Let’s talk about Thurstone & Co.: An information-theoretical model for comparative judgments, and its statistical translation

Jose Manuel Rivera Espejo

Tine van van Daal

Sven De De Maeyer

Steven Gillis

2024-11-08

Abstract

(to do)

# Introduction

In *comparative judgment* (CJ) studies, judges assess a specific trait or competence across various stimuli by performing pairwise comparisons (Thurstone 1927; Pollitt 2004, 2012a). Each comparison produces a dichotomous outcome, indicating which stimulus is perceived to manifest a higher trait level. For example, when assessing text quality, judges compare pairs of written texts (the stimuli) to determine the relative quality each text exhibit (the trait) (Laming 2004; Pollitt 2012b; Whitehouse 2012; van Daal et al. 2016; Lesterhuis 2018; Coertjens et al. 2017; Goossens and De Maeyer 2018; Bouwer et al. 2023).

Numerous studies have documented the effectiveness of CJ in assessing traits and competencies over the past decade. These studies have emphasized three aspects of the method’s effectiveness: its reliability, validity, and practical applicability. Research on reliability indicates that CJ requires a relatively small number of pairwise comparisons (S. Verhavert et al. 2019; Crompvoets, Béguin, and Sijtsma 2022) to produce trait scores that are as precise and consistent as those generated by other assessment methods (Coertjens et al. 2017; Goossens and De Maeyer 2018; Bouwer et al. 2023). Furthermore, evidence suggests that the reliability and time efficiency of CJ are comparable, if not superior, to those of other assessment methods when employing adaptive comparison algorithms (Pollitt 2012b; San Verhavert, Furlong, and Bouwer 2022; Mikhailiuk et al. 2021). On the other hand, research on validity suggests that scores generated by CJ can accurately represent the traits under measurement (Whitehouse 2012; van Daal et al. 2016; Lesterhuis 2018; Bartholomew et al. 2018; Bouwer et al. 2023). Finally, research on practical applicability highlights the method’s versatility across both educational and non-educational contexts (Jones 2015; Bartholomew et al. 2018; Jones et al. 2019; Marshall et al. 2020; Bartholomew and Williams 2020; Boonen, Kloots, and Gillis 2020).

Nevertheless, despite the growing number of CJ studies, unsystematic and fragmented research approaches in the literature have left several critical issues unaddressed. This research primarily focuses on three: the apparent disconnect between CJ’s measurement and structural model, the over-reliance on the assumptions of Thurstone’s Case 5 (1927) in CJ’s measurement model, and the unclear role of comparison algorithms on the method’s reliability and validity. The following sections will discuss each of these issues in detail, followed by the introduction of a theoretical model and its statistical translation, which aims to address all three concerns simultaneously.

# Three critical issues in CJ literature

## The disconnect between structural and measurement models

In a typical CJ study, the Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952; Luce 1959) serves as the measurement model for CJ (Andrich 1978; Bramley 2008). A measurement model specifies how manifest variables contribute to the estimation of latent variables (Everitt and Skrondal 2010). For example, when evaluating text quality, researchers use the BTL model to process the dichotomous outcomes resulting from the pairwise comparisons (the manifest variables) to estimate scores that reflect the underlying quality level of the texts (the latent variable) (Laming 2004; Pollitt 2012b; Whitehouse 2012; van Daal et al. 2016; Lesterhuis 2018; Coertjens et al. 2017; Goossens and De Maeyer 2018; Bouwer et al. 2023).

Researchers then typically use the estimated BTL scores, or their transformations, to conduct additional analyses or hypothesis tests. For example, these scores have been used to identify ‘misfit’ judges and stimuli (Pollitt 2012b; van Daal et al. 2017; Goossens and De Maeyer 2018), detect biases in judges’ ratings (Pollitt and Elliott 2003; Pollitt 2012b), calculate correlations with other assessment methods (Goossens and De Maeyer 2018; Bouwer et al. 2023), or test hypotheses related to the underlying trait of interest (Bramley and Vitello 2019; Boonen, Kloots, and Gillis 2020; Bouwer et al. 2023; van Daal et al. 2017; Jones et al. 2019; Gijsen et al. 2021).

However, the statistical literature advises caution when using estimated scores to conduct additional analyses or hypotheses tests. A key consideration is that BTL scores are parameter estimates that inherently carry uncertainty. Ignoring this uncertainty can introduce bias into the analysis and reduce the precision of hypothesis tests. Notably, the direction and magnitude of the bias are often unpredictable; results may be attenuated, exaggerated, or remain unaffected, depending on the amount of uncertainty present in the scores and the actual effects being tested (Kline 2023, 25; Hoyle 2023, 137). Furthermore, reduced precision in hypothesis tests weakens their statistical power, ultimately increasing the likelihood of committing type-I or type-II errors (McElreath 2020).

