Bayesian modeling of comparative judgment data with `R` and `Stan`: A tutorial for speech quality researchers

Jose Manuel Rivera Espejo[a,*], Tine Daal[a], Sven Maeyer[a], Steven Gillis[b]

[a] *University of Antwerp, Training and education sciences,*

[b] *University of Antwerp, Linguistics,*

Abstract

The Bradley-Terry-Luce (BTL) model is commonly used to analyze comparative judgment (CJ) data because it provides a simple method for measuring traits and conducting statistical inference. Its simplicity stems from two key features: (1) a reliance on an extensive set of simplifying assumptions about the traits, judges, and stimuli involved in CJ assessments; and (2) the use of ad hoc procedures to handle inferences, including hypothesis testing. However, recent literature questions whether these assumptions hold in modern CJ applications and whether the ad hoc procedures effectively fulfill their intended analytical purpose.

To address these concerns, Rivera et al. (2025) proposed an approach that extends the general form of Thurstone's law of comparative judgment. The approach enables the development of a model tailored to the assumed data-generating process of the CJ system under study, eliminating the need to rely on simplifying assumptions. Moreover, by integrating measurement and inference within a single analytical framework, the approach also removes the dependence on ad hoc hypothesis-testing procedures.

This tutorial illustrates the application of the proposed approach to a simulated dataset on speech quality. It offers detailed guidance on data simulation, model specification, estimation, and interpretation using the software `R` and `Stan`. While the tutorial assumes familiarity with CJ theory and practice, latent variable models, and causal inference, it does not require prior experience with Bayesian inference methods or the associated software. Ultimately, by following the outlined procedures, researchers can replicate this analysis and adapt the approach for more complex CJ studies.

Keywords: tutorial, causal inference, bayesian inference, thurstonian model, comparative judgement, statistical modeling

*Corresponding author

Email addresses: `JoseManuel.RiveraEspejo@uantwerpen.be` (Jose Manuel Rivera Espejo),

1. Introduction

Comparative judgment (CJ) is an assessment method in which judges evaluate a trait across different stimuli using pairwise comparisons (Thurstone, 1927b,a). Each comparison generates a dichotomous outcome that indicates which stimulus is perceived to exhibit a higher attribute level. For instance, judges might compare pairs of short speech samples (the stimuli) to evaluate the relative speech quality of children (the trait) (Boonen et al., 2020).

The Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959) is then employed to analyze the CJ data, as it provides a simple method for measuring traits and conducting statistical inference (Andrich, 1978; Pollitt, 2012). The method's simplicity stems from two key features. First, it relies on an extensive set of simplifying assumptions about the traits, judges, and stimuli involved in CJ assessments (Thurstone, 1927a; Bramley, 2008). Second, it employs ad hoc procedures to handle inferences, including hypothesis testing (Pollitt, 2012).

However, recent studies question whether these assumptions hold in modern CJ applications (Bramley, 2008; Kelly et al., 2022; Rivera et al., 2025) and whether the ad hoc procedures achieve their intended analytical purpose (Kelly et al., 2022; Rivera et al., 2025). For instance, Rivera et al. (2025, pp. 2) argues that while assuming equal dispersions and zero correlation between stimuli simplifies the trait measurement model, these assumptions may fail to capture the complexity of some traits or account for heterogeneous stimuli (Thurstone, 1927b; Andrich, 1978; van Daal et al., 2016; Lesterhuis et al., 2018; Chambers and Cunningham, 2022). As a result, these assumptions can compromise the reliability and accuracy of trait estimates (Ackerman, 1989; Zimmerman, 1994; McElreath, 2020; Wu et al., 2022; Miller, 2023; Hoyle, 2023). Moreover, the same authors note that although ad hoc procedures simplify CJ data analysis, relying on untested methods can also undermine the validity of statistical inferences drawn from the data (McElreath, 2020; Kline, 2023; Hoyle, 2023).

To address these concerns, Rivera et al. (2025) proposed an approach that extends the general form of Thurstone's law of comparative judgment (Thurstone, 1927b,a) using causal and Bayesian inference methods. The approach combines Thurstone's core theoretical principles with key CJ assessment design features, enabling the development of a model tailored to the assumed data-generating process of the CJ system under study. This tailoring effectively eliminates the need to rely on the simplifying assumptions of the BTL model. Moreover, by integrating measurement and inference within a single analytical framework, the approach also removes the dependence on ad hoc procedures. Ultimately, the approach has the potential to yield reliable trait estimates and accurate statistical inferences. However, this promise still needs to be empirically tested.

tine.vandaal@uantwerpen.be (Tine Daal), sven.demaeyer@uantwerpen.be (Sven Maeyer), steven.gillis@uantwerpen.be (Steven Gillis)

Thus, this tutorial applies the proposed approach to a simulated dataset on speech quality to evaluate whether the approach's promise holds in practice. Specifically, we are interested in the following research questions:. At the same time, it provides detailed guidance on data simulation, model specification, estimation, and interpretation using the software `R` and `Stan`. Notably, while the tutorial assumes familiarity with CJ theory and practice, latent variable models, and causal inference, it does not require prior experience with Bayesian inference methods or the associated software. Ultimately, by following the procedures here outlined, researchers can replicate the analysis and adapt the approach to more complex CJ studies.

The remainder of this manuscript is organized into four sections. Section 2 describes the model specification, the dataset simulation, inference procedure, and evaluation metrics relevant to the research questions. Section 3 summarizes the analysis, including parameter estimates, credible intervals, and comparisons with the standard BTL model. Next, Section 4 reviews the findings, outlines future research directions, and discusses the limitation of the study. Finally, Section 5 provides the concluding remarks.

