results

This section presents the results of the Bayesian inference procedures, with particular emphasis on answering the three research questions.

3.1 Predictive capabilities of the Beta-proportion GLLAMM compared to the Normal LMM (RQ1)

This research question evaluates the effectiveness of the Beta-proportion GLLAMM in handling the features of entropy scores by comparing its predictive accuracy to the Normal LMM. Models 1, 4, 7, and 10 are specifically chosen for this comparison because their assumptions exclusively address the features of the scores, without integrating additional covariate information. As detailed in Table 2, Model 1 is a Normal LMM that solely addresses data clustering. Building upon this, Model 4 introduces a robust feature. Conversely, Model 7 is a Beta-proportion GLLAMM that deals with boundedness, measurement error and data clustering, and Model 10 extends this model by incorporating a robust feature.

Figure 1 displays values for the `DIC`, `WAIC`, and `PSIS`. They also include the components `dWAIC` and `dPSIS`, highlighting the differences in out-of-sample deviations from the best-fitting model and its associated uncertainty. The associated Table 6 and Table 7 provide similar information, while also reporting the `pWAIC` and `pPSIS` values, indicating the penalization received by the models for their complexity (roughly associated with their number of parameters). Lastly, the tables show the `weight` of evidence, which summarizes the relative support for each model.

Overall, all criteria consistently point to Model 10 as the most plausible choice for the data. The model exhibits the lowest values for both `WAIC` and `PSIS`, establishing itself as the model with the least deviation from *perfect* predictive accuracy among those under comparison. Additionally, Figure 1 visually demonstrates the non-overlapping uncertainty (horizontal blue lines) in both `dWAIC` and `dPSIS` values for Models 1, 4, and 7 when compared to Model 10. This indicates that Model 10 significantly deviates the least from *perfect* predictive accuracy when compared to the rest of the models. Lastly, the `weight` of evidence in Table 6 and Table 7 underscores that 100% of the evidence aligns with and supports Model 10.

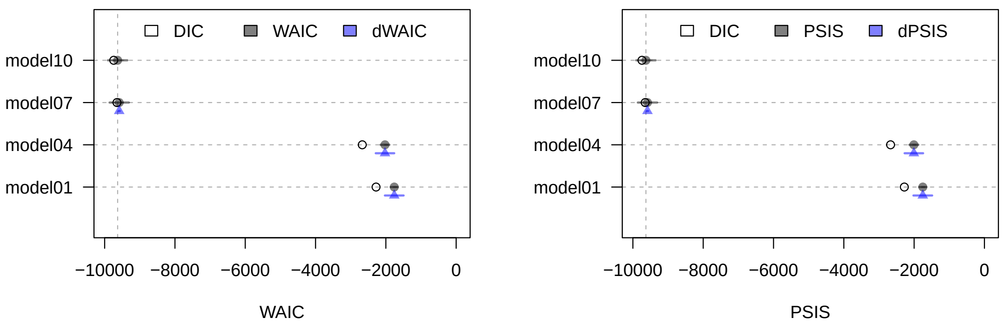

Figure 1: WAIC and PSIS model comparison plot. *Note:* Black and blue points describe point estimates, and continuous horizontal lines indicate the associated uncertainty.

Upon closer examination, the reasons behind the observed disparities in the models become more apparent. Specifically, Figure 2 highlights that the Normal LMM, as outlined in Model 4, fails to capture the underlying data patterns, resulting in predictions that are physically inconsistent, falling outside the outcome's range between zero and one. Further insight into this issue is provided by Figure 9 and Figure 11. Figure 9 displays Model 4's score prediction densities which bear no resemblance to the actual data densities. Furthermore, the top two panels in Figure 11 reveal that misspecification in the Normal LMM causes the model to be *more surprised* by 'extreme' entropy scores, leading to their identification as highly unlikely and influential observations. Consequently, the model is rendered unreliable due to the potential biases present in the parameter estimates. In contrast, the Beta-proportion GLLAMM appears to effectively capture the data patterns, generating predictions within the expected data range. This is evident in Figure 2 and complemented by Figure 10 and Figure 11. In Figure 10, Model 10 display prediction densities that bear more resemblance to the actual data densities. Furthermore, the bottom two panels in Figure 11 show the model is *less surprised* by 'extreme' scores, fostering more trust in the model's estimates.

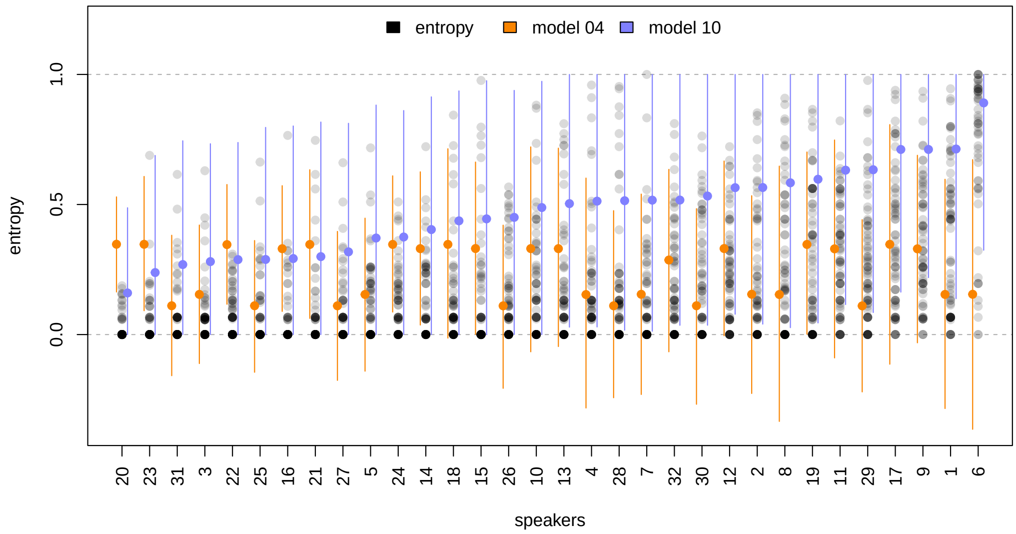

Figure 2: Entropy scores prediction for selected models. *Note:* Black dots show manifest entropy scores, orange dots and vertical lines show the point estimates and 95% highest probability density interval (HPDI) derived from Model 4, blue dots and vertical lines show similar information for Model 10.

3.2 Estimation of speakers' latent potential intelligibility from manifest entropy scores (RQ2)

The second research question aimed to demonstrate the application of the Beta-proportion GLLAMM in estimating the latent potential intelligibility of speakers. This was achieved by employing the general mathematical formalism outlined in Equation 6, along with additional specifications provided in Table 2. The Bayesian procedure successfully estimated the latent potential intelligibility of speakers under Model 10 through the structural equation:

$$SI_i = \alpha + e_i + u_i \tag{9}$$

Moreover, due to its implementation under Bayesian procedures, Model 10 provides the complete posterior distribution of the speakers' potential intelligibility scores. This provision, in turn, (1) enables the calculation of summaries, facilitating the ranking of individuals, and (2) supports the assessment of differences among selected speakers. In both cases, the model considers the inherent uncertainty of the estimates resulting from its measurement using multiple entropy scores.

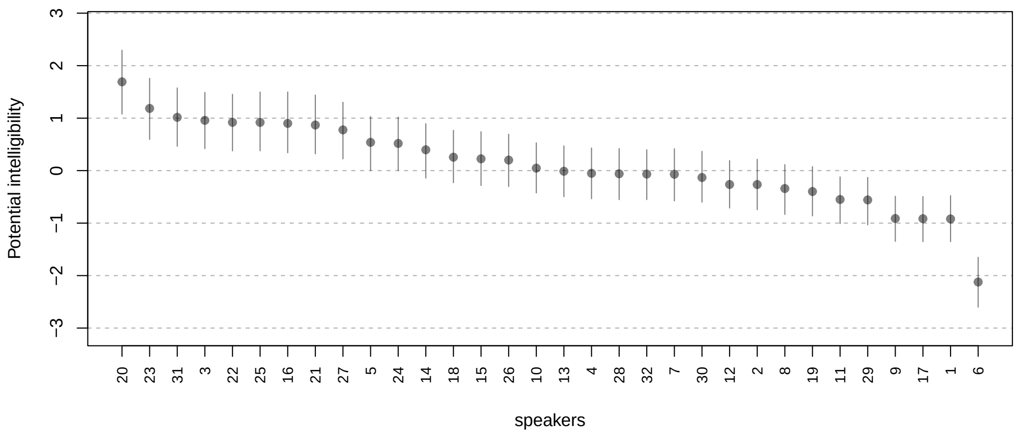

Figure 3: Model 10, latent potential intelligibility of speakers. *Note:* Black dots and vertical lines show mean point estimates and 95% HPDI intervals.