To mitigate these risks, principles from Structural Equation Modeling (SEM) (Hoyle 2023, 138) and Item Response Theory (IRT) (Fox 2010, chap. 6; van der Linden 2017, vol. 1, chap. 24) recommend conducting these analyses and tests within a structural model. A structural model specifies how different manifest or latent variables influence the latent variable of interest (Everitt and Skrondal 2010). This approach allows analyses that can account for both the scores and their uncertainties simultaneously, rather than treating them as separate elements. Therefore, an integrated approach that combines CJ’s measurement and structural models can offer significant advantages.

## The assumptions of Case 5 and the measurement model

## The role and impact of comparison algorithms

# Theory

## A theoretical model for CJ

## From theory to statistics

# Discussion

## Findings

## Limitations and further research

# Conclusion

# Declarations

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

**Financial interests:** The authors have no relevant financial interest to disclose.

**Non-financial interests:** Author XX serve on advisory broad of Company Y but receives no compensation this role.

**Ethics approval:** 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:** No data was utilized in this study.

**Code availability:** All the code utilized in this research is available in the digital document located at: <https://jriveraespejo.github.io/paper2_manuscript/>.

**AI-assisted technologies in the writing process:** The authors used ChatGPT, an AI language model, during the preparation of this work. They occasionally employed the tool to refine phrasing and optimize wording, ensuring appropriate language use and enhancing the manuscript’s clarity and coherence. The authors take full responsibility for the final content of the publication.

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

# Appendix

# References

Andrich, D. 1978. “Relationships Between the Thurstone and Rasch Approaches to Item Scaling.” *Applied Psychological Measurement* 2 (3): 451–62. <https://doi.org/10.1177/014662167800200319>.

Bartholomew, S., L. Nadelson, W. Goodridge, and E. Reeve. 2018. “Adaptive Comparative Judgment as a Tool for Assessing Open-Ended Design Problems and Model Eliciting Activities.” *Educational Assessment* 23 (2): 85–101. <https://doi.org/10.1080/10627197.2018.1444986>.

Bartholomew, S., and P. Williams. 2020. “STEM Skill Assessment: An Application of Adaptive Comparative Judgment.” In *Integrated Approaches to STEM Education. Advances in STEM Education*, edited by J. Anderson and Y. Li, 331–49. Springer. <https://doi.org/10.1007/978-3-030-52229-2_18>.

Boonen, N., H. Kloots, and S. Gillis. 2020. “Rating the Overall Speech Quality of Hearing-Impaired Children by Means of Comparative Judgements.” *Journal of Communication Disorders* 83: 1675–87. <https://doi.org/10.1016/j.jcomdis.2019.105969>.

Bouwer, R., M. Lesterhuis, F. De Smedt, H. Van Keer, and S. De Maeyer. 2023. “Comparative Approaches to the Assessment of Writing: Reliability and Validity of Benchmark Rating and Comparative Judgement.” *Journal of Writing Research* 15 (3): 497–518. <https://doi.org/10.17239/jowr-2024.15.03.03>.

Bradley, R., and M. Terry. 1952. “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.” *Biometrika* 39 (3-4): 324–45. <https://doi.org/10.2307/2334029>.

Bramley, T. 2008. “Paired Comparison Methods.” In *Techniques for Monitoring the Comparability of Examination Standards*, edited by P. Newton, J. Baird, H. Goldsteing, H. Patrick, and P. Tymms, 246--300. GOV.UK. <https://www.gov.uk/government/publications/techniques-for-monitoring-the-comparability-of-examination-standards>.

Bramley, T., and S. Vitello. 2019. “The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement.” *Assessment in Education: Principles, Policy and Practice* 71 (9): 1–25. <https://doi.org/10.1080/0969594X.2017.1418734>.

Coertjens, L., M Lesterhuis, S. Verhavert, R. Van Gasse, and S. De Maeyer. 2017. “Teksten Beoordelen Met Criterialijsten of via Paarsgewijze Vergelijking: Een Afweging van Betrouwbaarheid En Tijdsinvestering.” *Pedagogische Studien* 94: 283–303. <https://repository.uantwerpen.be/docman/irua/e71ea9/147930.pdf>.

Crompvoets, E., A. Béguin, and K. Sijtsma. 2022. “On the Bias and Stability of the Results of Comparative Judgment.” *Frontiers in Education* 6. <https://doi.org/10.3389/feduc.2021.788202>.

Everitt, B., and A. Skrondal. 2010. *The Cambridge Dictionary of Statistics*. Cambridge University Press.

Fox, J. P. 2010. *Bayesian Item Response Modeling, Theory and Applications*. Statistics for Social and Behavioral Sciences. Springer.

Gijsen, M., T. van Daal, Marije Lesterhuis, David Gijbels, and Sven De Maeyer. 2021. “The Complexity of Comparative Judgments in Assessing Argumentative Writing: An Eye Tracking Study.” *Frontiers in Education* 5. <https://doi.org/10.3389/feduc.2020.582800>.

