

Causes and effects in Dichotomous Comparative Judgments: an information-theoretical system of plausible mechanism

Jose Manuel Rivera Espejo^{a,*}, Tine van Daal^a, Sven De Maeyer^a, Steven Gillis^b

^a*University of Antwerp, Training and education sciences,*

^b*University of Antwerp, Linguistics,*

Abstract

(to do)

Keywords: causal inference, probability, Thurstone, comparative judgement, directed acyclic graph, structural causal models, statistical modeling

1. Introduction

Over the past decade, numerous studies have documented the effectiveness of the *comparative judgement* (CJ) method (Thurstone, 1927) for assessing competencies and traits. These studies have evaluated CJ from two main perspectives: its ability to produce reliable and valid trait scores, and its practical applicability. Research on reliability and validity has shown that CJ can generate precise and consistent scores that accurately represent the traits being measured (Pollitt, 2012a,b; Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018; Bramley and Vitello, 2019; Verhavert et al., 2019; Crompvoets et al., 2022; Bouwer et al., 2023). Regarding practical applicability, several studies have highlighted CJ's versatility across both educational and non-educational contexts, presenting it as an efficient and effective alternative for measurement and evaluation (Jones, 2015; Bartholomew et al., 2018; Jones et al., 2019; Marshall et al., 2020; Bartholomew and Williams, 2020; Boonen et al., 2020).

Despite the growing number of CJ studies, the unsystematic and fragmented research approaches employed in the literature have overlooked several critical issues concerning the method. These issues fall into three main categories: concerns about the structural model, the measurement model, and the experimental design of CJ. In the following sections, each issue will be discussed in detail, followed by the introduction of an approach that addresses all three concerns simultaneously.

*Corresponding author

Email addresses: JoseManuel.RiveraEspejo@uantwerpen.be (Jose Manuel Rivera Espejo), tine.vandaal@uantwerpen.be (Tine van Daal), sven.demaeyer@uantwerpen.be (Sven De Maeyer), steven.gillis@uantwerpen.be (Steven Gillis)

A key issue in the first category is the apparent disconnect between the method's structural and measurement models. In CJ literature, it is common to perform data analysis and hypothesis testing on scores estimated using the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959). Several studies use the scores generated by the BTL model or their transformations to identify 'misfit' judges and stimuli (Pollitt, 2012b; van Daal et al., 2017), detect 'bias' in judges' ratings (Pollitt and Elliott, 2003; Pollitt, 2012b), or test various hypotheses about the underlying trait being measured (Bramley and Vitello, 2019; Boonen et al., 2020; Bouwer et al., 2023; van Daal et al., 2017; Jones et al., 2019; Gijzen et al., 2021). However, since these scores are parameter estimates with inherent uncertainty, the statistical literature suggests that separating the analysis from this uncertainty may artificially inflate the precision and power of the results. This, in turn, could increase the risk of committing a type I error, where a null hypothesis is wrongly rejected (McElreath, 2020). To address this issue properly, the approach should follow a strategy similar to that used in Structural Equation Modeling (SEM), where data analysis and hypothesis testing occur at the structural model level, while the BTL model functions as the measurement model.

2. Theory

- 2.1. Let's talk about Thurstone co.*
- 2.2. A scientific model for the CJ*
- 2.3. From theory to statistical model*

3. Discussion

- 3.1. Findings*
- 3.2. Limitations and further research*

4. 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 board 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://jriverspejo.github.io/paper2_manuscript/.

Authors' contributions: *Conceptualization:* S.G., S.D.M., T.vD., and J.M.R.E.; *Methodology:* S.D.M., 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.D.M., and T.vD.; *Data curation:* J.M.R.E.; *Writing - original draft:* J.M.R.E.; *Writing - review & editing:* S.G., S.D.M., and T.vD.; *Visualization:* J.M.R.E.; *Supervision:* S.G. and S.D.M.; *Project administration:* S.G. and S.D.M.; *Funding acquisition:* S.G. and S.D.M.