Figure 3 displays the ranking of speakers in decreasing order based on point estimates of the latent potential intelligibility. These estimates are accompanied by their associated 95% highest probability density intervals (HPDI). The figure clearly indicates that clearly indicate that speaker 6 stands out as the least intelligible in the sample, followed farther behind by speaker 1, 17 and 9. In contrast, the figure highlights speaker 20 as the most intelligible, closely followed by speakers 23, 31 and 3. Conversely, Figure 4 shows the full posterior distribution for the comparison of potential intelligibility among selected speakers. The figure reveals that only the differences between speakers 6, 1, 17, and 9, along with the difference between speakers 20 and 3 are statistically significant, as their associated 95% HPDI did not overlap with zero (shaded area). The R code to derive these scores and generate the figure is available in the digital walk-through document (refer to Section 2.2.7 Open Science Statement).

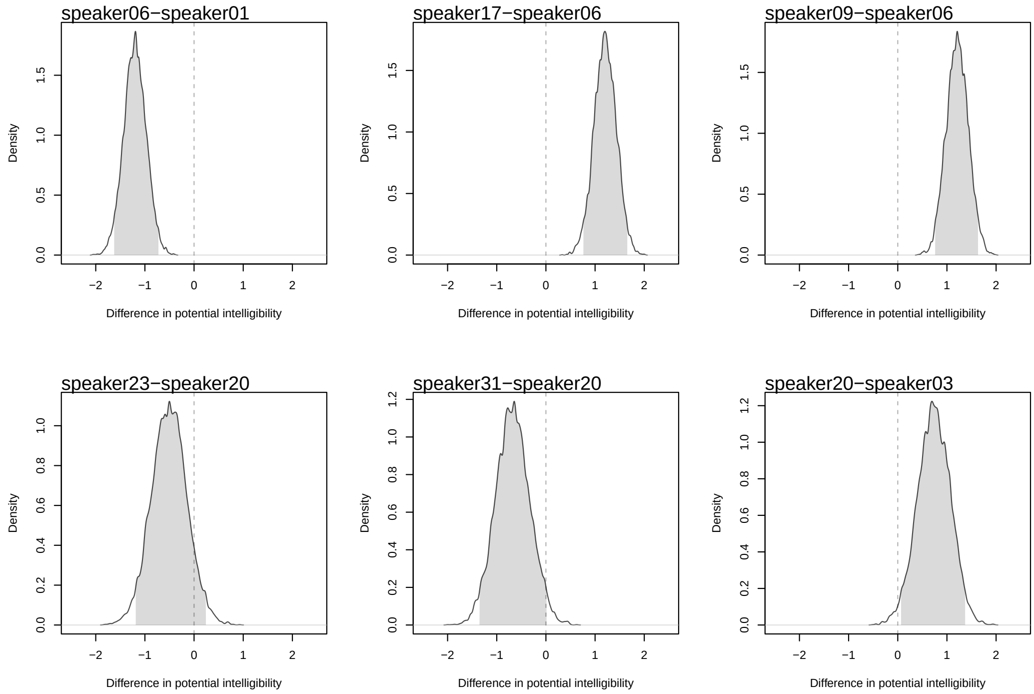

Figure 4: Model 10, potential intelligibility comparisons among selected speakers. *Note:* Shaded area describes the 95% highest probability density interval (HPDI)

3.3 Testing the influence of speaker-related factors on intelligibility (RQ3)

This research question illustrates how theories on intelligibility can be examined within the model's framework. Specifically, the focus centers on assessing the influence of speaker-related factors on intelligibility, such as chronological age and hearing status. Notably, despite RQ1 indicating the suitability of Beta-proportion GLLAMM models for entropy scores, existing statistical literature suggests that, in certain scenarios, models incorporating covariate adjustment exhibit robustness to misspecification in the functional form linking an outcome and covariates, commonly referred to as covariate-outcome relationship (Tackney et al., 2023). Consequently, this study compares all models detailed in Table 2. These models are characterized by different covariate adjustments on entropy scores or the latent potential intelligibility of speakers, namely chronological age and hearing status, while potentially exhibiting misspecification in the covariate-outcome relationship, as observed in the case of the Normal LMM.

Similar to RQ1, all criteria consistently identify the Beta-proportion GLLAMM outlined in models 11, 12 and 10 as the most plausible models for the data. The models exhibit the lowest values for both `WAIC` and `PSIS`, establishing them as the least deviating models among those under comparison. Moreover, Figure 5 depicts with horizontal blue lines the non-overlapping uncertainty for the models' `dWAIC` and `dPSIS` values. This reveals that, when compared to Model 11, most models exhibit significantly distinct predictive capabilities. Models 12 and 10, however, stand out as exceptions to this pattern. This observation suggests that Models 11, 12, and 10 display the least deviation from *perfect* predictive accuracy in contrast to the other models. Lastly, the `weight` of evidence in Tables Table 8 and Table 9, underscores that Model 11 accumulated the greatest support, followed by Model 12, and lastly, by Model 10.

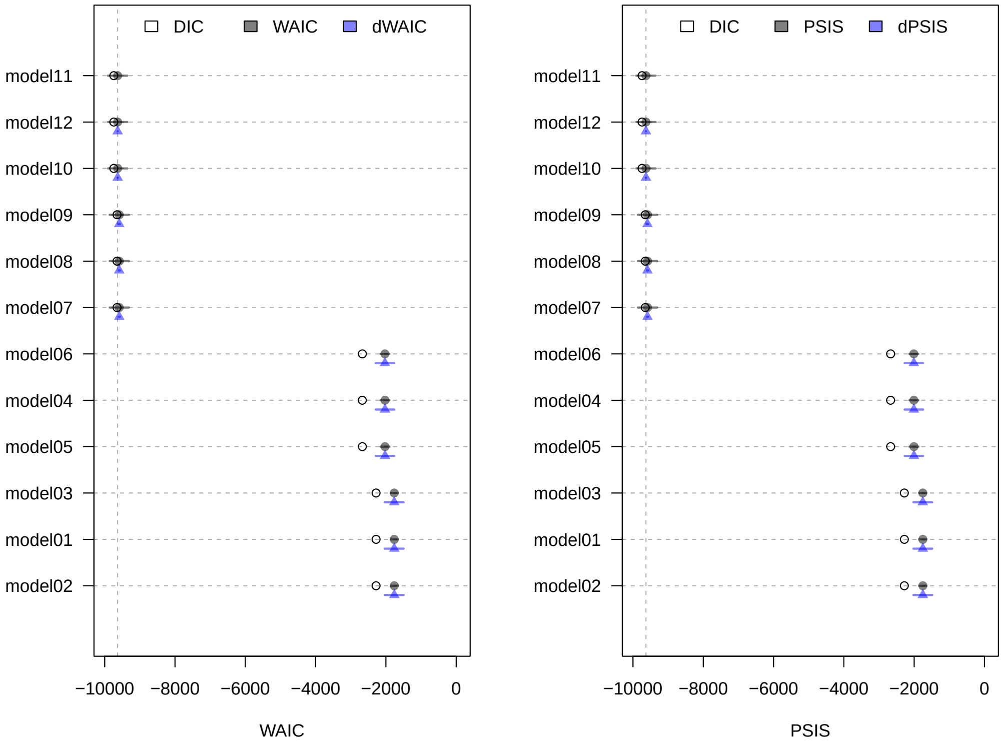

Figure 5: WAIC and PSIS model comparison plot. Note: Black and blue points describe point estimates, and continuous horizontal lines indicate the associated uncertainty.

A closer examination of two models within this comparison set reveal the reasons behind the largest observed disparities. The Normal LMM, as outlined in Model 6, continues to face challenges in capturing underlying data patterns, resulting in predictions that are physically inconsistent, falling outside the outcome's range. Additionally, the model persists in identifying highly unlikely and influential observations, making it inherently unreliable. In contrast, the Beta-proportion GLLAMM described by Model 12 appears to be less susceptible to 'extreme' scores, effectively capturing data patterns within the expected data range and thereby instilling greater confidence in the reliability of the model's estimates. This contrast is visually depicted in Figure 12, Figure 13, Figure 14, and Figure 15.