2. Methods

2.1. Model specification

2.2. Dataset simulation

2.3. Inference procedure

2.4. Evaluation metrics

3. Results

4. Discussion

4.1. Future research directions

4.2. Study limitations

5. Conclusion

Declarations

Funding: The Research Fund (BOF) of the University of Antwerp funded this project.

Financial interests: The authors declare no relevant financial interests.

Non-financial interests: The authors declare no relevant non-financial interests.

Ethics approval: The University of Antwerp Research Ethics Committee confirmed that this study does not require ethical approval.

Consent to participate: Not applicable

Consent for publication: All authors have read and approved the final version of the manuscript for publication.

Data availability: This study did not use any data.

Materials and code availability: A previous version of this manuscript, along with the associated materials and code (see the section titled `CODE LINK`), has been made publicly available at: https://jriveraespejo.github.io/paper3_manuscript/.

AI-assisted technologies in the writing process: The authors used various AI-based language tools to refine phrasing, optimize wording, and enhance clarity and coherence throughout the manuscript. They take full responsibility for the final content of the publication.

CRediT authorship contribution statement: *Conceptualization:* J.M.R.E, T.vD., S.DM., and S.G.; *Methodology:* J.M.R.E, T.vD., and S.DM.; *Software:* J.M.R.E.; *Validation:* J.M.R.E.; *Formal Analysis:* J.M.R.E.; *Investigation:* J.M.R.E; *Resources:* T.vD. and S.DM.; *Data curation:* J.M.R.E.; *Writing - original draft:* J.M.R.E.; *Writing - review and editing:* J.M.R.E., T.vD., S.DM., and S.G.; *Visualization:* J.M.R.E.; *Supervision:* S.G. and S.DM.; *Project administration:* S.G. and S.DM.; *Funding acquisition:* S.G. and S.DM.

References

Ackerman, T., 1989. Unidimensional irt calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement 13, 113–127. doi:10.1177/014662168901300201.

Andrich, D., 1978. Relationships between the thurstone and rasch approaches to item scaling. Applied Psychological Measurement 2, 451–462. doi:10.1177/014662167800200319.

Boonen, N., Kloots, H., 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:10.1016/j.jcomdis.2019.105969.

Bradley, R., Terry, M., 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39, 324–345. doi:10.2307/2334029.

Bramley, T., 2008. Paired comparison methods, in: Newton, P., Baird, J., Goldsteing, H., Patrick, H., Tymms, P. (Eds.), Techniques for monitoring the comparability of examination standards. GOV.UK., pp. 246—300. URL: https://assets.publishing.service.gov.uk/media/5a80d75940f0b62305b8d734/2007-comparability-exam-standards-i-chapter7.pdf.

Chambers, L., Cunningham, E., 2022. Exploring the validity of comparative judgement: Do judges attend to construct-irrelevant features? Frontiers in Education doi:10.3389/feduc.2022.802392.

Hoyle, R.e., 2023. Handbook of Structural Equation Modeling. Guilford Press.

Kelly, K., Richardson, M., Isaacs, T., 2022. Critiquing the rationales for using comparative judgement: a call for clarity. Assessment in Education: Principles, Policy & Practice 29, 674–688. doi:10.1080/0969594X.2022.2147901.

Kline, R., 2023. Principles and Practice of Structural Equation Modeling. Methodology in the Social Sciences, Guilford Press.

Lesterhuis, M., van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., De Maeyer, S., 2018. When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality. L1-Educational Studies in Language and Literature 18, 1–22. doi:10.17239/L1ESLL-2018.18.01.02.

Luce, R., 1959. On the possible psychophysical laws. The Psychologcal Review 66, 482–499. doi:10.1037/h0043178.

McElreath, R., 2020. Statistical Rethinking: A Bayesian Course with Examples in R and STAN. Chapman and Hall/CRC. doi:https://doi.org/10.1201/9780429029608.

Miller, J., 2023. Outlier exclusion procedures for reaction time analysis: The cures are generally worse than the disease. Journal of Experimental Psychology: General 152, 3189–3217. doi:10.1037/xge0001450.

Pollitt, A., 2012. The method of adaptive comparative judgement. Assessment in Education: Principles, Policy and Practice 19, 281—300. doi:10.1080/0969594X.2012.665354.

Rivera, J., van Daal, T., De Maeyer, S., Gillis, S., 2025. Let's talk about thurstone & Co.: an information-theoretical model for comparative judgments, and its statistical translation. URL: https://jriveraespejo.github.io/paper2_manuscript/. last accessed in 30-08-2025.

Thurstone, L., 1927a. A law of comparative judgment. Psychological Review 34, 482–499. doi:10.1037/h0070288.

Thurstone, L., 1927b. Psychophysical analysis. American Journal of Psychology , 368–89URL: https://brocku.ca/MeadProject/Thurstone/Thurstone_1927g.html. last accessed 20 december 2024.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., De Maeyer, S., 2016. Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. Assessment in Education: Principles, Policy & Practice 26, 59–74. doi:10.1080/0969594X.2016.1253542.

Wu, W., Niezink, N., Junker, B., 2022. A diagnostic framework for the bradley–terry model. Journal of the Royal Statistical Society Series A: Statistics in Society 185, S461–S484. URL: https://academic.oup.com/jrsssa/article-pdf/185/Supplement_2/S461/49421054/jrsssa_185_supplement_2_s461.pdf, doi:10.1111/rssa.12959.

Zimmerman, D., 1994. A note on the influence of outliers on parametric and nonparametric tests. The Journal of General Psychology 121, 391–401. doi:10.1080/00221309.1994.9921213.