5. Appendix

References

- 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](https://doi.org/10.1080/10627197.2018.1444986).
- Bartholomew, S., Williams, P., 2020. Stem skill assessment: An application of adaptive comparative judgment, in: Anderson, J., Li, Y. (Eds.), *Integrated Approaches to STEM Education*. Advances in STEM Education. Springer, pp. 331–349. doi:[10.1007/978-3-030-52229-2_18](https://doi.org/10.1007/978-3-030-52229-2_18).
- 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](https://doi.org/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. URL: <https://www.jowr.org/index.php/jowr/article/view/867>, doi:[10.17239/jowr-2024.15.03.03](https://doi.org/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. URL: <http://www.jstor.com/stable/2334029>, doi:[10.2307/2334029](https://doi.org/10.2307/2334029).
- Bramley, T., Vitello, S., 2019. The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice* 71, 1–25. doi:[10.1080/0969594X.2017.1418734](https://doi.org/10.1080/0969594X.2017.1418734).
- Crompvoets, E.A.V., Béguin, A.A., Sijtsma, K., 2022. On the bias and stability of the results of comparative judgment. *Frontiers in Education* 6. URL: [url{https://www.frontiersin.org/articles/10.3389/feduc.2021.788202}](https://www.frontiersin.org/articles/10.3389/feduc.2021.788202), doi:[10.3389/feduc.2021.788202](https://doi.org/10.3389/feduc.2021.788202).
- Gijzen, M., van Daal, T., Lesterhuis, M., Gijbels, D., De Maeyer, S., 2021. The complexity of comparative judgments in assessing argumentative writing: An eye tracking study. *Frontiers in Education* 5. URL: [url{https://www.frontiersin.org/articles/10.3389/feduc.2020.582800}](https://www.frontiersin.org/articles/10.3389/feduc.2020.582800), doi:[10.3389/feduc.2020.582800](https://doi.org/10.3389/feduc.2020.582800).
- Jones, I., 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](https://doi.org/10.1007/s10649-015-9607-1).
- 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. URL: <https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1002/berj.3519>, doi:[10.1002/berj.3519](https://doi.org/10.1002/berj.3519).
- Lesterhuis, M., 2018. The validity of comparative judgement for assessing text quality: An assessor’s perspective. Ph.D. thesis. University of Antwerp.
- Luce, R., 1959. On the possible psychophysical laws. *The Psychological Review* 66, 482–499. doi:[10.1037/h0043178](https://doi.org/10.1037/h0043178).
- Marshall, N., Shaw, K., Hunter, J., Jones, I., 2020. Assessment by comparative judgement: An application to secondary statistics and english in new zealand. *New Zealand Journal of Educational Studies* 55, 49–71. doi:[10.1007/s40841-020-00163-3](https://doi.org/10.1007/s40841-020-00163-3).
- McElreath, R., 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. Chapman and Hall/CRC.
- Pollitt, A., 2012a. Comparative judgement for assessment. *International Journal of Technology and Design Education* 22, 157–170. doi:[10.1007/s10798-011-9189-x](https://doi.org/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](https://doi.org/10.1080/0969594X.2012.665354).
- Pollitt, A., Elliott, G., 2003. Finding a proper role for human judgement in the examination system. URL: <https://www.cambridgeassessment.org.uk/Images/109707-monitoring-and-investigating-comparability-a-proper-role-for-human-judgement.pdf>. research & Evaluation Division.
- Thurstone, L., 1927. A law of comparative judgment. *Psychological Review* 34, 482–499. doi:[10.1037/h0070288](https://doi.org/10.1037/h0070288).
- 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](https://doi.org/10.1080/0969594X.2016.1253542).
- van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M., Donche, V., De Maeyer, S., 2017. The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. *Frontiers in Education* 2. URL: <https://www.frontiersin.org/articles/10.3389/feduc.2017.00044>, doi:[10.3389/feduc.2017.00044](https://doi.org/10.3389/feduc.2017.00044).

- Verhavert, S., Bouwer, R., Donche, V., De Maeyer, S., 2019. A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy and Practice* 26, 541–562. doi:[10.1080/0969594X.2019.1602027](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. URL: https://filestore.aqa.org.uk/content/research/CERP_RP_CW_24102012_0.pdf?download=1. aQA Education.