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: Probability, Directed Acyclic Graphs, Bayesian methods, Thurstonian model, Comparative judgement, Structural Causal Models, Statistical modeling

1. Introduction

In *comparative judgment* (CJ) studies, judges assess a specific trait or attribute across various stimuli by performing pairwise comparisons (Thurstone, 1927b,a). Each comparison produces a dichotomous outcome, indicating which stimulus is perceived to exhibit 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, 2012; 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 (Verhavert et al., 2019; Crompvoets et al., 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, 2012; Verhavert et al., 2022; Mikhailiuk et al., 2021). On the other hand, research on

*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)

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 et al., 2020).

Nevertheless, despite the increasing number of CJ studies, unsystematic and fragmented research approaches have left several critical issues unaddressed. This research primarily focuses on three: the over-reliance on Thurstone's Case V assumptions in the statistical analysis of CJ data, the apparent disconnect between CJ's trait measurement and hypothesis testing, 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.

2. Three critical issues in CJ literature

In its most general form, Thurstone's theory (1927a) posits that the dichotomous outcome resulting from comparing two stimuli is determined by two factors: the discriminal process of each stimulus and their discriminal difference. The *discriminal process* refers to the psychological effect each stimulus has on the judges, or more simply stated, the judges' perception of the trait level of each stimulus. Thurstone assumes that the discriminal process for each stimulus follows a Normal distribution. In this distribution, the mode (mean), known as the *modal discriminal process*, represents the position of the stimulus on the trait continuum, while the dispersion, known as the *discriminal dispersion*, reflects variability in the stimulus' perceived trait level. Figure 1 shows example distributions of discriminal process for two stimuli (objects).

However, since the discriminal mode and dispersion of a single stimulus are not directly observable except through comparison, the *law of comparative judgment* becomes essential. This law asserts that when assessing a specific trait by comparing two stimuli, the stimulus positioned further along the continuum is perceived as having a higher level of that trait. Thus, the observed dichotomous outcome is determined by the distribution of the difference between the stimuli's discriminal processes, called the *discriminal difference*. Figure 2 shows an example distribution of the discriminal difference for two stimuli (objects).

Importantly, the theory's general form primarily addresses pairwise comparisons of stimuli made by a single judge (Thurstone, 1927a, pp. 267). Consequently, to enhance its practical applicability, Thurstone introduced five distinct cases, each defined by progressively simplifying assumptions. Table 1 summarizes these cases, focusing on key assumptions such as the distribution of discriminal processes, the similarity of discriminal dispersions across stimuli, the correlation between stimuli, and which judges perform the comparisons. For a comprehensive discussion of this progression, refer to Thurstone (1927a) and Bramley (2008, pp. 248-253).

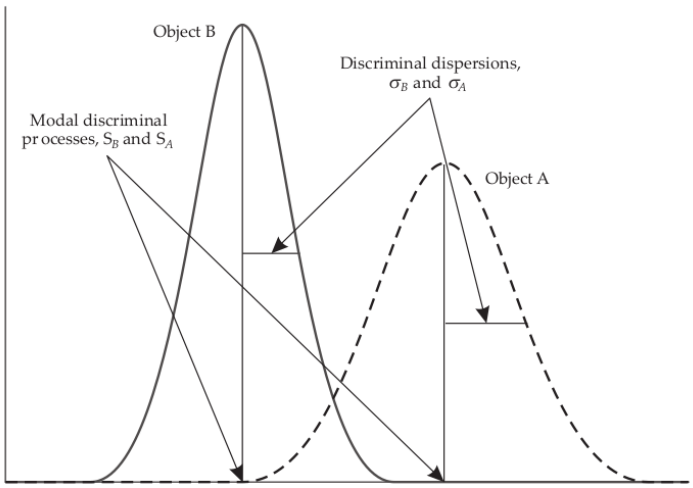


Figure 1: Example distributions of discriminal processes for two stimuli (objects). Extracted from Bramley (2008, pp. 249).

