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

January 27, 2022

Abstract

High level description of the research.

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 [20, 42, 37, 14]. Because intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered [14], 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" [9].

The literature suggest there are two perspectives from which speech intelligibility (SI) can be assessed: the speaker and listener's perspective [4, 5]. 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) [32]. 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 [4, 5]. On both cases, the stimuli/representation (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) [17]. 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, known also as the relative holistic method.

It is clear 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 [5, 12]. However, given OR's demanding process in terms of time, number of listeners required, and finally cost, 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 [5], they suffer from a lack of intra- and inter-rater reliability[4] [28, 19, 17, 5]. Additionally, the literature inform us the procedure does not allow to assess subtle differences in the representations [5], while the scales derived from them are usually coarse, where children reach the higher levels fairly quickly [31], e.g. the Speech Intelligibility Rating (SIR) [10, 26].

In this context, CJ has received a growing attention, because it tackles directly the issues with the HJ procedures: it fosters reliable [40] and valid scores [7, 24], while the judges can focus only on the relevant aspects of the compared representations, i.e. the "just noticeable difference" [24]. 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 [15], it does not require a high level of expertise [24, 4], it encourages to tackle hard to operationlize or open-ended tasks [29, 30, 24], and the measurement of competencies [39], among others.

[1] ordered based on increasing level of ecological validity [13, 11]

[2] validity is defined as the extent to which scores are appropriate for their intended interpretation and use [24].

[3] measuring something consistently [source].

[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.

2 Research questions

Considering the previous, this proposal seeks to lay the ground 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 Judgement assumptions

On the one hand, HJ methods have their assumptions rooted in the Classical Test Theory (CTT)[5], 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.

On the other hand, CJ methods hinges on two principles: the law of comparative judgments [34], and the consensus of judges [24]. 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 [24]. This claim is supported by the fact that different listeners differ in the focus and broadness of their judgments [24], and that each representation is assessed by multiple judges, implying the final score is a reflection of the judges' collective expertise [30]. This only means that by combining various judgments, we come closer to the "true" rankings of SI [23].

3.1.2 Procedures

The HJ procedure consisted on two psycho-physical stages: (1) select and mentally represent the stimulus' information, independent of other stimuli, and (2) rate the representation, according to a task. Therefore, under this procedure, listeners rate the stimulus' SI in an absolute manner, e.g. on a 5-point scale going from "very unintelligible" to "very intelligible" [10, 21], or on a 10-point scale [5, 12].

In contrast, CJ is composed of three interrelated 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 [36]. As a result, in CJ-D, the listeners rate a pair

[5]`https://www.ncme.org/resources/glossary`

[6]for a detailed explanation of the law, see Thurstone [34], Verhavert [39] (p. 22-29)

of stimuli in a dichotomous way, e.g. if stimulus A is more intelligible than B you observe a one (1) in the outcome variable, and zero otherwise (0) [6]. On the other hand, under CJ-O, the listeners rate both stimuli on an ordinal scale, e.g. a 3-point scale where the outcome variable maps to the following preference relationships: $A >> B$, $A = B$, $A << B$, where $>>$, $<<$, and $=$ indicate the level of preference and indifference between the pairs, respectively [35, 1].

3.1.3 Experimental settings

On both procedures, the experimental settings for the **judgment task** followed the next steps [4, 5]:

1. the listener took a seat in front of a computer screen, located at his(her) home, workplace, or the experimental laboratory, and open Comproved[7]
2. two set of instructions were presented on the computer screen:
   (a) how to perform the task, and
   (b) the aspects not to consider for the task.
3. the listener hear the stimuli through high quality headphones, set at a comfortable volume.
4. the listener rate which stimulus sounded better, by selecting the appropriate button or slider[8] on the computer screen.

On the other hand, for the **transcription task**, the followed steps were similar as the previous. However, at the fourth step, the listeners did not rated the stimulus, but rather produced their orthographic transcriptions.

3.2 Stimuli

The stimuli consisted of children's utterances (sentences of similar length) recovered from a larger corpus of *spontaneously spoken speech* collected by CLiPS over the last twenty years. More specifically, we use a portion of the corpus that consisted of 10 utterances recordings for 32 7-year old children, telling a story cued by the picture book "Frog, where are you" [25] to a caregiver "who does not know the story". The utterances were selected making sure they did not have syntactically ill-formed or incomplete sentences, any background noise, cross-talk, long hesitations, revisions or non-words [5]. As a result, the data set consisted in a total of 320 utterances[9], that were presented to the listeners in a random order based on a pairing algorithm implemented in Comproved.

