

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, 2012b; Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018a; 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, 2012b; Verhavert et al., 2022; Mikhailiuk et al., 2021). Meanwhile, 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)

Preprint submitted to *Psychometrika*

December 6, 2024

validity suggests that scores generated by CJ can accurately represent the traits under measurement (Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018a; Bartholomew et al., 2018; Bouwer et al., 2023), while research on practical applicability highlights the method’s versatility across both educational and non-educational contexts (Kimbell, 2012; Jones and Inglis, 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. The present study primarily focuses on two: the overreliance on the assumptions of Thurstone’s Case V in the statistical analysis of CJ data, and the apparent disconnect between CJ’s approach to trait measurement and hypothesis testing. The following sections begin with a brief overview of Thurstone’s theory and a detailed examination of these issues. Subsequently, the study introduces a theoretical model for CJ that builds upon Thurstone’s theory, alongside its statistical translation, designed to address the two concerns simultaneously.

2. Thurstone’s theory

In its most general form, Thurstone’s theory addresses pairwise comparisons of stimuli made by a single judge (Thurstone, 1927a, pp. 267). The theory posits that two key factors determine the dichotomous outcome of these comparisons: the discriminial process of each stimulus and their discriminial difference. The *discriminal process* represents the psychological impact each stimulus has on judges or, more simply, their perception of the stimulus trait. According to the theory, the discriminial process for each stimulus follows a Normal distribution, where its mode (mean), called the *modal discriminial process*, indicates the stimulus position on the trait continuum, while its dispersion, known as the *discriminal dispersion*, reflects variability in the perceived trait of the stimulus.

Figure 1 illustrates the discriminial processes along a quality trait continuum for two written texts. The figure shows that these processes individually follow a Normal distribution. Furthermore, it shows that the modal discriminial process for Text B is positioned further along the continuum than Text A ($S_B > S_A$). Additionally, the figure highlights the broader distribution of Text B compared to Text A as a result of its larger discriminial dispersion ($\sigma_B > \sigma_A$).

However, since the discriminial processes of the stimuli are not directly observable, the theory introduces the *law of comparative judgment*. This law posits that in pairwise comparisons, a judge perceives the stimulus positioned further along the trait continuum as having a higher level of that trait. This emphasizes that the relative distance between stimuli, rather than their absolute positions on the continuum, determines the outcome of the pairwise comparison. Indeed, the theory assumes that the observed dichotomous outcome arises from the difference between the underlying discriminial processes of the stimuli, referred to as the *discriminal difference*. Since the individual discriminial processes follow a Normal distribution, the discriminial difference also follows a Normal distribution (Andrich, 1978), with its mode (mean) representing the relative separation between the stimuli, while its dispersion reflects the variability in that relative separation.

Figure 2 shows the distribution of the discriminial difference for the texts depicted in Figure 1. The figure illustrates that the judge perceives Text B as having significantly

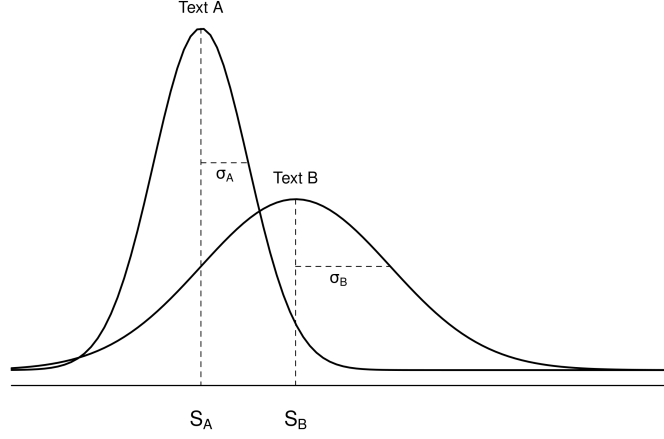


Figure 1: Discriminational processes for two written texts

higher quality than Text A, as indicated by their discriminational difference ($S_B - S_A > 0$) and the shaded gray area under the curve $P(B > A)$. Therefore, the dichotomous outcome of this comparison would likely favor Text B over Text A.

