

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

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 judgment* (CJ) method (Thurstone, 1927) for assessing competencies and traits. These studies have evaluated CJ from three main perspectives: its ability to produce reliable and valid trait scores, its practical applicability, and its time efficiency. Research on reliability and validity shows that CJ can generate precise and consistent scores (Pollitt, 2012a,b; Coertjens et al., 2017; Goossens and De Maeyer, 2018; Verhavert et al., 2019; Crompvoets et al., 2022; Bouwer et al., 2023) that accurately represent the traits being measured (Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018; Bouwer et al., 2023). Research on practical applicability highlights CJ's versatility across both educational and non-educational contexts, presenting it as an efficient and effective alternative for measurement and evaluation (Pollitt, 2004; Jones, 2015; Bartholomew et al., 2018; Jones et al., 2019; Marshall et al., 2020; Bartholomew and Williams, 2020; Boonen et al., 2020). Lastly, research on time efficiency suggests that CJ can offer at least equal, if not significant, time savings when evaluating stimuli compared to traditional marking methods (Pollitt, 2012a,b; Coertjens et al., 2017; Goossens and De Maeyer, 2018).

Nevertheless, despite the growing number of studies on CJ, unsystematic and fragmented research approaches in the literature have left several critical issues unaddressed.

*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)

Preprint submitted to *Psychometrika*

October 25, 2024

This research primarily focuses on three: the apparent disconnect between CJ’s structural and measurement model, the over-reliance on the assumptions of Thurstone’s Case 5 (1927) in CJ’s measurement model, and the unclear role and impact 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 aim to address all three concerns simultaneously.

2. Three critical issues in CJ literature

2.1. *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 (Andrich, 1978; Bramley, 2008). This model specifies how latent variables are estimated from manifest variables (Everitt and Skrondal, 2010). In CJ, multiple judges engage in several rounds of pairwise comparisons to assess the relative manifestation of a trait between two stimuli. Each comparison generates a dichotomous outcome, indicating which stimulus is perceived to exhibit a higher degree of the trait. The BTL model then uses these observed outcomes to estimate scores that represent the latent trait of interest (Pollitt, 2012a,b; Whitehouse, 2012; Jones, 2015; van Daal et al., 2016; Lesterhuis, 2018; Boonen et al., 2020; Bouwer et al., 2023).

Moreover, researchers often use these BTL-generated scores or their transformations separately, conducting further analyses or hypothesis testing. The literature shows how these scores have been employed 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 scoring methods (Goossens and De Maeyer, 2018; Bouwer et al., 2023), or test hypotheses related to the trait of interest (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, the statistical literature cautions against using estimated scores for separate analyses and tests. A key consideration is that BTL-generated scores are parameter estimates that inherently carry uncertainty. Ignoring this uncertainty when conducting separate analyses and tests can inflate their precision and statistical power, increasing the risk of committing a type I error (McElreath, 2020), which is when a null hypothesis is incorrectly rejected (Everitt and Skrondal, 2010). To mitigate this risk, principles from Structural Equation Modeling (SEM) (Hoyle, 2023; Kline, 2023) and Item Response Theory (IRT) (de Ayala, 2009; Fox, 2010; van der Linden, 2017) recommend conducting these analyses and tests within a structural model that accounts for both the scores and their uncertainties, rather than treating them separately. Thus, an integrated approach combining CJ’s structural and measurement models can offer significant advantages.

2.2. *The assumptions of Case 5 and the measurement model*

1. the BTL model is the statistical articulation of Case 5

From early on in the literature, it has been clear that the BTL model represents a statistical articulation of Thurstone’s Case 5 (1927). Talk about Pollitt and Elliott (2003) and Bramley (2008)

2. But this model hold the greater number of assumptions

Although Case 5 was originally articulated to produce a “rather coarse scaling” of traits (Thurstone, 1927, p. 269), its assumptions have become predominant in the literature, but Thurstone’s Case 5 is the case that holds the greater amount of simplifying assumptions.

