Finding Thurstone: modeling comparative judgment data with `R` (and `Stan`)

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 classical BTL analysis has become the standard approach for analyzing comparative judgment (CJ) data because it provides a simple method for measuring traits and conducting related analysis. This simplicity arises from two key features. First, the approach relies on the Bradley-Terry-Luce (BTL) model to estimate latent traits. Second, it uses of ad hoc procedures to conduct related analysis. However, recent studies question whether the BTL model assumptions hold in contemporary CJ applications and whether the ad hoc procedures effectively fulfill their intended analytical goals.

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 for simplifying assumptions. Moreover, by integrating measurement and inference within a single analytical framework, it also removes the dependence on ad hoc analytical procedures. Despite these advantages, the approach still requires empirical validation.

Thus, this study empirically validates the proposed Information-Theoretical model for CJ, benchmarked against the classical BTL analysis, and demonstrates its practical implementation. The document includes a structured tutorial based on a simulated speech-quality dataset, providing guidance on data simulation, prior specification, model estimation, and interpretation using `Stan`, `R`, and the interface packages `cmdstan` and `brms`. Ultimately, the study equips researchers with practical tools to apply the model to 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), `tine.vandaal@uantwerpen.be` (Tine Daal), `sven.demaeyer@uantwerpen.be` (Sven Maeyer), `steven.gillis@uantwerpen.be` (Steven Gillis)

1. Introduction

Comparative judgment (CJ) has emerged as a valuable methodology for measuring latent traits across diverse fields, including education (Kimbell, 2012; Jones and Inglis, 2015; van Daal et al., 2016; Bartholomew et al., 2018), political sciences (ZuccoJr. et al., 2019), linguistics (Boonen et al., 2020), and criminology (Seymour and Hernandez, 2025). In CJ studies, judges actively compare pairs of stimuli to determine which stimulus exhibits more of the latent trait of interest (Thurstone, 1927b,a).

A specific data analysis workflow has become the standard approach for analyzing CJ data (see, e.g., Thwaites and Paquot, 2024). In this study, we refer to this workflow as the *classical BTL analysis*. Researchers favor this approach because it provides a simple method for measuring traits and conducting related analyses (Andrich, 1978; Pollitt, 2012). This simplicity, in turn, arises from two key features. First, the approach relies on the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959) to estimate latent traits. The model facilitates trait estimation by imposing several simplifying assumptions about traits, judges, and stimuli in CJ assessments (Thurstone, 1927a; Bramley, 2008). Second, the approach uses ad hoc procedures to conduct related analysis, including data summarization and statistical inference (Pollitt, 2012).

Recent studies, however, question whether the assumptions of the BTL model hold in contemporary CJ applications and whether the ad hoc procedures achieve their intended analytical goals (Bramley, 2008; Kelly et al., 2022; Rivera et al., 2025). For instance, Rivera and colleagues (2025) argue that while the assumptions of equal dispersions and zero correlations between stimuli simplify trait measurement, they may fail to represent complex traits or heterogeneous stimuli adequately (Thurstone, 1927b; Andrich, 1978; van Daal et al., 2016; Lesterhuis et al., 2018; Chambers and Cunningham, 2022). As a result, such 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 data analyses, the use of untested methods can also undermine the validity of statistical inferences derived from CJ 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). This approach leverages causal and Bayesian inference methods to combine Thurstone's core theoretical principles with key design features of CJ assessment. By doing so, it enables the development of a model tailored to the assumed data-generating process of the CJ system under study. This tailoring effectively removes 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 eliminates the dependence on ad hoc analytical procedures. Ultimately, this approach has the potential to produce reliable trait estimates and accurate statistical inferences. However, its effectiveness still required empirical validation.

1.1. Research goals

Building on the introduction, this study pursues two closely related research goals. The first goal is to *empirically validate* the proposed Information-Theoretical model for CJ (hereafter, ITCJ analysis) by evaluating the accuracy and reliability of its trait estimates and inference parameters, benchmarked against the classical BTL analysis (hereafter, CBTL analysis). The second goal emerges as a practical byproduct of this validation: to demonstrate how to implement the model in practice. To this end, the document provides a structured tutorial based on a simulated speech-quality dataset, offering guidance on data simulation, prior specification, model estimation, and interpretation using `Stan`(Stan Development Team., 2026b,a), `R` (R Core Team, 2015), and the interface packages `cmdstan` (Gabry et al., 2025) and `brms` (Bürkner, 2017, 2018). By combining model validation and practical instruction, the study can evaluate the methodological performance of the ITCJ analysis and provide researchers with practical tools to apply it to more complex CJ studies.