Considering the results in Figure 5, the model comparisons favor three distinct models: Model 10, 11 and 12. Model 10, supported by 20.4% of the evidence, estimates a single intercept α and no slope to explain the potential intelligibility of speakers (refer to Table 3). In contrast, supported by 45.1% of the evidence, Model 11 in Table 4 estimates distinct intercepts for each hearing status group, namely $\alpha_{HS[1]}$ for NH speakers and $\alpha_{HS[2]}$ for the HI/CI counterparts, while maintaining a single slope that gauges the impact of age on potential intelligibility estimates. The 95% HPDI for the comparison of intercepts $\alpha_{HS[2]} - \alpha_{HS[1]}$ reveal significant differences between

NH and HI/CI speakers. Lastly, with evidence of 34.1%, Model 12 in Table 5 estimates one intercept and slope per hearing status group, namely $\alpha_{HS[1]}$ and $\beta_{A,HS[1]}$ for the NH speakers, and $\alpha_{HS[2]}$ and $\beta_{A,HS[2]}$ for the HI/CI counterparts. The 95% HPDI for the comparison of intercepts and slopes reveal significant differences solely in the slopes between NH and their HI/CI counterparts ($\beta_{A,HS[2]} - \beta_{A,HS[1]}$).

However, a discerning reader can notice that these models yield conflicting conclusions regarding the influence of chronological age and hearing status on intelligibility. Model 10 implies no influence of chronological age and hearing status on the potential intelligibility of speakers. A visual inspection of Figure 6, however, reveals the reason for the model's low support. Model 10 fails to capture the prevalent increasing age pattern observed in potential intelligibility estimates. In contrast, Model 11 identifies significant differences in potential intelligibility between NH and HI/CI speakers. The model further suggests that with the progression of chronological age, HI/CI speakers lag behind in intelligibility development, with no opportunity to catch up to their NH counterparts within the analyzed age range, as depicted in Figure 7. Finally, Model 12 indicates no significant differences in intelligibility between NH and HI/CI speakers at 68 months of age (around 6 years old). However, the model reveals distinct evolution patterns of intelligibility per unit of chronological age between different hearing status groups, with HI/CI speakers displaying a slower rate of development compared to their NH counterparts within the analyzed age range. The latter is evident in Figure 8.

Table 3: Model 10, parameter estimates and 95% highest probability density intervals (HPDI)

Parameter	Posterior mean	95% HPDI
α	0.01	[-0.09, 0.1]

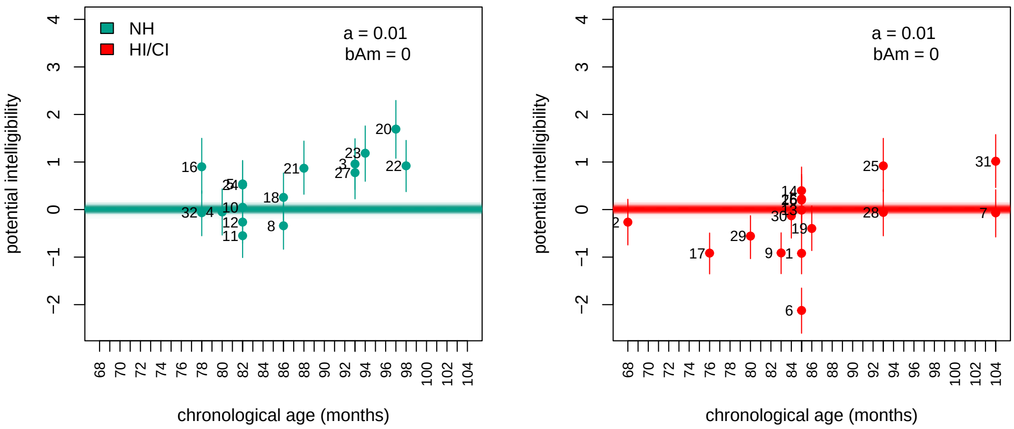

Figure 6: Model 10, Potential intelligibility per chronological age and hearing status. *Note:* Colored dots denote mean point estimates, vertical lines describe the 95% highest probability density intervals (HPDI), thick discontinuous line indicate the regression line, thin continuous lines denote regression lines samples from the posterior distribution, and numbers indicate the speaker index.

Table 4: Model 11, parameter estimates and 95% highest probability density intervals (HPDI)

Parameter	Posterior mean	95% HPDI
α	0.01	[-0.08, 0.11]
$\alpha_{HS[1]}$	0.53	[0.11, 0.94]
$\alpha_{HS[2]}$	-0.03	[-0.43, 0.39]
β_A	0.07	[0.05, 0.1]
Contrasts		
$\alpha_{HS[2]} - \alpha_{HS[1]}$	-0.55	[-1, -0.15]

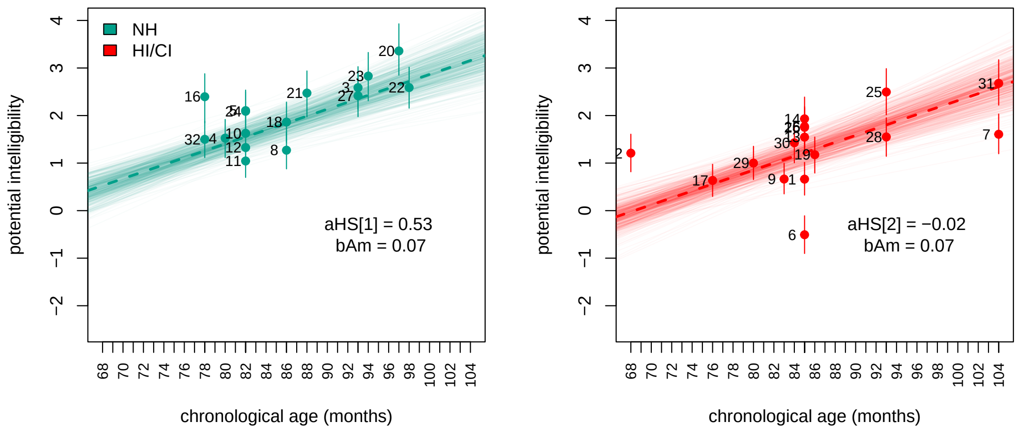

Figure 7: Model 11, Potential intelligibility per chronological age and hearing status. *Note:* Colored dots denote mean point estimates, vertical lines describe the 95% highest probability density intervals (HPDI), thick discontinuous line indicate the regression line, thin continuous lines denote regression lines samples from the posterior distribution, and numbers indicate the speaker index.

Table 5: Model 12, parameter estimates and 95% highest probability density intervals (HPDI)

Parameter	Posterior mean	95% HPDI
α	0.01	[-0.09, 0.11]
$\alpha_{HS[1]}$	0.21	[-0.28, 0.72]
$\alpha_{HS[2]}$	0.23	[-0.24, 0.69]
$\beta_{A,HS[1]}$	0.10	[0.07, 0.13]
$\beta_{A,HS[2]}$	0.06	[0.03, 0.09]
Contrasts		
$\alpha_{HS[2]} - \alpha_{HS[1]}$	0.01	[-0.61, 0.74]

Table 5: Model 12, parameter estimates and 95% highest probability density intervals (HPDI)

Parameter	Posterior mean	95% HPDI
$\beta_{A,HS[2]} - \beta_{A,HS[1]}$	-0.04	[-0.08, 0]

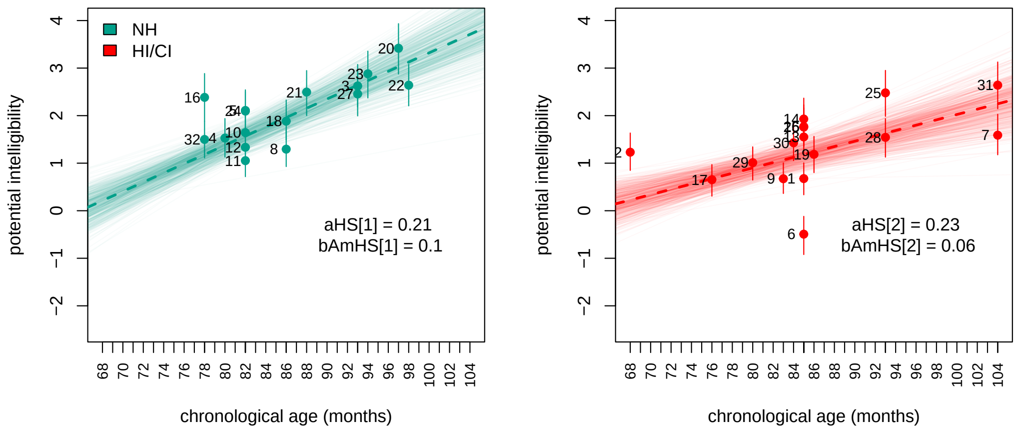

Figure 8: Model 12, Potential intelligibility per chronological age and hearing status. Note: Colored dots denote mean point estimates, vertical lines describe the 95% highest probability density intervals (HPDI), thick discontinuous line indicate the regression line, thin continuous lines denote regression lines samples from the posterior distribution, and numbers indicate the speaker index.