Similar designs were used by Boonen et al. [4] and Faes et al. [12]. 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 a pairwise comparison.

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

[8] The default setting for the slider is on the far left, however, evidence form the field of Behavioral Economics show that default settings might impact the rating procedure. Therefore, it would be advisable to randomize the default setting of the slider. source: Ariely, Thaler, Khaneman

[9] 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 [22], an experimental unit is the item under study upon which something is changed, and a replicate run is the experiment conducted with the same factor settings, but using different experimental units.

3.3 Comparisons / assessments

In terms the number of comparison per representation (stimulus) required, Verhavert [39] provided compelling evidence, that under CJ, between 17 and 20 comparisons were enough to achieved a Scale Separation Reliability (SSR) of 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[10].

3.4 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. We expect the CJ tasks to be 4-times more demanding, in terms of time and effort, than the HJ task, therefore, we allocate 4-times more judges in 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 judgment method.

On the other hand, for the transcription task, 100 transcribers participated in the experiment. The participants and stimuli were divided into five groups. As a consequence, each group of 20 students ($100/5$) transcribed 64 stimuli on their series ($320/5$), resulting in 20 transcriptions per utterance ($64 \times 100/320$)[11].

4 Statistical analysis

4.1 Data

4.1.1 Outcomes

On the one hand, the outcome from the **transcription task** was obtained following a two step procedure [5]. First, we aligned the participant's orthographic transcriptions at the utterance level. This step was repeated for every one of the 6400 transcription (64×100, see Section 3.4).

[10]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 [22], 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.

Lastly, we computed the entropy measure of the aligned transcriptions using equation (1) [33]. The last step resulted in 320 entropy measures[11].

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 $(H \rightarrow 0)$, while utterances yielding a low degree of agreement were considered as exhibiting low intelligibility and therefore registered a higher entropy $(H \rightarrow 1)$ [5, 12]. In that sense, entropy was defined as:

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

where N denotes the number of transcribers, p_i the utterance's probability of occurrence, and n the total number of events, i.e. two for the CJ-D case, and more than two for the CJ-O and HJ procedures.

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

It is important to mention that besides the exclusion of corrupted observations. e.g. no available rating, no other experimental run was eliminated before the modeling process. This decision departs from what it is observed in previous research, e.g. Boonen et al. [4] decided to eliminate observations based on misfit analysis [24], while van Daal [36] and Boonen et al. [5] did the same based on outlier analysis. For the case of misfit analysis, we argue that such procedures cannot be used without caution. The literature points out that in the context of CJ, these statistics are always relative, i.e. they depend on other stimulus and judges included in the assessment [29, 30], while they have been proven to be less sensitive, as they are calculated with a low number of judgments per representation [29]. On the other hand, for the case of outlier analysis, we argue that outlying observations cannot be identified properly outside the context of a full model [27], i.e. what can behave as an outlier based on a univariate analysis, can behave as expected under the appropriate model. Moreover, as stated by McElreath [27], outliers are interesting cases to analyze. Considering the previous, if we still manage to identify outlying observations within the context of the proposed models (see Section 4.2), the researcher would rather adjust the model, so it can be robust against the influence of such outliers.

4.1.2 Covariates

The sampled children registered the following characteristics regarding their hearing status:

1. Normal Hearing (NHC)
2. Hearing Impaired (HIC):
    - Cochlear Implant (CIC)
    - Hearing Aid (HAC)
    - Auditory Brain stem Implant (ABI)

[11] under DoE literature, the design corresponds to entropy measures for 32 experimental units with 10 replicate runs and 20 duplicates, making a total of 320 experimental runs.

4.2 Statistical modeling

Considering the objectives outlined in Section 2, this section describes the model selection procedures and the statistical models considered.

4.2.1 Model selection

Following the successful and comprehensive analysis in van Daal [36] and Lesterhuis [24], this research will also use the Information-Theoretic Approach (ITA) [3, 8] for the selection of competing models.