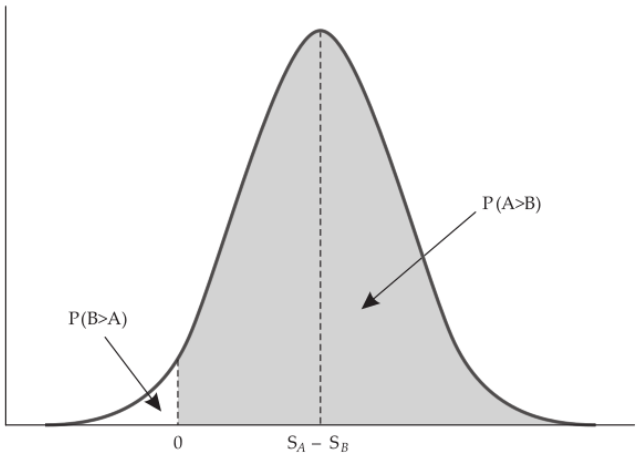


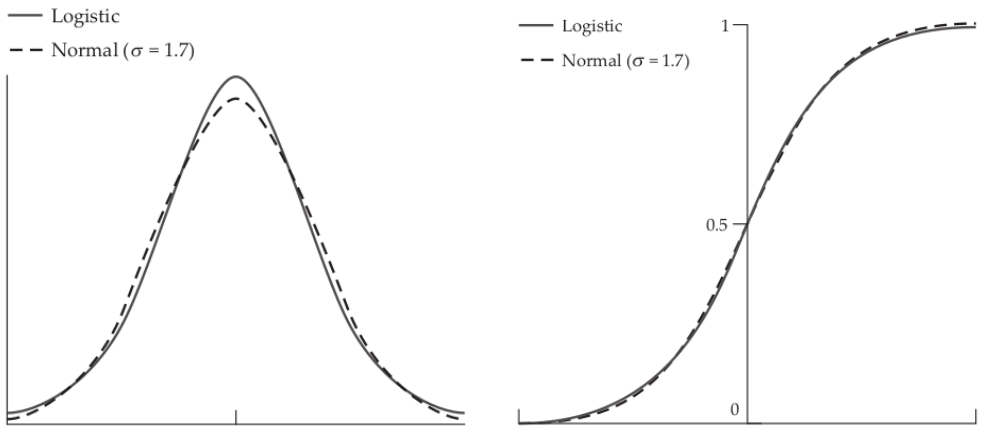
Figure 2: Distribution of the discriminal difference between two stimuli (objects). Extracted from Bramley (2008, pp. 251).

Table 1: Thurstones cases and asumptions

Assumption	Thurstone's					BTL
	Case I	**Case II**	**Case III**	**Case IV**	**Case V**	**model**
Discriminal process (distribution)	Normal	Normal	Normal	Normal	Normal	Logistic
Discriminal dispersion (between stimuli)	Different	Different	Different	Similar	Equal	Equal
Correlation (between stimuli)	Constant	Constant	Zero	Zero	Zero	Zero
Which judges compare?	Single	Multiple	Multiple	Multiple	Multiple	Multiple

2.1. The Case V and the statistical analysis of CJ data

Surprisingly, despite its reliance on the largest number of simplifying assumptions (Bramley, 2008, pp. 253), Case V remains the most widely used case in the CJ literature. This popularity is largely due to its simplified statistical representation in the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959). The BTL model mirrors Case V's assumptions, with one key difference: while Case V assumes a Normal distribution for the stimuli's discriminal processes, the BTL model uses the more mathematically tractable Logistic distribution (Bramley, 2008, pp. 254) (see Table 1). This substitution has little effect on the model's estimation or interpretation, as the Normal and Logistic distributions differ by a scaling factor of approximately 1.7 (van der Linden, 2017, pp. 16) (refer to Figure 3).



(a) Probability density

(b) Cummulative probability

Figure 3: Probability density and cumulative probability of the logistic and Normal distributions. Extracted from Bramley (2008, pp. 254-255).

However, Case V (and its BTL model counterpart) was originally developed to provide a "rather coarse scaling" of traits (Thurstone, 1927a, pp. 269), prioritizing statistical simplicity over precision in trait measurement. As a result, its assumptions may not suit applications beyond the psychophysical contexts for which it was created. Thur-

stone himself cautioned that its use "should not be made without (an) experimental test" (Thurstone, 1927a, pp. 270), acknowledging that some assumptions could prove problematic with more complex, less homogeneous stimuli, such as handwriting or English compositions (Thurstone, 1927b, pp. 374). Consequently, given that current CJ applications often deal with complex, less homogeneous stimuli, two key assumptions of Case V may not consistently hold in theory or practice: the zero correlation and equal dispersion between stimuli.

The *zero correlation between stimuli* represents two additional assumtions about the method:

2.2. The disconnect between trait measurement and hypothesis testing

Building on the previous section, it is evident that the BTL model typically functions as the measurement model for the trait of interest (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 texts' quality level (the latent variable) (Laming, 2004; Pollitt, 2012; 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, 2012; van Daal et al., 2017; Goossens and De Maeyer, 2018), detect biases in judges' ratings (Pollitt and Elliott, 2003; Pollitt, 2012), 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 et al., 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, pp. 25; Hoyle, 2023, pp. 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, pp. 138) and Item Response Theory (IRT) (Fox, 2010, chap. 6; van der Linden, 2017, 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 BTL 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.

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.

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.

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.

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.

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. URL: https://www.jowr.org/index.php/jowr/article/view/867, 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. URL: http://www.jstor.com/stable/2334029, 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://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.

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., Béguin, 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}, doi:10.3389/feduc.2021.788202.

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.

Gijsen, 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}, doi: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.

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.

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.

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, 89–96. doi:10.2304/plat.2003.3.2.89.

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 Psychologcal Review 66, 482–499. doi: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.

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

Mikhailiuk, A., Wilmot, C., Perez-Ortiz, M., Yue, D., Mantiuk, R., 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization, in: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2559–2566. doi:10.1109/ICPR48806.2021.9412676.

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.

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., 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 25 january 2025.

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.

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.

van der Linden, W. (Ed.), 2017. Handbook of Item Response Theory: Models. volume 1 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.

Verhavert, S., Furlong, A., Bouwer, R., 2022. The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. Frontiers in Education 6. URL: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2021.785919, doi:10.3389/feduc.2021.785919.

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.