4 Discussion

4.1 Findings

This study examined the suitability of the Bayesian Beta-proportion GLLAMM for the quantitative measuring and testing of research theories related to speech intelligibility using entropy scores. The initial findings supported the assertion that Beta-proportion GLLAMMs consistently outperformed Normal LMMs in predicting entropy scores, underscoring its superior predictive performance. The results emphasized that models neglecting the outcomes' measurement error and boundedness lead to underfitting and misspecification issues, even when robust features are integrated. This is clearly illustrated by the Normal LMMs.

Secondly, the study showcased the Beta-proportion GLLAMM's proficiency in estimating the latent potential intelligibility of speakers based on manifest entropy scores. Implemented under Bayesian procedures, the proposed model offered a valuable advantage over frequentist methods by further providing the full posterior distribution of the speakers' potential intelligibility. This provision facilitated the calculation of summaries, aiding individual rankings, and supported the comparisons among selected speakers. In both scenarios, the proposed model accounted for the inherent uncertainty in the intelligibility estimates.

Thirdly, the study illustrated how the proposed model assessed the impact of speaker-related factors on potential intelligibility. The results suggested that multiple models were plausible for the observed entropy scores, indicating that different speaker-related factor theories were viable for the data, with some presenting contradictory conclusions about the influence of those factors on intelligibility. However, even when unequivocal support for one theory was not possible, the divided support among these models informed that certain statistical issues may be hindering the model's ability to distinguish among individuals and, ultimately, among models. These issues encompassed the insufficient sample size of speakers, the inadequate representation of the population of speakers, and the imprecise measurement of the latent variable of interest.

Ultimately, this study introduced researchers to innovative statistical tools that enhanced existing research models. These tools not only assessed the predictability of empirical phenomena but also quantitatively measured the latent trait of interest, namely potential intelligibility, facilitating the comparison of research theories related to this trait. However, the presented tools introduce new challenges for researchers seeking their implementation. These challenges emerge from two distinct aspects: one methodological and the other practical. In the methodological domain, researchers need familiarity with Bayesian methods and the principled formulation of assumptions regarding the data-generating process and research inquiries. This entails understanding and addressing each of the data and research challenges within the context of a statistical (probabilistic) model. Conversely, in the practical domain, researchers need familiarity with probabilistic programming languages (PPLs), which are designed for specifying and obtaining inferences from probabilistic models -the core of Bayesian methods. To ensure the successful utilization of this new statistical tool, this study addresses both challenges by providing comprehensive, step-by-step guidance in the form of a digital walk-through document (refer to Section 2.2.7 Open Science Statement).

4.2 Limitations and further research

This study provides valuable insights into the use of a novel approach to simultaneously address the different data features of entropy scores in speech intelligibility research. However, it is important to acknowledge the limitations of this study and explore potential avenues for future research. Firstly, the study interprets potential intelligibility as an unobserved latent trait of speakers influencing the likelihood of observing a set of entropy scores. These scores, in turn, reflect the transcribers' ability to decode words in sentences produced by the same speakers. Despite this practical approach, the construct validity of the latent trait heavily depends on the listeners' appropriate understanding and execution of the transcription task. Construct validity, as defined by Cronbach & Meehl (1955), refers to the extent to which a set of manifest variables accurately represents a concept that cannot be directly measured. Considering the study assumes the transcription task set by Boonen and colleagues (2021) was properly understood and executed, it expects that potential intelligibility reflects the overall speech intelligibility of speakers. However, the study does not delve into the general epistemological considerations regarding the connection between the latent variable and the concept.

Secondly, the study revealed a notable lack of unequivocal support for one of the models among the compared set. This outcome may be attributed to factors such as the insufficient sample size of speakers, the inadequate representation of the populations of speakers (referred to as selection bias), and the imprecise measurement of the latent variable. Small sample size and selection bias yield data with limited outcome and covariates ranges, leading to biased and imprecise parameter estimates (Everitt & Skrondal, 2010). Moreover, fueled by the reduced measurement precision, these issues can result in models with diminished statistical power and a higher risk of type I or type II errors (McElreath, 2020). Consequently, future research should consider conducting power analyses for the proposed models. This entails assessing the impact of expanding the speakers' pool on testing research theories, or increasing the number of speech samples, transcriptions, and listeners to enhance the precision of potential intelligibility estimates. With these insights, future investigations should contemplate increasing the speaker sample with a group that adequately represents the population of interest. However, this must be done while mindful of the pragmatic limitations associated with transcription tasks, specifically considering the costs and time-intensiveness of the procedure.

Thirdly, the study presented an illustrative example for the investigation of research theories within the model's framework. However, it did not offer an exhaustive evaluation of all factors influencing intelligibility, which are thoroughly explored in the works of Niparko et al. (2010), Boons et al. (2012), Gillis (2018), and Fagan et al. (2020). Consequently, the study cannot discard the presence of unobservable variables that might bias the parameter estimates, potentially impacting the inferences provided. Hence, future research should consider integrating appropriate causal hypotheses about these factors into the proposed models, as proper covariate adjustment facilitates the production of unbiased and precise parameter estimates (Cinelli et al., 2022; Deffner et al., 2022).

Lastly, this study proposes two directions for future exploration in speech intelligibility research. Firstly, there is an opportunity to investigate alternative methods for assessing speech intelligibility beyond transcription tasks and entropy scores. The experimental design of transcription tasks imply that the procedure may be time-intensive and costly. Thus, exploring less time-intensive or more cost-effective procedures, that still offer comparable precision in intelligibility estimates, could benefit both researchers and speech therapists alike. An illustrative example of such a method is Comparative Judgment (CJ), where judges compare and score the perceived intensity of a trait between two stimuli (Thurstone, 1927). In the context of the intelligibility trait, the stimuli under assessment could be the speech samples uttered by two speakers. Nevertheless, CJ serve as an ideal example as the method has gained increasing attention within the realm of educational assessment, with several studies providing evidence for its validity in assessing various task within student works, as demonstrated by examples in Pollitt (2012); Pollitt_2012b, Lesterhuis (2018), van Daal (2020), and Verhavert et al. (2019).

Conversely, a second avenue for exploration involves integrating diverse data types and evaluation methods to assess individuals' intelligibility. This can be accomplished by leveraging two features of Bayesian methods: their flexibility and the concept of Bayesian updating. Bayesian methods possess the flexibility to simultaneously handle various data types. Additionally, through Bayesian updating, researchers can integrate information from the posterior distribution of parameters as priors in models for subsequent evaluations. Ultimately, this could enable researchers to assess speakers' intelligibility progress without committing to a specific data type or evaluation method. This advancement could mirror the emergence of second-generation Structural Equation Models proposed by Muthén (2001), where models facilitate the combined estimation of categorical and continuous latent variables. However, in the context of future research, the proposal would facilitate the estimation of latent variables using a combination of data types and evaluation methods, contingent upon the fulfillment of construct validity by those evaluation methods.

5 Conclusion

This study highlights the effectiveness of the Bayesian Beta-proportion GLLAMM to collectively address several key data features when investigating unobservable and complex traits, using speech intelligibility and entropy scores as an example. The results demonstrate the proposed model consistently outperforms the Normal LMM in predicting the empirical phenomena. Moreover, it exhibits the ability to quantify the latent potential intelligibility of speakers, allowing for the ranking and comparison of individuals based on the latent trait while accommodating associated uncertainties. Additionally, the proposed model facilitates the exploration of research theories concerning the influence of speaker-related factors on potential intelligibility. The study indicates that integrating and comparing these theories within the model's framework is a straightforward task.

However, the introduction of these innovative statistical tools presents new challenges for researchers seeking implementation. These challenges encompass the principled formulation of assumptions about the data-generating processes and research inquiries, along with the need for familiarity with probabilistic programming languages (PPLs) essential for implementing Bayesian methods. Nevertheless, the study suggests several promising avenues for future research, including power analysis, causal hypothesis formulation, and exploration and integration of novel evaluation methods for assessing intelligibility. The insights derived from this study hold implications for both researchers and data analysts interested in quantitatively measuring and testing theories related to nuanced, unobservable constructs, while also considering the appropriate prediction of the empirical phenomena.