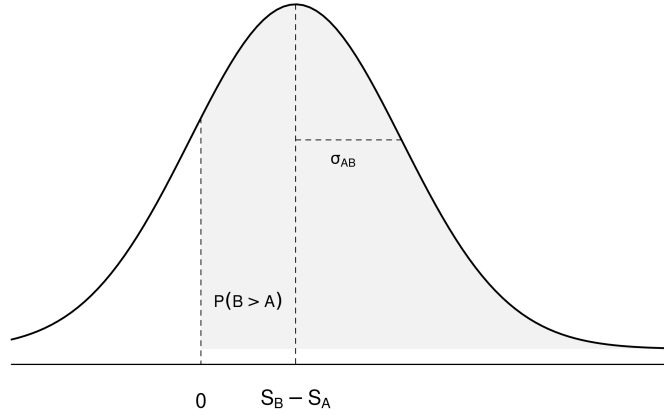


Figure 2: Discriminational difference for two written texts

3. The two critical issues in CJ literature

This section examines the two critical issues in the CJ literature that serve as the primary focus of this study. The first is the overreliance on Thurstone's Case V assumptions in the statistical analysis of CJ data. The second is the apparent disconnect between CJ's approach to trait measurement and hypothesis testing.

Table 1: Thurstone’s cases and assumptions

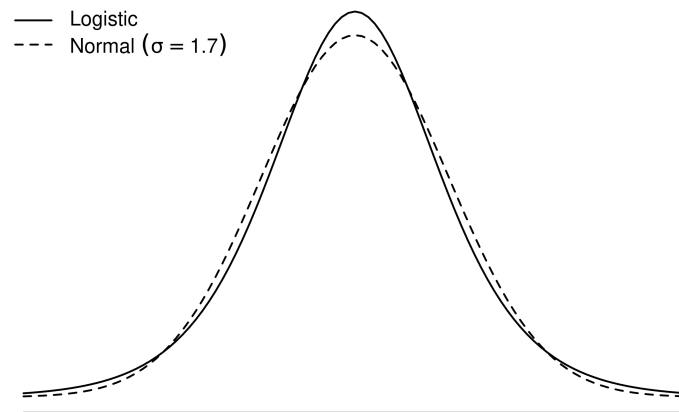
Assumption	Thurstone’s					BTL model
	Case I	Case II	Case III	Case IV	Case V	
Discriminal process (distribution)	Normal	Normal	Normal	Normal	Normal	Logistic
Discriminal dispersion (between stimuli)	Different	Different	Different	Similar	Equal	Equal
Correlation (between stimuli)	Constant	Constant	Zero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

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

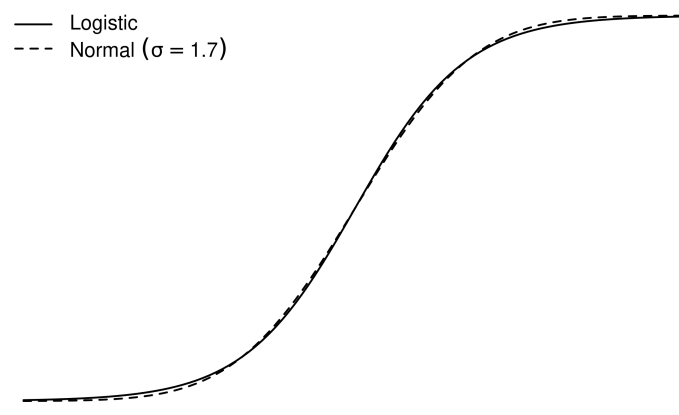
The previous section outlines the general form of Thurstone’s theory, which applies to a CJ design where a single judge evaluates multiple stimuli. For the practical application of the theory, Thurstone developed four additional cases derived from this general form, where each successive case incorporates additional simplifying assumptions. Case I represents the general form of the theory. Case II extends this by allowing multiple judges to make comparisons rather than restricting the comparisons to a single judge. Case III introduces the assumption of zero correlation between stimuli. Case IV builds on this by assuming that the stimuli have similar dispersions. Finally, Case V replaces this assumption with the condition that the stimuli have equal discriminative dispersions. Table 1 summarizes these cases and their assumptions. For a detailed discussion of this progression, refer to [Thurstone \(1927a\)](#) and [Bramley \(2008, pp. 248-253\)](#).

