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 and colleagues (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 or *CBTL analysis*. Researchers favor this approach because it provides a simple method for measuring traits and conducting related analyses (Andrich, 1978; Pollitt, 2012b). This simplicity 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 present 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, 2012b).

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 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), referred to as the Information-Theoretical model for CJ (hereafter, ITCJ analysis). 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.

1.1. Research goals

The ITCJ analysis (Rivera et al., 2025) shows theoretical promise for yielding reliable trait estimates and accurate statistical inferences. However, as noted by the authors, this promise has not yet been empirically validated. Thus, the present study addresses this gap by pursuing two closely related research goals. The first goal is to *empirically validate* the proposed ITCJ analysis by evaluating the accuracy and reliability of its trait estimates and inference parameters, benchmarked against the 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 evaluates the methodological performance of the ITCJ analysis and equips 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

Pairwise comparison data, and more specifically CJ data, can be analyzed using two main approaches: the CBTL and the ITCJ analysis. The CBTL approach applies a sequence of separate analytical steps to estimate traits and draw inferences. In contrast, the ITCJ analysis uses a single, systematic, and integrated approach to achieve the same objectives. This section describes the two approaches in detail.

2.1. The CBTL analysis

The CBTL approach implements a sequence of separate analytical steps, each serving a distinct purpose (Pollitt, 2012a,b; Jones et al., 2019; Boonen et al., 2020; Chambers and Cunningham, 2022; Bouwer et al., 2023). This multi-step procedure typically unfolds as follows. First, analysts apply the

BTL model to the CJ data to produce two outputs: (1) point estimates of stimulus traits along with their standard errors, and (2) residuals at the stimulus level. These outputs provide the foundation for the subsequent analyses.

Second, researchers summarize or fit regression models to the stimulus point estimates. This step serves multiple purposes, including aggregating stimulus-level estimates to the individual level, partitioning variability between and within individuals, and drawing inferences about factors that influence trait values. For example, Boonen et al. (2020) applied a multilevel regression model to the stimulus point estimates to examine whether children's age or hearing status affects their intelligibility scores.

Third, analysts summarize or fit regression models to the BTL residuals. This step helps to aggregate the remaining variability at the judge level, partition residual variability between and within judges, test for systematic biases, and identify potential misfitting judgments, stimuli, or judges. For instance, Wu (2025) fitted an analysis of variance (ANOVA) model to the infit statistic for each rater to examine the potential effects of raters' expertise on their judgments. The infit statistics is a weighted average of the squared Pearson residuals (Wright and Masters, 1982).

While this stepwise approach is the standard practice in fields such as education (see, Wu, 2025) and linguistics (see, Thwaites and Paquot, 2024), it presents several limitations. First, each stage treats outputs from previous steps as fixed data, rather than acknowledging their status as uncertain parameter estimates. Failure to account for this uncertainty can introduce bias and decrease the precision of inferences. The direction and magnitude of these biases can be unpredictable: results may be attenuated, amplified, or remain unaffected depending on the uncertainty in the scores and the actual effects being tested (McElreath, 2020; Kline, 2023; Hoyle, 2023). Moreover, the loss of precision diminishes statistical power and increases the likelihood of committing type I or type II errors (McElreath, 2020). Second, the procedure lacks theoretical coherence, as it combines models with different underlying assumptions. For example, third-step analyses treat residuals as outputs that ostensibly capture deviations from expected comparison outcomes, potentially reflecting judge-specific tendencies or idiosyncrasies in judgments, without providing evidence on whether these assumptions hold.

2.2. The ITCJ analysis

The ITCJ analysis addresses the aforementioned limitations by providing a unified and systematic approach to analyzing CJ data (Rivera et al., 2025). It starts with a general *Directed Acyclic Graph (DAG)* and a corresponding *Structural Causal Model (SCM)* (Morgan and Winship, 2014; Gross et al., 2018; Neal, 2020), which together establish a coherent theoretical foundation for CJ analysis by explicitly representing the relationships among observed judgments, discriminal differences, stimulus traits, individual traits, judge biases, and the sampling and comparison mechanisms. Next, the

approach adapts the general SCM and DAG to the assumed data-generating process of the CJ system under study. Then, it derives one or more bespoke *probabilistic* and *statistical* models tailored to that system. Finally, it uses one or more statistical models to estimate traits and conduct statistical inference. Figure 1 illustrates an example of a general DAG structure.