6 Appendix

6.1 Entropy scores calculation

This section exemplifies the entropy calculation procedure. For that purpose, the words in position two, four and five observed in Table 1 were used. These words were assumed present in the first sentence, produced by the first speaker assigned to the first block, and transcribed by five listeners ($w = \{2,4,5\}$, $s = 1$, $i = 1$, $b = 1$, $J = 5$). For the word 2, the first four listeners identified the word type *jongen* (T_{j1}), while the last identified the word type *hond* (T_{j2}). Therefore, two word types were identified ($K = 2$), with proportions equal to $\{p_1, p_2\} = \{4/5, 1/5\} = \{0.8, 0.2\}$, and entropy score equal to:

$$H_{2111} = \frac{0.8 \cdot log_2(0.8) + 0.2 \cdot log_2(0.2)}{log_2(5)} \approx 0.3109$$

For the word 4, two listeners identified the word type *een* (T_{j1}), one listener the word type *de* (T_{j2}), and another the word *geen* (T_{j3}). A blank space *[B]* is a symbol that defines the absence of a word in a space where a word is expected, as compared with other transcriptions, during the alignment procedure. Notice that for calculation purposes, because the blank space is not expected in such position, this is considered as a different word type. Consequently four word types were registered ($K = 4$), with proportions equal to $\{p_1, p_2, p_3, p_4\} = \{2/5, 1/5, 1/5, 1/5\} = \{0.4, 0.2, 0.2, 0.2\}$ and entropy score equal to:

$$H_{4111} = \frac{0.4 \cdot log_2(0.4) + 3 \cdot 0.2 \cdot log_2(0.2)}{log_2(5)} \approx 0.8277$$

Lastly, for word 5, each listener transcribed a different word. it is important to highlight that when a listener does not identify a complete word, or part of it, (s)he is instructed to write *[X]* in that position. However, for the calculation of the entropy score, if more than one listener marks an unidentifiable word with *[X]*, each one of them is considered a different word type. This is done to avoid the artificial reduction of the entropy score, as *[X]* values already indicate the word's lack of intelligibility. . Consequently, five word types were observed, T_{j1} =*kikker*, T_{j2} =*[X]*, T_{j3} =*kokkin*, T_{j4} =*kikkers*, T_{j5} =*[X]* ($K = 5$), with proportions equal to $\{p_1, p_2, p_3, p_4, p_5\} = \{1/5, 1/5, 1/5, 1/5, 1/5\} = \{0.2, 0.2, 0.2, 0.2, 0.2\}$, and entropy score equal to:

$$H_{5111} = \frac{5 \cdot 0.2 \cdot log_2(0.2)}{log_2(5)} = 1$$

739 ## 6.2 Tables

Table 6: WAIC comparison for selected models

Model	DIC	WAIC	SE	dWAIC	dSE	pWAIC	weight
10	-9741.66	-9630.63	276.64	0.00		55.52	1
7	-9649.54	-9586.00	274.50	44.63	17.89	31.77	0
4	-2670.62	-2024.84	127.02	7605.78	263.22	322.89	0
1	-2278.68	-1761.10	101.80	7869.53	266.54	258.79	0

Table 7: PSIS comparison for selected models

Model	DIC	PSIS	SE	dPSIS	dSE	pPSIS	weight
10	-9741.66	-9629.27	276.74	0.00		56.19	1
7	-9649.54	-9585.92	274.56	43.36	17.67	31.81	0
4	-2670.62	-2007.66	128.57	7621.61	263.60	331.48	0
1	-2278.68	-1753.71	102.09	7875.57	266.54	262.48	0

Table 8: WAIC comparison for all models

Model	DIC	WAIC	SE	dWAIC	dSE	pWAIC	weight
11	-9741.51	-9632.24	276.80	0.00		54.63	0.46
12	-9741.49	-9631.66	276.82	0.58	1.00	54.91	0.34
10	-9741.66	-9630.63	276.64	1.61	2.97	55.52	0.20
9	-9649.15	-9586.67	274.35	45.56	18.01	31.24	0.00
8	-9649.05	-9586.41	274.33	45.83	18.01	31.32	0.00
7	-9649.54	-9586.00	274.50	46.24	18.19	31.77	0.00
6	-2669.28	-2027.11	126.86	7605.13	263.15	321.08	0.00
4	-2670.62	-2024.84	127.02	7607.40	263.22	322.89	0.00
5	-2669.28	-2024.58	127.06	7607.66	263.24	322.35	0.00
3	-2279.58	-1762.08	101.79	7870.16	266.68	258.75	0.00
1	-2278.68	-1761.10	101.80	7871.14	266.64	258.79	0.00
2	-2279.35	-1760.36	101.86	7871.88	266.69	259.49	0.00

Table 9: PSIS comparison for all models

Model	DIC	PSIS	SE	dPSIS	dSE	pPSIS	weight
11	-9741.51	-9631.16	276.88	0.00		55.17	0.46
12	-9741.49	-9630.70	276.90	0.47	1.01	55.39	0.36
10	-9741.66	-9629.27	276.74	1.89	2.84	56.19	0.18
9	-9649.15	-9586.58	274.41	44.58	17.91	31.28	0.00
8	-9649.05	-9586.33	274.39	44.83	17.91	31.36	0.00
7	-9649.54	-9585.92	274.56	45.24	18.10	31.81	0.00

Table 9: PSIS comparison for all models

Model	DIC	PSIS	SE	dPSIS	dSE	pPSIS	weight
6	-2669.28	-2009.22	128.46	7621.94	263.52	330.03	0.00
4	-2670.62	-2007.66	128.57	7623.50	263.60	331.48	0.00
5	-2669.28	-2006.49	128.71	7624.67	263.62	331.39	0.00
3	-2279.58	-1754.43	102.07	7876.73	266.68	262.57	0.00
1	-2278.68	-1753.71	102.09	7877.46	266.64	262.48	0.00
2	-2279.35	-1752.86	102.13	7878.30	266.68	263.24	0.00

6.3 Figures

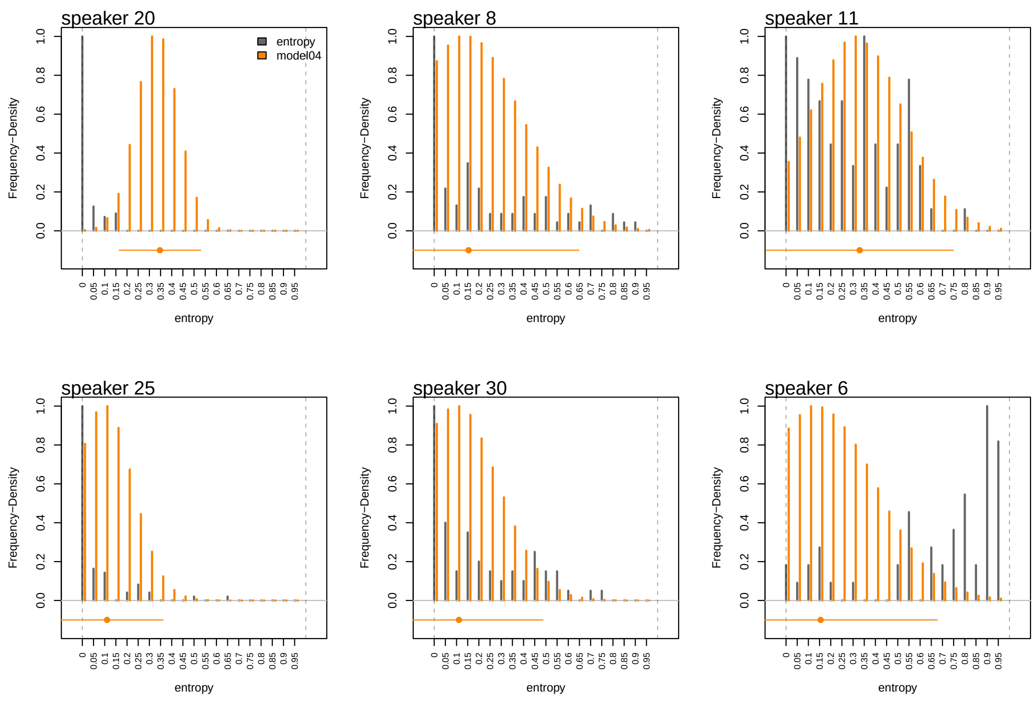

Figure 9: Model 4: Entropy scores density for selected speakers. *Note:* Black bars denote the true data density, orange bars describe the predicted data density