Despite its reliance on the largest number of simplifying assumptions ([Bramley, 2008, pp. 253](#); [Kelly et al., 2022, pp. 677](#)), Case V remains the most widely used case in the CJ literature. This popularity stems mainly from its simplified statistical representation in the Bradley-Terry-Luce (BTL) model ([Bradley and Terry, 1952](#); [Luce, 1959](#)). The BTL model mirrors the assumptions of Case V, with one key difference: while Case V assumes a Normal distribution for the discriminative processes of the stimuli, the BTL model uses the more mathematically tractable Logistic distribution ([Andrich, 1978](#); [Bramley, 2008, pp. 254](#)) (see Table 1). This substitution has little impact on the model’s estimation or interpretation, as the Normal and Logistic distributions share similar statistical properties, differing only by a scaling factor of approximately 1.7 ([van der Linden, 2017a, pp. 16](#)) (see Figure 3).

Nevertheless, Thurstone originally developed Case V to provide a “rather coarse scaling” of traits ([Thurstone, 1927a, pp. 269](#)), prioritizing statistical simplicity over precision in trait measurement ([Kelly et al., 2022, pp. 677](#)). Moreover, he warned against its untested application, stating that its use “should not be made without (an) experimental test” ([Thurstone, 1927a, pp. 270](#)), acknowledging that some assumptions might be problematic when complex traits or heterogeneous stimuli are involved ([Thurstone, 1927b, pp. 376](#)). Consequently, given that modern CJ applications frequently involve such traits and stimuli, two main assumptions of Case V and, by extension, of the BTL model may not consistently hold in theory or practice: the assumption of equal dispersion and zero correlation between stimuli.



(a) Probability density



(b) Cumulative probability

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

3.1.1. The assumption of equal dispersions between stimuli

The discriminative dispersions of stimuli are crucial in determining the dichotomous outcomes of pairwise comparisons. Discrepancies in these dispersions shape the distribution of the discriminative difference, directly influencing the comparison outcome. For instance, Figure 4 illustrates how more uncertainty in the trait perception of a text over another broadens the distribution of their discriminative difference. This broadening affects the area under the curve, $P(B > A)$, determining the comparison outcome. Additionally, the figure shows that when there are no discrepancies in the texts' discriminative dispersions, meaning the *texts have equal dispersions* ($\sigma_B - \sigma_A = 0$), the likelihood that the discriminative difference distinctly favors one text over another increases.

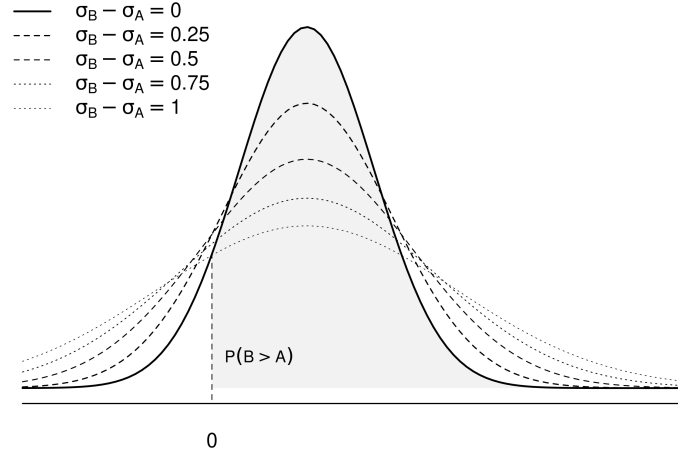


Figure 4: The effect of different dispersions between stimuli on the distribution of the discriminative difference