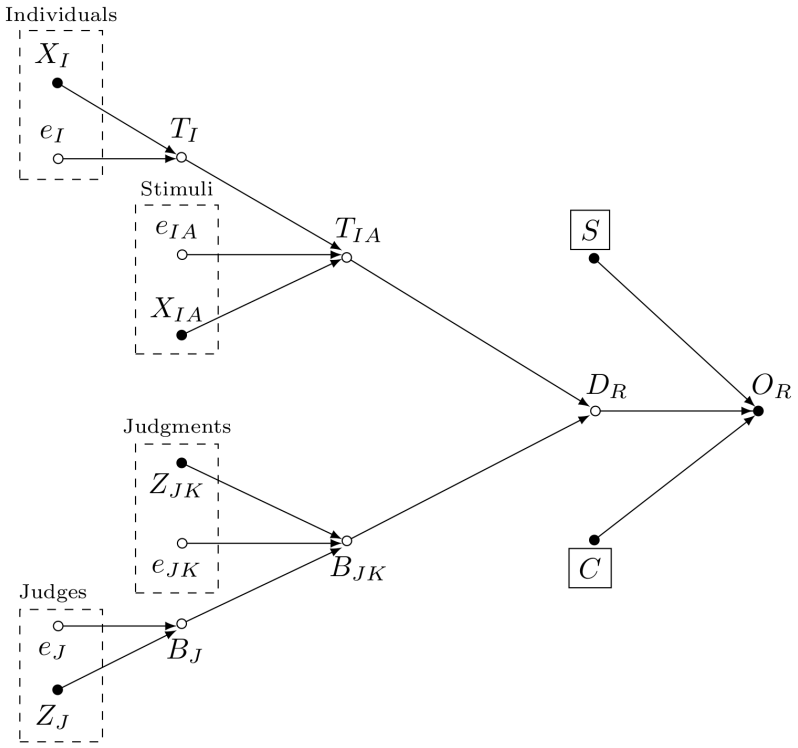

Figure 1: ITCJ analysis: DAG

In this representation, O_R denotes the observed judgment outcome vector, and D_R represents the discriminal difference vector. T_{IA} captures the vector of stimulus traits, and T_I represents the vector of individual traits. The vector of judgment biases is represented by B_{JK}, while the vector of judges' biases by B_J. Covariates at the stimulus and individual levels appear as X_{IA} and X_I, respectively; while Z_{JK} and Z_J represent covariates at the judgment and judge levels. The error terms (e_{IA}, e_I, e_{JK}, e_J) capture residual variation at each level.

The DAG also represents the sampling and comparison mechanisms as the vectors S and C, two conditioned variables that determine how population outcomes become "observed" outcomes. Importantly, the DAG depicts S and C as independent from all other variables by showing no arrows pointing into them. Regarding S, this indicates that the DAG applies to *Simple Random Sampling (SRSg)* designs, where each repeated judgment, judge, stimulus, and individual has an equal probability of inclusion within their respective groups (Lawson, 2015). Regarding C, the DAG applies to *Random Allocation Comparative Designs* (Bramley, 2015) or *Incomplete Block Designs* (Lawson, 2015), where every repeated judgment has an equal chance of being included in the sample.

Researchers can then translate this DAG representation into SCMs and probabilistic forms that express the joint distribution of a complex CJ system as a product of simpler *conditional probability distributions (CPDs)* (Pearl et al., 2016; Neal, 2020; see also Rivera et al., 2025, section 5), as illustrated in Figure 2.