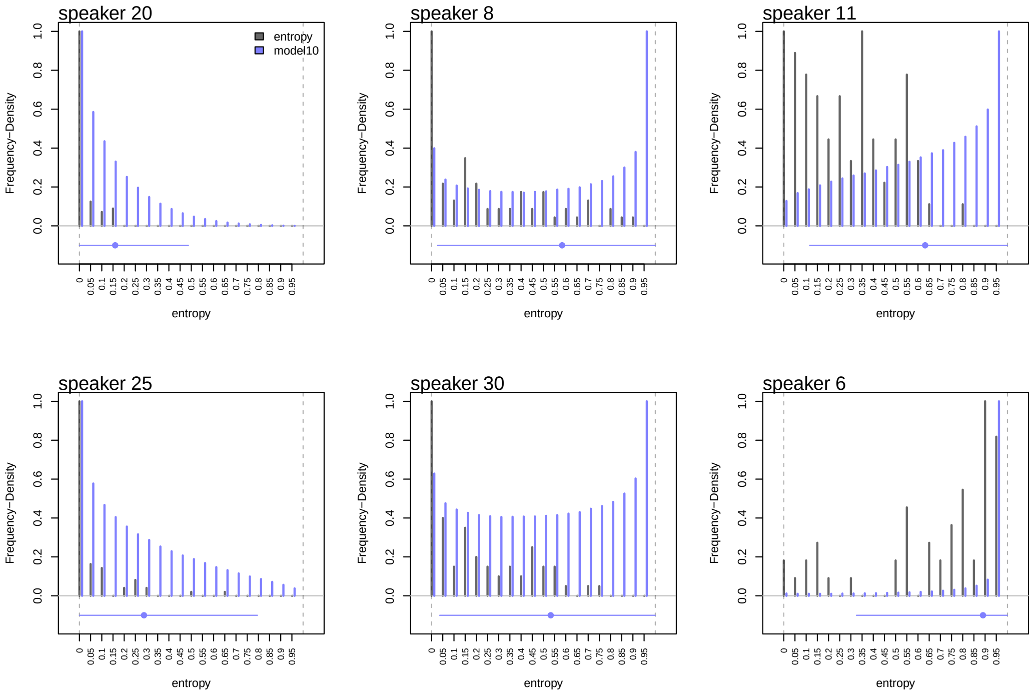

Figure 10: Model 10: Entropy scores density for selected speakers. *Note:* Black bars denote the true data density, blue bars describe the predicted data density

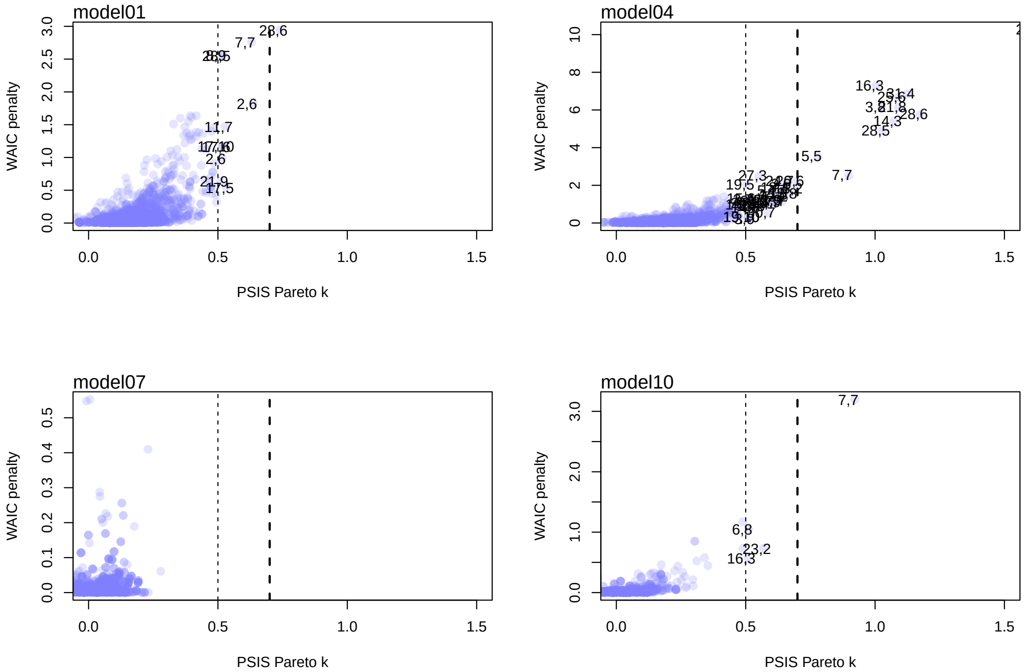

Figure 11: Outlier identification and analysis for selected models. Note: Thin and thick vertical discontinuous line indicate threshold of 0.5 and 0.7, respectively. Number pair texts indicate the observation pair of speaker and sentence index.

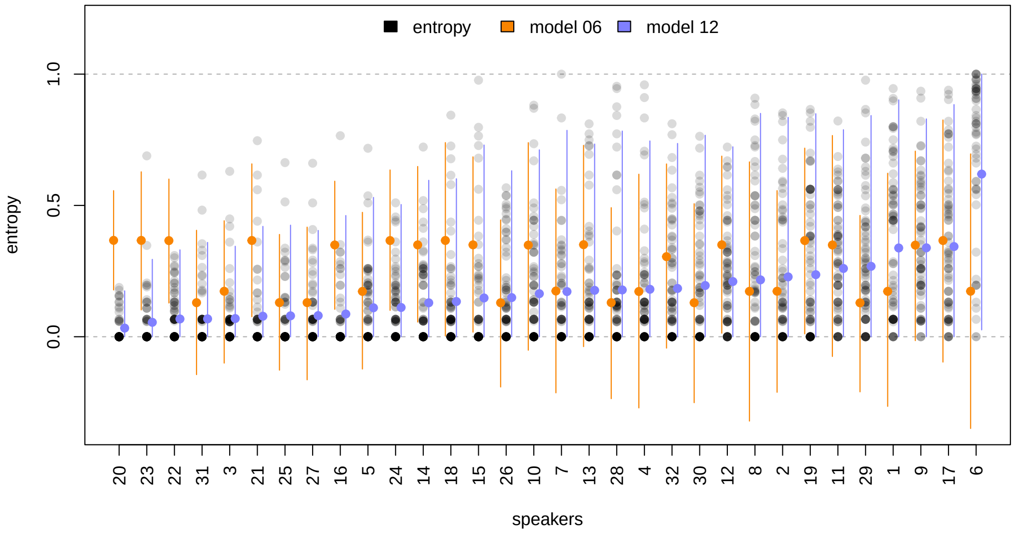

Figure 12: Entropy scores prediction for selected models. Note: Black dots show manifest entropy scores, orange dots and vertical lines show the point estimates and 95% highest probability density intervals (HPDI) derived from model 6, blue dots and vertical lines show similar information for model 12.

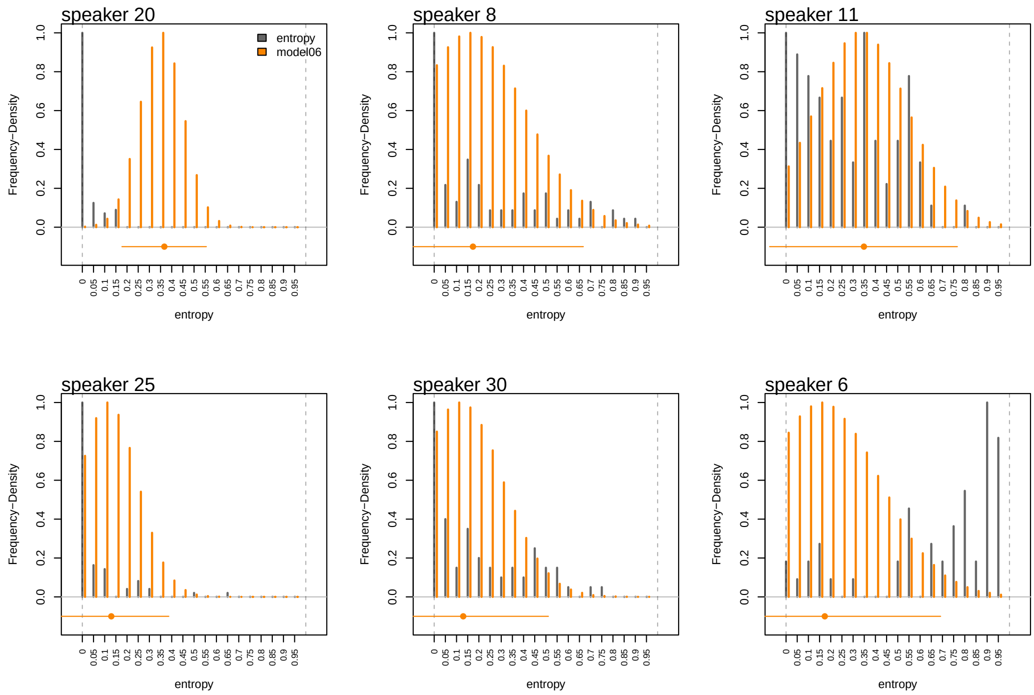

Figure 13: Model 6: Entropy scores density for selected speakers. Note: Black bars denote the true data density, orange bars describe the predicted data density