Thurstone (1927) justify the use of case V, on multiple assumptions, but the most important for our purpose are three:

- (a) related to Case 3, it assumes the correlation between stimuli is zero, this translates into the cancellation of judges effects by mean of opposing and equally weighed ‘mood’ and ‘simultaneous contrasts’ effects. This is demonstrated using the additive nature of the logit scale, which helps to cancel ‘bias’ judges effects.
- (b) related to Case 3, it assumes the correlation between stimuli is zero, this also translates into the idea that stimuli are the main focus of estimation and analysis, but what happens when the focus of analysis is the individuals that generated those stimuli. Meaning there is an amount of correlation that it is not accounted for.
- (c) related to case 5, discriminial dispersions of the stimuli are equal, but it is not hard to imagine that certain individuals can produce good quality texts with more precision than other individuals.

3. What problems does this bring?

This leaves issues such as the presence of judge’ biases hinted by Bramley (2008) and Kelly et al. (2022), and evidenced by Pollitt and Elliott (2003), ?, and ?

Reliability is a necessary but not sufficient condition for validity. Reliability can exist without validity but validity cannot exist without reliability (Perron and Gillespie, 2015).

2.3. The role and impact of comparison algorithms

3. Theory

3.1. A theoretical model for CJ

3.2. From theory to statistics

4. Discussion

4.1. Findings

4.2. Limitations and further research

5. 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.

6. Appendix

References

- Andrich, D., 1978. Relationships between the thurstone and rasch approaches to item scaling. *Applied Psychological Measurement* 2, 451–462. doi:[10.1177/014662167800200319](https://doi.org/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](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., 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://www.gov.uk/government/publications/techniques-for-monitoring-the-comparability-of-examination-standards>.
- 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).
- Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., De Maeyer, S., 2017. Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering. *Pedagogische Studien* 94, 283–303. URL: <https://repository.uantwerpen.be/docman/irua/e71ea9/147930.pdf>.
- Crompvoets, E.A.V., Béguin, A.A., Sijtsma, K., 2022. On the bias and stability of the results of comparative judgement. *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).
- de Ayala, R., 2009. *The Theory and Practice of Item Response Theory*. Methodology in the Social Sciences, The Guilford Press.
- Everitt, B., Skrondal, A., 2010. *The Cambridge Dictionary of Statistics*. Cambridge University Press.
- Fox, J., 2010. *Bayesian Item Response Modeling, Theory and Applications*. Statistics for Social and Behavioral Sciences, Springer.
- 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).
- Goossens, M., De Maeyer, S., 2018. How to obtain efficient high reliabilities in assessing texts: Rubrics vs comparative judgement, in: Ras, E., Guerrero Roldán, A. (Eds.), *Technology Enhanced Assessment*, Springer International Publishing. pp. 13–25. doi:[10.1007/978-3-319-97807-9_2](https://doi.org/10.1007/978-3-319-97807-9_2).
- Hoyle, R.e., 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, 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).
- 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](https://doi.org/10.1080/0969594X.2022.2147901).
- Kline, R., 2023. *Principles and Practice of Structural Equation Modeling*. Methodology in the Social Sciences, Guilford Press.
- 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.
- Perron, B., Gillespie, D., 2015. Reliability and Measurement Error, in: *Key Concepts in Measurement*. Oxford University Press. doi:[10.1093/acprof:oso/9780199855483.003.0004](https://doi.org/10.1093/acprof:oso/9780199855483.003.0004).
- Pollitt, A., 2004. Let's stop marking exams, in: *Proceedings of the IAEA Conference, University of Cambridge Local Examinations Syndicate, Philadelphia*. URL: <https://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf>.
- 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).
- van der Linden, W. (Ed.), 2017. *Handbook of Item Response Theory*. volume 1-3 of *Statistics in the Social and Behavioral Sciences Series*. CRC Press.
- 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.