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)

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 discriminal process of each stimulus and their discriminal 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 discriminal process for each stimulus follows a Normal distribution, where its mode (mean), called the *modal discriminal 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 discriminal 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 discriminal 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 discriminal dispersion ($\sigma_B > \sigma_A$).

However, since the discriminal processes of the stimuli are not directly observable, the theory introduces the *law of comparative judgment*. This law states that in pairwise comparisons, a judge perceives the stimulus positioned further along the trait continuum as having a higher level of that trait. This law 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 discriminal processes of the stimuli, referred to as the *discriminal difference*. Since the individual discriminal processes follow a Normal distribution, the discriminal difference also follows a Normal distribution (Andrich, 1978). In this distribution, the mode (mean) represents the relative separation between the stimuli and the dispersion captures the variability of that relative separation.

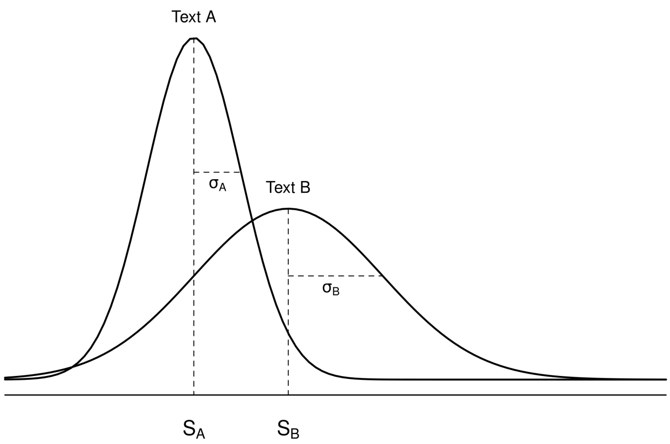


Figure 1: Discriminal processes for two written texts

Figure 2 illustrates the distribution of the discriminal difference for the texts presented in Figure 1. The figure indicates that the judge perceives Text B as having significantly higher quality than Text A. This conclusion is evident from the positive difference in the modal discriminal processes of the texts ($S_B - S_A > 0$) and the area under the curve where the discriminal difference distinctly favors Text B over Text A, denoted as $P(B > A)$ (shaded gray area). As a result, the dichotomous outcome of this comparison is more likely to favor Text B over Text A.

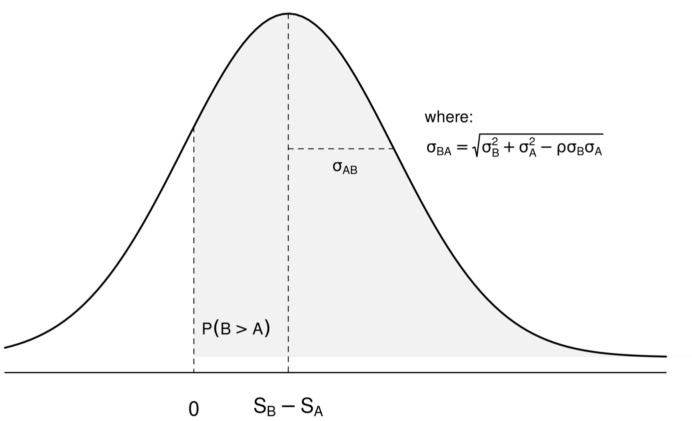


Figure 2: Discriminal difference for two written texts

Table 1: Thurstones cases and asumptions

Assumption	**Thurstone's** **Case I**	**Case II**	**Case III**	**Case IV**	**Case V**	**BTL 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
How many judges compare?	Single	Multiple	Multiple	Multiple	Multiple	Multiple

3. The two critical issues in CJ literature

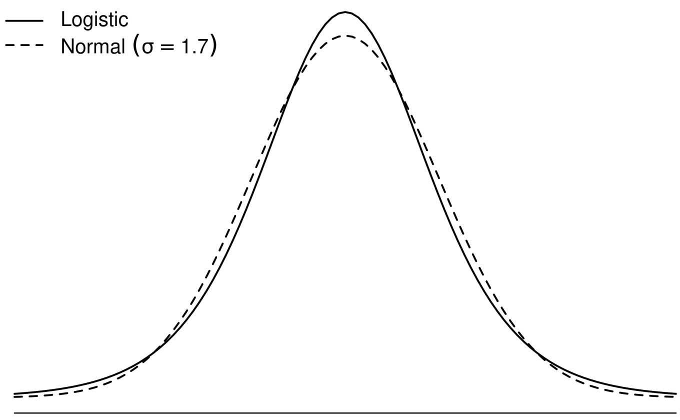
This section examines the two critical issues in the CJ literature that serve as the primary focus of this study. The first is the overreliance on Thurstone's Case V assumptions in the statistical analysis of CJ data. The second is the apparent disconnect between CJ's approach to trait measurement and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