However, Thurstone contended that this assumption may not hold when researchers assess complex traits or heterogeneous stimuli (Thurstone, 1927b, pp. 376), as these traits and stimuli can introduce judgment discrepancies due to their unique features. Indeed, evidence of such violation may already exist in the CJ literature as misfit statistics. These statistics measure the judgment discrepancies associated with a given stimulus (Pollitt, 2004, pp. 12; Goossens and De Maeyer, 2018, pp. 20). For instance, labeling texts as “misfits” indicates that comparisons involving these texts result in more judgment discrepancies than comparisons involving other texts (Pollitt, 2012a,b; van Daal et al., 2016; Goossens and De Maeyer, 2018). This finding implies that the discriminative differences associated with “misfits” texts usually display a broader dispersion, suggesting that their discriminative processes also exhibit more variation than other texts. Notably, this reasoning also applies to “misfit” judges, whose evaluations reflect substantial deviations from the shared consensus due to the unique characteristics of the stimuli or the judges themselves.

3.1.2. The assumption of zero correlation between stimuli

Similar to the discriminative dispersions, the correlation between the discriminative processes is also crucial in determining the dichotomous outcomes of pairwise comparisons. This

correlation shapes the distribution of the discriminial difference, directly influencing the comparison outcome. For instance, Figure 5 illustrates how the stronger dependence of the trait perception in one text over another narrows the distribution of their discriminial difference. This narrowing impacts the area under the curve, $P(B > A)$, determining the comparison outcome. Additionally, the figure shows that when the trait perception in one text is independent of another, meaning the *texts have zero correlation*, the likelihood that the discriminial difference distinctly favors one text over another is lower compared to when the texts are highly correlated.

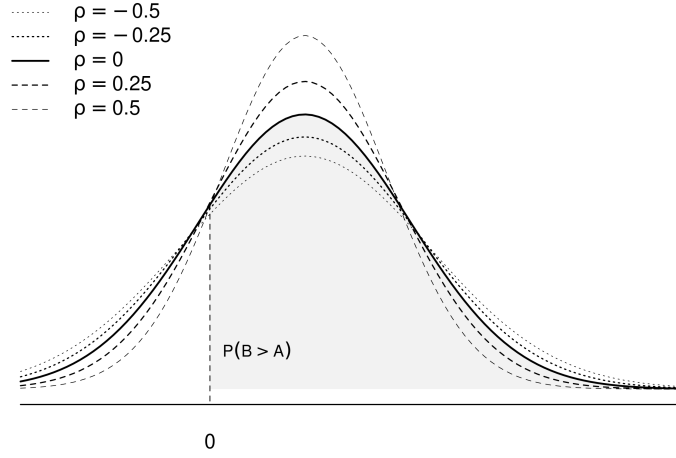


Figure 5: The effect of correlation on the distribution of the discriminial difference

Notably, Thurstone’s Case V and the BTL model assume independent trait perceptions across stimuli. Thurstone attributed this independence to the cancellation of potential judges’ biases, driven by two opposing and equally weighted effects occurring during the pairwise comparisons (Thurstone, 1927a, pp. 268). Andrich (1978) mathematically demonstrated this cancellation using the BTL model under the assumption of discriminial processes with additive biases. However, it is easy to imagine at least two scenarios where the zero correlation assumption almost certainly does not hold: when the pairwise comparison involves multidimensional, complex traits with heterogeneous stimuli and when an additional hierarchical structure is relevant to the stimuli.

In the first scenario, the intricate aspects of multidimensional, complex traits may introduce dependencies between the stimuli due to certain judges’ biases that resist cancellation. Research on text quality suggests that when judges evaluate these traits, they often rely on various intricate characteristics of the stimuli to form their judgments (van Daal et al., 2016; Lesterhuis, 2018b; Chambers and Cunningham, 2022). These additional relevant characteristics, which are unlikely to be equally weighted or opposing, can unevenly influence judges’ perceptions, creating biases in their judgments and, ultimately, introducing dependencies between stimuli (van der Linden, 2017b, pp. 346). For example, this could occur when a judge assessing the argumentative quality of a text places more weight on grammatical accuracy than other judges, ultimately favoring texts with fewer errors but weaker arguments. While direct evidence for this specific scenario

is lacking, studies such as [Pollitt and Elliott \(2003\)](#) demonstrate the presence of such biases, supporting the idea that the factors influencing pairwise comparisons may not always cancel out.