The remainder of this manuscript is organized into five sections. Section 2 reviews the two analytical approaches commonly applied to CJ data: the CBTL and ITCJ analyses. Section 3 details the assumed data-generating process for the simulated dataset, the simulation procedure, the practical implementation of each analytical approach, and the evaluation criteria aligned with the research goals. Section 4 presents the data description and modeling results. Section 5 interprets the findings, outlines future research directions, and considers the study limitations. Finally, Section 6 offers the concluding remarks.

2. A tale of two analytical approaches

2.1. The classical BTL analysis

2.2. The Information-Theoretical model for CJ

3. Methods

3.1. Step 1, from Theory to Design: Data-generating assumptions

3.2. Step 2, from Design to Data: Data simulation

3.3. Step 5, from Estimator and Sample to Estimate(s): The analysis aproaches

3.3.1. The CBTL analysis

3.3.2. The ITCJ analysis

3.3.2.1. Model 1.

3.3.2.2. Model 2.

3.3.2.3. Model 3.

3.3.2.4. Model 4.

3.3.2.5. Model 5.

3.3.2.6. Model 6.

3.4. Step 6, from Estimate(s) to Diagnostics and Posterior predictives: The evaluation criteria

4. Results

4.1. Data description

4.2. Data modeling

4.2.1. The CBTL analysis

4.2.2. The ITCJ analysis

4.2.2.1. Model 1.

4.2.2.2. Model 2.

4.2.2.3. Model 3.

4.2.2.4. Model 4.

4.2.2.5. Model 5.

4.2.2.6. Model 6.

4.2.2.7. Model comparison.

5. Discussion

5.1. Future research directions

5.2. Study limitations

6. 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, materials and code availability: A previous version of this manuscript, along with the associated data, 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.

7. Appendix

7.1. Appendix A: Stationarity, converge and mixing

7.2. Appendix B: Misfit observations

7.3. Appendix C: Sample size calculations

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.

Bartholomew, S., Nadelson, L., Goodridge, W., Reeve, E., 2018. Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. Educational Assessment 23, 85–101. doi:10.1080/10627197.2018.1444986.

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.

Bürkner, P.C., 2017. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80, 1–28. doi:10.18637/jss.v080.i01.

Bürkner, P.C., 2018. Advanced Bayesian multilevel modeling with the R package brms. The R Journal 10, 395–411. doi:10.32614/RJ-2018-017.

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.

Gabry, J., Češnovar, R., Johnson, A., Bronder, S., 2025. cmdstanr: R Interface to 'CmdStan'. URL: https://mc-stan.org/cmdstanr/.

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

Jones, I., Inglis, M., 2015. The problem of assessing problem solving: can comparative judgement help? Educational Studies in Mathematics 89, 337–355. doi:10.1007/s10649-015-9607-1.

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.

Kimbell, R., 2012. Evolving project e-scape for national assessment. International Journal of Technology and Design Education 22, 135–155. doi:10.1007/s10798-011-9190-4.

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.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

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.

Seymour, R.G., Hernandez, F., 2025. Scalable bayesian inference for bradley–terry models with ties: an application to honour based abuse. Journal of Applied Statistics 52, 1695–1712. doi:10.1080/02664763.2024.2436608.

Stan Development Team., 2026a. Stan Reference Manual, version 2.38.0. Vienna, Austria. URL: https://mc-stan.org.

Stan Development Team., 2026b. Stan Userś Guide, version 2.38.0. Vienna, Austria. URL: https://mc-stan.org.

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.

Thwaites, P., Paquot, M., 2024. Comparative judgement for advancing research in applied linguistics. Research Methods in Applied Linguistics 3, 100142. URL: https://www.sciencedirect.com/science/article/pii/S277276612400048X, doi:https://doi.org/10.1016/j.rmal.2024.100142.

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.

ZuccoJr., C., Batista, M., Power, T.J., 2019. Measuring portfolio salience using the bradley–terry model: An illustration with data from brazil. Research & Politics 6, 2053168019832089. doi:10.1177/2053168019832089.