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

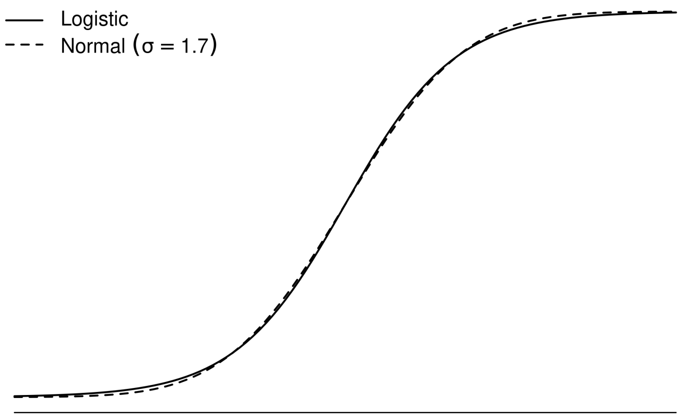
The previous section outlines the general form of Thurstone's theory, which applies to a CJ design where a single judge evaluates multiple stimuli. For the practical application of the theory, Thurstone developed four additional cases derived from this general form, where each successive case incorporates 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 the stimuli have similar dispersions. Finally, Case V replaces this assumption with the condition that the stimuli have equal discriminal 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 stimuli's discriminal processes, 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). He explicitly warned against



(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).

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.

3.1.1. The assumption of equal dispersions between stimuli

The discriminal dispersions of stimuli are crucial in determining the dichotomous outcomes of pairwise comparisons. Discrepancies in these dispersions shape the distribution of the discriminal difference, directly influencing the comparison outcome. Figure 4 illustrates how more uncertainty in the trait perception of one text relative to another broadens the distribution of their discriminal difference. This broadening affects the area under the curve where the discriminal difference distinctly favors one text over the other, $P(B > A)$. Additionally, the figure shows that when the discriminal dispersions of the texts are equal ($\sigma_B - \sigma_A = 0$), the discriminal difference is more likely to favor one text over the other than when their dispersions differ (shaded gray area).

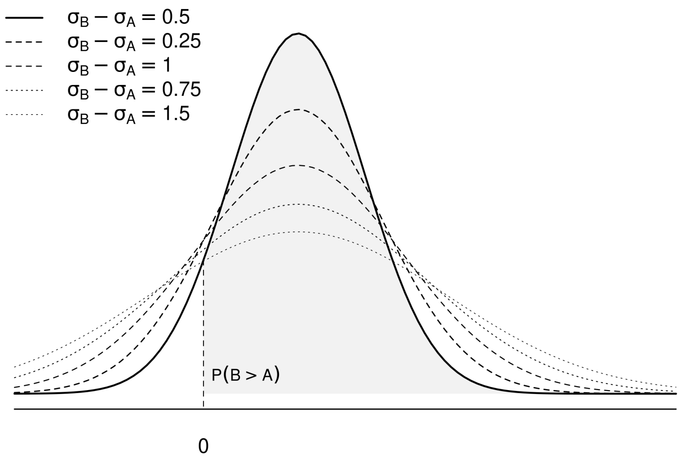


Figure 4: The effect of different dispersions between stimuli on the distribution of the discriminal difference

However, Thurstone contended that the assumption of equal dispersions 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 (van Daal et al., 2016; Lesterhuis, 2018b; Chambers and Cunningham, 2022). 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 others (Pollitt, 2012a,b; van Daal et al., 2016; Goossens and De Maeyer, 2018). This finding implies that discriminal differences associated with "misfits" texts usually display a broader dispersion, suggesting that their discriminal 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. Moreover, the presence of these "misfit" judges and their deviations can introduce additional statistical and measurement issues, as discussed in Section 3.1.2.