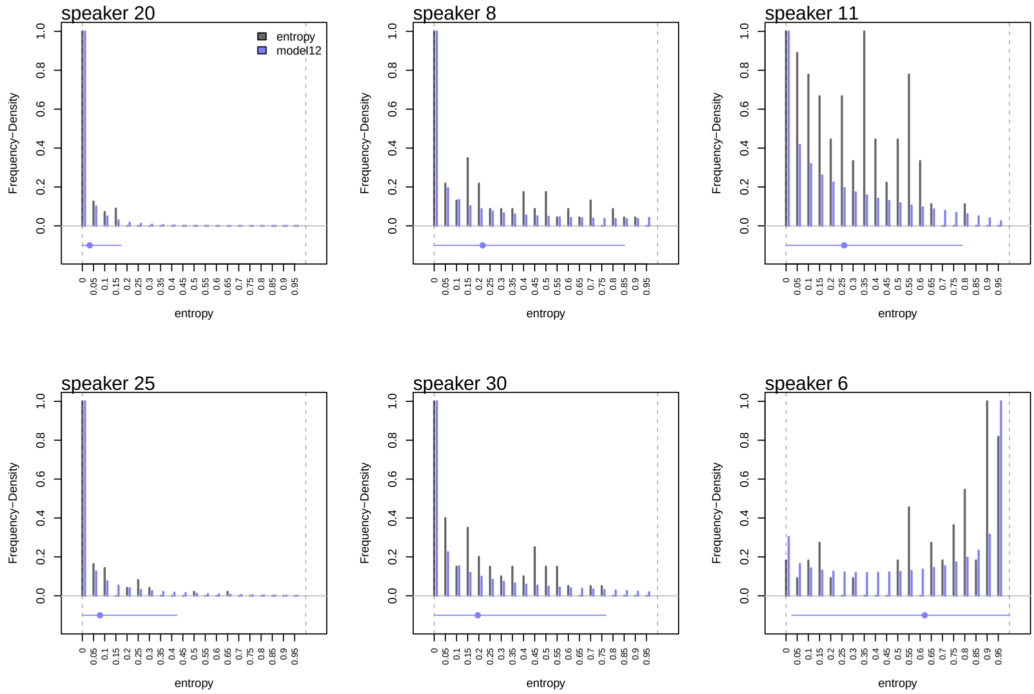

Figure 14: Model 12: Entropy scores density for selected speakers. Note: Black bars denote the true data density, blue bars describe the predicted data density

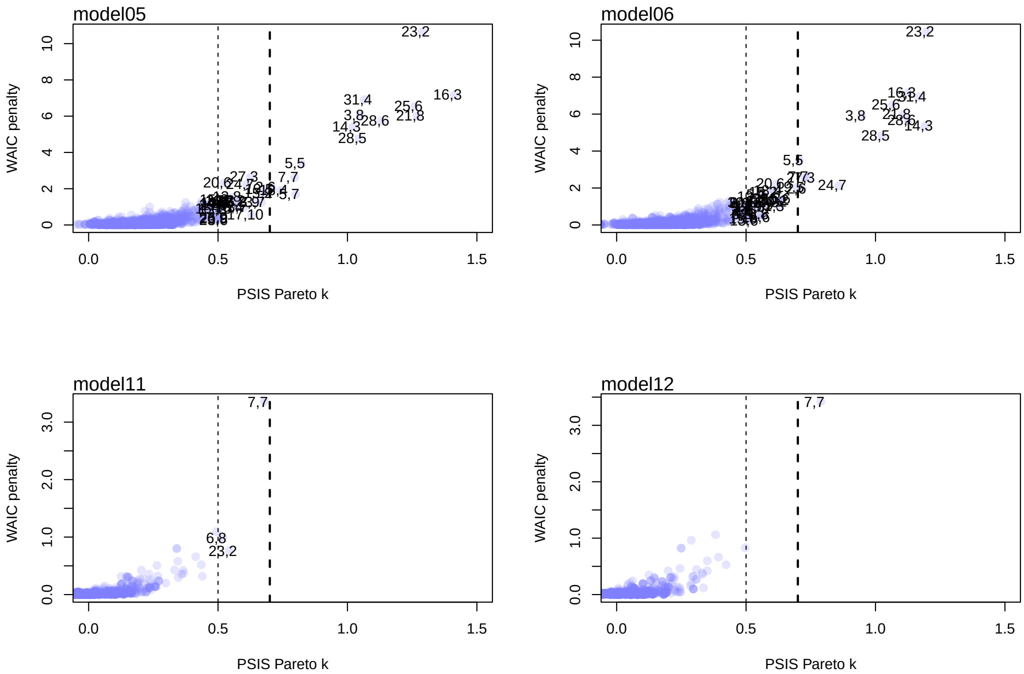

Figure 15: Outlier identification and analysis for selected models. Note: Thin and thick vertical discontinuous line indicate threshold of 0.5 and 0.7, respectively. Number pair texts indicate the observation pair of speaker and sentence index.

Declarations

Funding: The project was founded through the Research Fund of the University of Antwerp (BOF).

Conflict of interests: The authors declare no conflict of interest.

Ethics approval: This is an observational study. The University of Antwerp Research Ethics Committee has confirmed that no ethical approval is required.

Consent to participate: Not applicable

Consent for publication: All authors have read and agreed to the published version of the manuscript.

Availability of data and materials: The data is delivered upon request, while the user-defined functions are available in the `code` folder from this walk-through located at: https://github.com/jriveraespejo/paper1_manuscript

Code availability: The code is available in the different notebooks of this document located at: https://jriveraespejo.github.io/paper1_manuscript/

Authors' contributions: *Conceptualization:* S.G., S.dM., and J.M.R.E; *Data curation:* J.M.R.E.; *Formal Analysis:* J.M.R.E.; *Funding acquisition:* S.G. and S.dM; *Investigation:* S.G.; *Methodology:* S.G., S.dM., and J.M.R.E; *Project administration:* S.G. and S.dM.; *Resources:* S.G. and S.dM.; *Software: J.M.R.E.*; *Supervision:* S.G. and S.dM.; *Validation:* J.M.R.E.; *Visualization:* J.M.R.E.; *Writing - original draft:* J.M.R.E.; *Writing - review & editing:* S.G. and S.dM.

References

Baker, F. (1998). An investigation of the item parameter recovery characteristics of a gibbs sampling procedure. *Applied Psychological Measurement*, *22*(22), 153–169. https://doi.org/10.1177/01466216980222005

Baldwin, S., & Fellingham, G. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Journal of Psychological Methods*, *18*(2), 151–164. https://doi.org/10.1037/a0030642

Bayes, C., Bazán, J., & García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, *7*(4), 841–866. https://doi.org/10.1214/12-ba728

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

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

Carrasco, J., Ferrari, S., & Arellano-Valle, R. (2012). Errors-in-variables beta regression models. Retrieved from /url%7Bhttps://arxiv.org/abs/1212.0870%7D

Castellanos, I., Kronenberger, W., Beer, J., Henning, S., Colson, B., & 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. https://doi.org/10.1179/1754762813Y.0000000043

Chin, S., & 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. https://doi.org/10.3109/02699206.2014.926997

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

Choi, IH. (2023). The impact of measurement noninvariance across time and group in longitudinal item response modeling. *Asia Pacific Education Review*. https://doi.org/10.1007/s12564-023-09907-4

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *SSRN*. https://doi.org/http://dx.doi.org/10.2139/ssrn.3689437

Cox, R., McDaniel, D., Kent, J., & 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. https://doi.org/10.1044/jshr.3202.347

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Brito Trindade, D., Espinheira, P. L., Pinto Vasconcellos, K. L., Farfán Carrasco, J. M., & do Carmo Soares de Lima, M. (2021). Beta regression model nonlinear in the parameters with additive measurement errors in variables. *PLOS ONE*, *16*(7), 1–28. https://doi.org/10.1371/journal.pone.0254103

Deffner, D., Rohrer, J., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3). https://doi.org/10.1177/25152459221106366

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in bayesian growth mixture modeling. *Journal of Structural Equation Modeling*, *21*, 239–252. https://doi.org/10.1080/10705511.2014.882686

Depaoli, S. (2021). *Bayesian structural equation modeling*. The Guilford Press.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Dieteren, C., Bonfrer, I., Brouwer, W., & van Exel, J. (2023). Public preferences for policies promoting a healthy diet: A discrete choice experiment. *European Journal of Health Economics*, *24*, 1429–1440. https://doi.org/10.1007/s10198-022-01554-7

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. https://doi.org/10.1044/0161-1461(2010/09-0081)

Everitt, B., & Skrondal, A. (2010). *The cambridge dictionary of statistics*. Cambridge University Press.