$$
\begin{aligned}
O_R &:= f_O(D_R, S, C) \\
D_R &:= f_D(T_{IA}, B_{JK}) \\
T_{IA} &:= f_T(T_I, X_{IA}, e_{IA}) \\
T_I &:= f_T(X_I, e_I) \\
B_{JK} &:= f_B(B_J, Z_{JK}, e_{JK}) \\
B_J &:= f_B(Z_J, e_J)
\end{aligned}
\qquad
\begin{aligned}
&P(O_R \mid D_R, S, C) \\
&P(D_R \mid T_{IA}, B_{JK}) \\
&P(T_{IA} \mid T_I, X_{IA}, e_{IA}) \\
&P(T_I \mid X_I, e_I) \\
&P(B_{JK} \mid B_J, Z_{JK}, e_{JK}) \\
&P(B_J \mid Z_J, e_J)
\end{aligned}
$$

$$
\begin{aligned}
e_I &\perp \{e_J, e_{IA}, e_{JK}\} \\
e_J &\perp \{e_{IA}, e_{JK}\} \\
e_{IA} &\perp e_{JK}
\end{aligned}
\qquad
P(e_I)P(e_{IA})P(e_J)P(e_{JK})
$$

(a) SCM (b) Probabilistic model

Figure 2: ITCJ analysis. SCM (left) and probabilistic (right) representations for DAG in Figure 1.

Critically, the approach allows tailoring this general structure to a specific CJ context, enabling development of parsimonious models that match the assumed data-generating process without imposing unnecessary constraints. For example, Rivera et al. (2025) modeled a CJ assessment designed to evaluate the impact of different teaching methods on students' writing ability. In this case, the observed outcome was binary, so the model assumed O_R followed a Bernoulli distribution. The discriminal difference (D_R) was determined by the texts' discriminal processes (T_{IA}) and the judgment biases (B_{JK}). Student-level variables X_I, such as teaching method, were included, whereas text-level variables X_{IA} (e.g., text length) were not gathered. Similarly, judge-level variables Z_J, like judgment expertise, were incorporated, while judgment-level variables Z_{JK} (e.g., number of judgments per judge) were absent. Finally, the probabilistic assumptions for the idiosyncratic errors (e_I, e_{IA}, e_J, e_{JK}) resolved indeterminacies in *location*, *orientation*, and *scale* for the variables T_I, T_{IA}, B_J, B_{JK}, as required in latent variable models (Depaoli, 2021; de Ayala, 2009).

Lastly, researcher can derive one or more *bespoke* statistical models tailored to the CJ system of interest, as demonstrated in Rivera et al. (2025). At this stage, the ITCJ analysis differs fundamentally from the CBTL approach in how it handles parameter estimation and inference. Rather than fitting multiple separate models, the approach simultaneously estimates all parameters within a single coherent framework using *Bayesian inference*. This joint estimation accounts for

uncertainty at all levels and enables direct inference about quantities of interest without relying on post-hoc procedures (McElreath, 2020).

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.

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., De Maeyer, S., 2023. Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. Journal of Writing Research 15, 497–518. doi:10.17239/jowr-2024.15.03.03.

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.

Bramley, T., 2015. Investigating the reliability of adaptive comparative judgment. URL: http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf. cambridge Assessment Research Report.

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.

de Ayala, R., 2009. The Theory and Practice of Item Response Theory. Methodology in the Social Sciences, The Guilford Press.

Depaoli, S., 2021. Bayesian Structural Equation Modeling. Methodology in the social sciences, The Guilford Press.

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

Gross, J., Yellen, J., Anderson, M., 2018. Graph Theory and Its Applications. Textbooks in Mathematics, Chapman and Hall/CRC. doi:https://doi.org/10.1201/9780429425134. 3rd edition.

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

Jones, I., Bisson, M., Gilmore, C., Inglis, M., 2019. Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? British Educational Research Journal 45, 662–680. doi:10.1002/berj.3519.

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.

Lawson, J., 2015. Design and Analysis of Experiments with R. Chapman and Hall/CRC.

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.

Morgan, S., Winship, C., 2014. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Analytical Methods for Social Research. 2 ed., Cambridge University Press.

Neal, B., 2020. Introduction to causal inference from a machine learning perspective. URL: https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf. last accessed 30 April 2024.

Pearl, J., Glymour, M., Jewell, N., 2016. Causal Inference in Statistics: A Primer. John Wiley & Sons, Inc.

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

Pollitt, A., 2012b. 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.

Wright, B., Masters, G., 1982. Rating scale analysis. MESA Press. URL: https://research.acer.edu.au/measurement/2.

Wu, Q., 2025. Comparative judgment: Building a shared consensus over rater variation in assessing second language writing performance. Sage Open 15, 21582440251346346. doi:10.1177/21582440251346346.

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.