Therefore, assuming equal dispersions between stimuli, despite its violation, can cause Case V (and the BTL model) to inflate the reliability of the comparison outcome. Moreover, ignoring the difference in dispersions can lead to the neglect of critical differences in the reliability of the trait across stimuli, resulting in inaccurate conclusions about the trait's estimates (McElreath, 2020, pp. 370). Furthermore, if researchers acknowledge that misfit statistics help identify these critical differences, the usual practice in the CJ literature of excluding stimuli based on these statistics (Pollitt, 2012b; van Daal et al., 2017; Goossens and De Maeyer, 2018) risks discarding valuable information and introducing bias into the trait's estimates (Zimmerman, 1994; McElreath, 2020, chap. 12). The direction and magnitude of these biases are unpredictable, as they depend on the specific stimuli researchers exclude from the analysis. Together, these oversights undermine the reliability of the trait and ultimately compromise its validity (Perron and Gillespie, 2015, pp. 2).

3.1.2. The assumption of zero correlation between stimuli

Similar to the discriminal dispersions, the correlation between discriminal processes plays a crucial role in determining the dichotomous outcomes of pairwise comparisons. This correlation also shapes the distribution of the discriminal difference, directly influencing the comparison outcome. Figure 5 illustrates how the dependence in trait perception between two texts narrows the distribution of their discriminal difference. This narrowing affects the area under the curve where the discriminal difference distinctly favors one text over the other, $P(B > A)$. Moreover, the figure shows that when two texts are independent or uncorrelated ($\rho = 0$), their discriminal difference is less likely to favor one text over the other than when the texts are highly correlated (shaded gray area).

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 discriminal 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

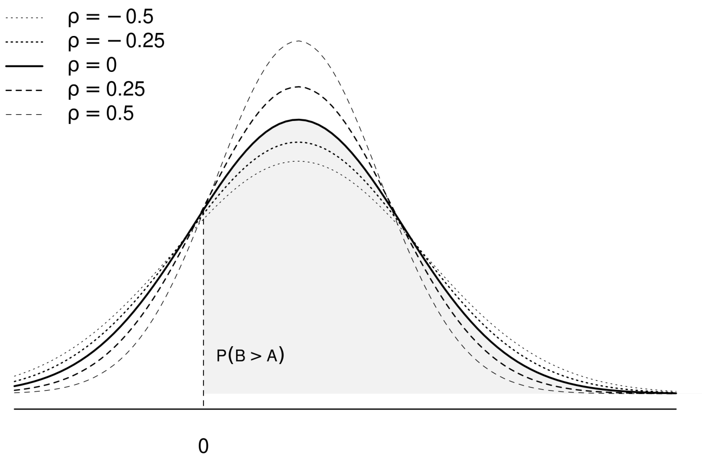


Figure 5: The effect of correlation on the distribution of the discriminal difference

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 its 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; Gijsen et al., 2021). This approach, however, can introduce additional statistical and measurement issues, as discussed in Section 3.2.

In any case, the psychometric and statistical literature emphasizes the need to address factors that create dependencies between stimuli, as failing to do so can affect the reliability of the comparison outcomes and lead to inaccurate conclusions about the trait's estimates. For instance, researchers who overlook additional relevant traits, such as judges' biases, can cause dimensional mismatches in the statistical model used for analysis. This mismatch can artificially inflate the reliability of the trait (Hoyle, 2023, pp. 341) or, even worse, introduce bias into the trait's estimates (Ackerman, 1989). Furthermore, as discussed in Section 3.1.1, researchers who exclude judges based on misfit statistics can risk discarding valuable information, further biasing the trait's estimates (Zimmerman, 1994; McElreath, 2020, chap. 12). Lastly, researchers who fail to account for hierarchical (grouping) structures can reduce the precision of model parameter estimates, which may amplify the overestimation of reliability (Hoyle, 2023, pp. 482). Together, these issues undermine the reliability of the trait and compromise the validity of its estimates (Perron and Gillespie, 2015, pp. 2).

3.2. The disconnect between trait measurement and hypothesis testing

Building on the previous section, it is clear that, despite its limitations, the BTL model is commonly used as the 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 the 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; Gijsen 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://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 and 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.

References

Ackerman, T., 1989. Unidimensional irt calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement 13, 113–127. doi:10.1177/014662168901300201.

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. 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. 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://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.

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.

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.

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.

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., 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.

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.

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.

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.

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., 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 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.

Perron, B., Gillespie, D., 2015. Reliability and Measurement Error, in: Key Concepts in Measurement. Oxford University Press. Pocket guides to social work research methods. chapter 4. doi: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.

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.

Zimmerman, D., 1994. A note on the influence of outliers on parametric and nonparametric tests. The Journal of General Psychology 121, 391–401. doi:10.1080/00221309.1994.9921213.