Goossens, M., and S. De Maeyer. 2018. “How to Obtain Efficient High Reliabilities in Assessing Texts: Rubrics Vs Comparative Judgement.” In *Technology Enhanced Assessment*, edited by E. Ras and A. Guerrero Roldán, 13–25. Springer International Publishing. <https://doi.org/10.1007/978-3-319-97807-9_2>.

Hoyle, R. (eds.). 2023. *Handbook of Structural Equation Modeling*. Guilford Press.

Jones, I. 2015. “The Problem of Assessing Problem Solving: Can Comparative Judgement Help?” *Educational Studies in Mathematics* 89 (3): 337–55. <https://doi.org/10.1007/s10649-015-9607-1>.

Jones, I., M. Bisson, C. Gilmore, and M. Inglis. 2019. “Measuring Conceptual Understanding in Randomised Controlled Trials: Can Comparative Judgement Help?” *British Educational Research Journal* 45 (3): 662–80. <https://doi.org/10.1002/berj.3519>.

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

Laming, D. 2004. “Marking University Examinations: Some Lessons from Psychophysics.” *Psychology Learning & Teaching* 3 (2): 89–96. <https://doi.org/10.2304/plat.2003.3.2.89>.

Lesterhuis, M. 2018. “The Validity of Comparative Judgement for Assessing Text Quality: An Assessor’s Perspective.” PhD thesis, University of Antwerp.

Luce, R. 1959. “On the Possible Psychophysical Laws.” *The Psychologcal Review* 66 (2): 482–99. <https://doi.org/10.1037/h0043178>.

Marshall, N., K Shaw, J. Hunter, and I. Jones. 2020. “Assessment by Comparative Judgement: An Application to Secondary Statistics and English in New Zealand.” *New Zealand Journal of Educational Studies* 55: 49–71. <https://doi.org/10.1007/s40841-020-00163-3>.

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

Mikhailiuk, A., C. Wilmot, M. Perez-Ortiz, D. Yue, and R. Mantiuk. 2021. “Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization.” In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2559–66. <https://doi.org/10.1109/ICPR48806.2021.9412676>.

Pollitt, A. 2004. “Let’s Stop Marking Exams.” In *Proceedings of the IAEA Conference*. Philadelphia: University of Cambridge Local Examinations Syndicate. <https://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf>.

———. 2012a. “Comparative Judgement for Assessment.” *International Journal of Technology and Design Education* 22 (2): 157--170. <https://doi.org/10.1007/s10798-011-9189-x>.

———. 2012b. “The Method of Adaptive Comparative Judgement.” *Assessment in Education: Principles, Policy and Practice* 19 (3): 281--300. <https://doi.org/10.1080/0969594X.2012.665354>.

Pollitt, A., and G. Elliott. 2003. “Finding a Proper Role for Human Judgement in the Examination System.” University of Cambridge Local Examinations Syndicate. <https://www.cambridgeassessment.org.uk/Images/109707-monitoring-and-investigating-comparability-a-proper-role-for-human-judgement.pdf>.

Thurstone, L. 1927. “A Law of Comparative Judgment.” *Psychological Review* 34 (4): 482–99. <https://doi.org/10.1037/h0070288>.

van Daal, T., M. Lesterhuis, L. Coertjens, V. Donche, and S. De Maeyer. 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 (1): 59–74. <https://doi.org/10.1080/0969594X.2016.1253542>.

van Daal, T., M. Lesterhuis, L. Coertjens, MT. van de Kamp, V. Donche, and Sven De Maeyer. 2017. “The Complexity of Assessing Student Work Using Comparative Judgment: The Moderating Role of Decision Accuracy.” *Frontiers in Education* 2. <https://doi.org/10.3389/feduc.2017.00044>.

van der Linden, W., ed. 2017. *Handbook of Item Response Theory: Models*. Vol. 1. Statistics in the Social and Behavioral Sciences Series. CRC Press.

Verhavert, San, Antony Furlong, and Renske Bouwer. 2022. “The Accuracy and Efficiency of a Reference-Based Adaptive Selection Algorithm for Comparative Judgment.” *Frontiers in Education* 6. <https://doi.org/10.3389/feduc.2021.785919>.

Verhavert, S., R. Bouwer, V. Donche, and S. De Maeyer. 2019. “A Meta-Analysis on the Reliability of Comparative Judgement.” *Assessment in Education: Principles, Policy and Practice* 26 (5): 541–62. <https://doi.org/10.1080/0969594X.2019.1602027>.

Whitehouse, C. 2012. “Testing the Validity of Judgements about Geography Essays Using the Adaptive Comparative Judgement Method.” Centre for Education Research & Policy. <https://filestore.aqa.org.uk/content/research/CERP_RP_CW_24102012_0.pdf?download=1>.