Faes, J., De Maeyer, S., & Gillis, S. (2021). Speech intelligibility of children with an auditory brainstem implant: A triple-case study, 1–50.

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

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

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

Figueroa-Zúñiga, J., Carrasco, J., Arellano-Valle, R., & Ferrari, S. (2018). A bayesian approach to errors-in-variables beta regression. *Brazilian Journal of Probability and Statistics*, *32*(3), 559–582. https://doi.org/10.1214/17-bjps354

Figueroa-Zúñiga, J., Bayes, C., Leiva, V., & Liu, S. (2021). Robust beta regression modeling with errors-in-variables: A bayesian approach and numerical applications. *Statistical Papers*. https://doi.org/10.1007/s00362-021-01260-1

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

Freeman, V., Pisoni, D., Kronenberger, W., & 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. https://doi.org/10.1093/deafed/enx001

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC.

Ghosh, A. (2019). Robust inference under the beta regression model with application to health care studies. *Journal of Statistical Methods in Medical Research*, *28*(3), 871–888. https://doi.org/10.1177/0962280217738142

Gillis, S. (2018). Speech and language in congenitally deaf children with a cochlear implant. In E. Dattner & D. Ravid (Eds.), *Handbook of communication disorders: Theoretical, empirical, and applied linguistic perspectives* (pp. 765–792). De Gruyter Mouton. https://doi.org/10.1515/9781614514909-038

Holmes, W., Bolin, J., & Kelley, K. (2019). *Multilevel modeling using r (2nd edition)*. Chapman; Hall/CRC. https://doi.org/10.1201/9781351062268

Jeffreys, H. (1998). *Theory of probability*. Oxford University Press.

Jenkins, S. (2000). Cultural and linguistic miscues: A case study of international teaching assistant and academic faculty miscommunication. *International Journal of Intercultural Relations*, *24*(4), 477–501. https://doi.org/10.1016/S0147-1767(00)00011-0

Kangmennaang, J., Siiba, A., & Bisung, E. (2023). Does trust mediate the relationship between experiences of discrimination and health care access and utilization among minoritized canadians during COVID-19 pandemic? *Journal of Racial and Ethnic Health Disparities*. https://doi.org/10.1007/s40615-023-01809-w

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

Kent, R. D., Miolo, G., & Bloedel, S. (19943). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology*, *3*(2), 81–95. https://doi.org/10.1044/1058-0360.0302.81

Khwaileh, F., & Flipsen, P. (2010). Single word and sentence intelligibility in children with cochlear implants. *Clinical Linguistics & Phonetics*, *24*(9), 722–733. https://doi.org/10.3109/02699206.2010.490003

Kim, S., & Cohen, A. (1999). Accuracy of parameter estimation in gibbs sampling under the two-parameter logistic model. In *Annual meeting of the american educational research association*. American Educational Research Association. Retrieved from /url%7Bhttps://eric.ed.gov/?id=ED430012%7D

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. Retrieved from /url%7Bhttp://www.jstor.org/stable/2236703%7D

Lagerberg, T., Asberg, J., Hartelius, L., & Persson, C. (2014). Assessment of intelligibility using children's spontaneous speech: Methodological aspects. *International Journal of Language and Communication Disorders*, *49*(2), 228–239. https://doi.org/10.1111/1460-6984.12067

Lambert, P., Sutton, A., Burton, P., Abrams, K., & Jones, D. (2006). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Journal of Statistics in Medicine*, *24*(15), 2401–2428. https://doi.org/10.1002/sim.2112

Lebl, J. (2022). *Basic analysis i & II: Introduction to real analysis, volumes i & II*. Retrieved from /url%7Bhttps://www.jirka.org/ra/html/frontmatter-1.html%7D

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

Lopes, S., Shi, L., Pan, X., Gu, Y., Dengler-Crish, C., Yan Li, Y., et al. (2023). Meditation and cognitive outcomes: A longitudinal analysis using data from the health and retirement study 2000–2016. *Mindfulness*, *14*, 1705–1717. https://doi.org/10.1007/s12671-023-02165-w

MacWhinney, B. (2020). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates. https://doi.org/10.21415/3mhn-0z89

Martin, J., & McDonald, R. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for heywood cases. *Psychometrika*, (40), 505–517. https://doi.org/10.1007/BF02291552

Mayer, M. (1969). *Frog, where are you?* Dial Books for Young Readers. Retrieved from /url%7Bhttps://books.google.be/books?id=Asi5KQAACAAJ%7D

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and STAN*. Chapman; Hall/CRC.

Montag, J., AuBuchon, A., Pisoni, D., & 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. https://doi.org/10.1044/2014/_JSLHR-H-14-0190

Munro, M. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, *20*(2), 139–154. https://doi.org/10.1017/S0272263198002022

Munro, M., & Derwing, T. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, *48*(2), 159–182. https://doi.org/10.1111/1467-9922.00038

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). American Psychological Association. https://doi.org/10.1037/10409-010

Niparko, J., Tobey, E., Thal, D., Eisenberg, L., Wang, N., Quittner, A., & Fink, N. (2010). Spoken language development in children following cochlear implantation. *JAMA*, *303*(15), 1498–1506. https://doi.org/10.1001/jama.2010.451

Ockey, G., Papageorgiou, S., & French, R. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening*, *30*(1-2), 84–98. https://doi.org/0.1080/10904018.2015.1056877

Pereira, J., Nobre, W., Silva, I., & Schmidt, A. (2020). Spatial confounding in hurdle multilevel beta models: The case of the brazilian mathematical olympics for public schools. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(3), 1051–1073. https://doi.org/10.1111/rssa.12551

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2), 157--170. https://doi.org/10.1007/s10798-011-9189-x

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from /url%7Bhttp://www.R-project.org/%7D

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004a). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. https://doi.org/https://www.doi.org/10.1007/BF02295939

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004b). *GLLAMM manual.* UC Berkeley Division of Biostatistics. Retrieved from /url%7Bhttp://www.biostat.jhsph.edu/~fdominic/teaching/bio656/software-gllamm.manual.pdf%7D

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004c). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301–323. https://doi.org/https://www.doi.org/10.1016/j.jeconom.2004.08.017

Seaman, S. jr., J., & Stamey, J. (2011). Hidden dangers of specifying noninformative priors. *The American Statistician*, *66*(2), 77–84. https://doi.org/10.1080/00031305.2012.695938

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

Shmueli, G., & Koppius, O. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572. https://doi.org/10.2307/23042796

Simas, A. B., Barreto-Souza, W., & Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, *54*(2), 348–366. https://doi.org/https://doi.org/10.1016/j.csda.2009.08.017

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Chapman & Hall/CRC Press.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stan Development Team. (2021). *Stan modeling language users guide and reference manual, version 2.26*. Vienna, Austria. Retrieved from /url%7Bhttps://mc-stan.org%7D

Tackney, M., Morris, T., White, I., Leyrat, C., Diaz-Ordaz, K., & Williamson, E. (2023). A comparison of covariate adjustment approaches under model misspecification in individually randomized trials. *Trials*, *24*(14). https://doi.org/10.1186/s13063-022-06967-6

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

Unlu, H., & Aktas, S. (2017). Beta regression for the indicator values of well-being index for provinces in turkey. *Journal of Engineering Technology and Applied Sciences*, *2*(2), 101–111. https://doi.org/10.30931/jetas.321165

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.

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. https://doi.org/10.3366/E1753854809000305

Varonis, E., & Susan, G. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics*, *6*(1), 71–90. https://doi.org/10.1093/applin/6.1.71

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, PC. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

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

Verkuilen, J., & Smithson, M. (2013). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Ournal of Educational and Behavioral Statistics*, *37*(1), 82–113. https://doi.org/10.3102/1076998610396895

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897. Retrieved from /url%7Bhttps://www.jmlr.org/papers/volume14/watanabe13a/watanabe13a.pdf%7D

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

Zhang, J., Du, W., & Huang, F. and. (2023). Longitudinal study of dietary patterns and hypertension in adults: China health and nutrition survey 1991–2018. *Hypertension Research*, *46*, 2264–2271. https://doi.org/10.1038/s41440-023-01322-x