In the second scenario, the shared context or inherent connections created by additional hierarchical structures may further introduce dependencies between stimuli, a statistical phenomenon commonly known as clustering ([Everitt and Skrondal, 2010](#)). Although the CJ literature acknowledges the presence of such hierarchical structures, the statistical handling of this extra source of dependency between stimuli has been inadequate. For example, when CJ data includes multiple samples of stimuli from the same individuals, researchers often rely on (average) estimated BTL scores to conduct subsequent analyses and tests at the individual hierarchical level ([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](#)). This approach, however, has the significant limitation of ignoring the uncertainty associated with the BTL scores, which generates additional statistical and measurement issues, as discussed in section Section 3.2.

3.2. The disconnect between trait measurement and hypothesis testing

Building on the previous section, it is evident that, besides its inconveniences, the BTL model commonly functions as the trait’s measurement model in CJ assessments. 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 texts (the latent variable) ([Laming, 2004](#); [Pollitt, 2012b](#); [Whitehouse, 2012](#); [van Daal et al., 2016](#); [Lesterhuis, 2018a](#); [Coertjens et al., 2017](#); [Goossens and De Maeyer, 2018](#); [Bouwer et al., 2023](#)).

Researchers then typically use these 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., 2016](#); [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 et al., 2020](#); [Bouwer et al., 2023](#); [van Daal et al., 2017](#); [Jones et al., 2019](#); [Gijzen et al., 2021](#)).

However, the statistical literature advises caution when using estimated scores for additional analyses and tests. A key consideration is that BTL scores are parameter estimates that inherently carry uncertainty. Ignoring this uncertainty can bias the analysis and reduce the precision of hypothesis tests. Notably, the direction and magnitude of such biases are often unpredictable. Results may be attenuated, exaggerated, or remain unaffected depending on the degree of uncertainty in the scores and the actual effects being tested ([Kline, 2023](#), pp. 25; [Hoyle, 2023](#), pp. 137). Finally, the reduced precision in hypothesis tests diminishes their statistical power, increasing the likelihood of committing type-I or type-II errors ([McElreath, 2020](#)).

4. An updated theoretical and statistical model for CJ

4.1. The theoretical model

4.2. From theory to statistics

5. Discussion

5.1. Findings

5.2. Limitations and further research

6. 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: The authors have no relevant non-financial interest to disclose.

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/.

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.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 and 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.

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. 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. 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://assets.publishing.service.gov.uk/media/5a80d75940f0b62305b8d734/2007-comparability-exam-standards-i-chapter7.pdf>.
- 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).
- 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](https://doi.org/10.3389/feduc.2022.802392).
- 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. doi:[10.3389/feduc.2021.788202](https://doi.org/10.3389/feduc.2021.788202).
- Everitt, B., Skrondal, A., 2010. *The Cambridge Dictionary of Statistics*. Cambridge University Press.
- 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. 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., 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1007/s10798-011-9190-4).
- 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](https://doi.org/10.2304/plat.2003.3.2.89).
- Lesterhuis, M., 2018a. The validity of comparative judgement for assessing text quality: An assessor's perspective. Ph.D. thesis. University of Antwerp. URL: <https://hdl.handle.net/10067/1548280151162165141>.
- Lesterhuis, M., 2018b. 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 Psychological 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., 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](#).
- 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](#).
- 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 20 december 2024.
- 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. doi:[10.3389/feduc.2017.00044](#).
- van der Linden, W. (Ed.), 2017a. *Handbook of Item Response Theory: Models*. volume 1 of *Statistics in the Social and Behavioral Sciences Series*. CRC Press.
- van der Linden, W. (Ed.), 2017b. *Handbook of Item Response Theory: Statistical Tools*. volume 2 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. 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.