First, we will translate our working hypotheses into statistical models. In the present research, this step will be supported by the use of Directed Acyclic Graphs (DAG) and probabilistic programming [18]. A DAG is the simplest representation of a Graphical Causal Models (GCM), a heuristic model that contains information not purely statistical, but unlike a detailed statistical model, it allow us to deduce which variable relationships can provide valid causal inferences [16, 27], i.e. 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 a DAG are only valid if the assumed DAG is correct. On the other hand, the probabilistic programming will serve as the algebraic formalist to specify our probabilistic models.

Second, in order to select between competing models, we need to set an appropriate measure of what makes a model a better approximation of reality. McElreath [27] goes all the way to argument that the best measure of model fit is the out-of-sample predictive accuracy. In that sense, this research will embrace the full flexibility of our bayesian implementation (see Section 4.3) and use two criteria that provide a better approximations of the out-of-sample (cross-validated) deviance[12]: (1) the Widely Applicable Information Criterion (WAIC) [41], and (2) the Pareto-smoothed importance sampling cross-validation (PSIC) [38].

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

4.2.2 Models

We will consider two interrelated models: (1) a measurement error model for entropy, and (2) a measurement model for the CJ and HJ procedures, respectively.

Measurement error model for entropy:
As it is described in Section 4.1.1, our data set is composed of 320 entropy measures, nested within 32 children with 10 replicates per child. Each entropy measure was bounded in the continuum $[0, 1]$, as expected from equation (1).

Previous research have already used the entropy measure as an outcome [5, 12]. However, on those cases, the authors decided to aggregate the measure to a mean value, in order to ease its handling in modeling process. We argue this pre-aggregating procedure could be pernicious for a proper statistical inference, as ”anytime we use an average value, discarding the uncertainty around that average, we risk overconfidence and spurious inference” [27].

This claim is easier to understand using a though experiment within our research. For example, imagine we have two children with the same mean entropy, but the second child shows

[12] van Daal [36] used the Akaike’s Information Criterion (AIC) [2] with similar purposes.

more variability in the measure than the first. It is clear from the example that the average entropy measure informs about the child's average SI, indicating that both children have a similar level. However, the variability around such mean entropy also informs about the child's SI, as a higher variability imply transcribers agreed less about the second child intelligibility across the 10 utterances. A similiar intuition was presented in Boonen et al. [5], but the paper only used the information in a descriptive analysis, rather than integrate it to the modeling process.

We argue that the estimation of such measurement error model is trivial under the bayesian framework, and we present it in the following lines.

First, figure 1 depicts the DAG representation of the model. The figure shows the s'th observed entropy measure H_{is}^{O} nested within the i'th child, where $i = 1, \dots, N_c$, $s = 1, \dots, N_s$, with $N_c = 32$ and $N_s = 10$. Additionally, the figure reveals the observed entropy represents multiple instances of a "true" entropy H_i^T for each child, but measured with error (e_i). Finally, we notice covariates are set to explain the "true" entropy.

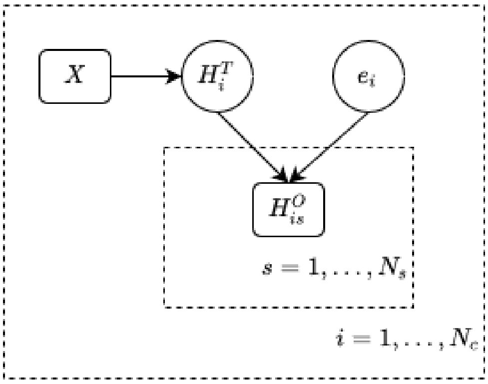

Figure 1: DAG for the measurement error model of entropy. Circles represent latent variables, squares observed values or covariates, and large squares the nesting within specific units.

(in process)

Measurement models for the CJ and HJ procedures:
(in process)

4.3 Estimation procedure

(in process)

4.4 Evaluation

4.4.1 validity

(in process)

4.4.2 reliability

(in process)

4.4.3 efficiency (time)

(in process)

References

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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).

[12] 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).

[13] 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.

[14] 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.

[15] 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.

[16] Hernán, M. and Robins, J. [n.d.]. *Causal Inference: What If*, Chapman and Hall/CRC.

[17] 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$_JSLHR-20-00008$.

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

[19] 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.

[20] 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.

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

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

[23] 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.

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

[25] Mayer, M. [1969]. *Frog, where are You?*, Boy, a Dog, and a Frog, Dial Books for Young Readers.

[26] 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.

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

[28] 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).

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

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

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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.

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

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