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, 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,

*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)

2012; Verhavert et al., 2022; Mikhailiuk et al., 2021). Meanwhile, research on 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).

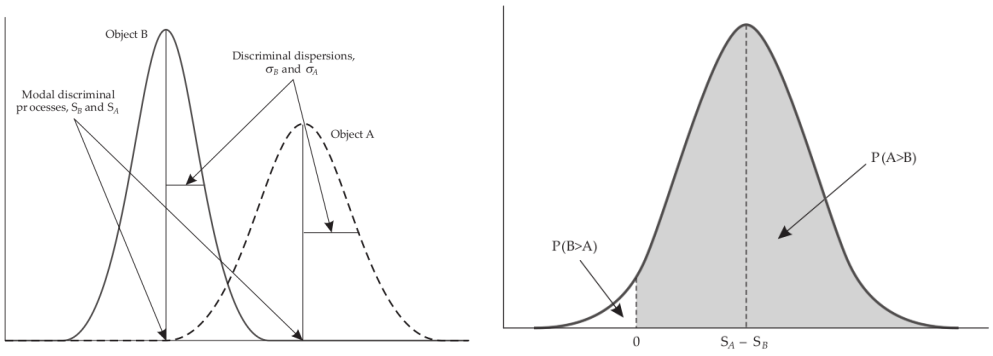
Nevertheless, despite the increasing number of CJ studies, unsystematic and fragmented research approaches have left several critical issues unaddressed. The present study primarily focuses on three: the over-reliance on the assumptions of Thurstone's Case V in the statistical analysis of CJ data, the apparent disconnect between CJ's trait measurement and hypothesis testing, and the unclear role of the diverse assessment design features on CJ's reliability and validity. The following sections begin with a brief overview of Thurstone's theory and a detailed discussion 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 all three concerns simultaneously.

2. Thurstone's theory

In its most general form, Thurstone's theory (1927a) suggests that two factors determine the dichotomous outcome of pairwise comparisons: the discriminal process of each stimulus and their discriminal difference. The *discriminal process* refers to the psychological effect each stimulus exerts on the judges, or more simply, the underlying perception of the stimulus' trait level. According to the theory, the discriminal process for each stimulus follows a Normal distribution. The mode (mean) of this distribution, referred to as the *modal discriminal process*, represents the stimulus' position on the trait continuum. Meanwhile, the dispersion of the distribution, referred to as the *discriminal dispersion*, reflects the variability in the perceived trait level of the stimulus.

However, since the discriminal process of a single stimulus is not directly observable, the *law of comparative judgment* becomes essential. This law states that in pairwise comparisons, the stimulus positioned further along the trait continuum is perceived as having a higher level of that trait. Thus, the theory assumes the observed dichotomous outcome is determined by the distribution of the difference between the underlying discriminal processes of the stimuli, referred to as the *discriminal difference*. This indicates that the outcome depends on the relative distance between stimuli, rather than their absolute positions on the trait continuum.

These concepts are more easily understood through an example. For instance, in the context of evaluating text quality, Figure 1a could depict the underlying discriminal process distributions for two written texts, highlighting differences in their discriminal dispersions and modal discriminal processes along the quality trait continuum. Furthermore, Figure 1b could display the discriminal difference distribution for these texts, showing that text A is perceived to exhibit significantly higher quality than text B, as indicated by the shaded gray area. Consequently, the dichotomous outcome of this comparison would probably favor text A.



(a) Discriminal processes

(b) Discriminal difference

Figure 1: Example distribution of discriminal processes and their discriminal difference for two written texts (stimuli or objects). Extracted from Bramley (2008, pp. 249-251).

Table 1: Thurstones cases and asumptions

	Thurstone's					BTL
Assumption	**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
How many judges compare?	Single	Multiple	Multiple	Multiple	Multiple	Multiple

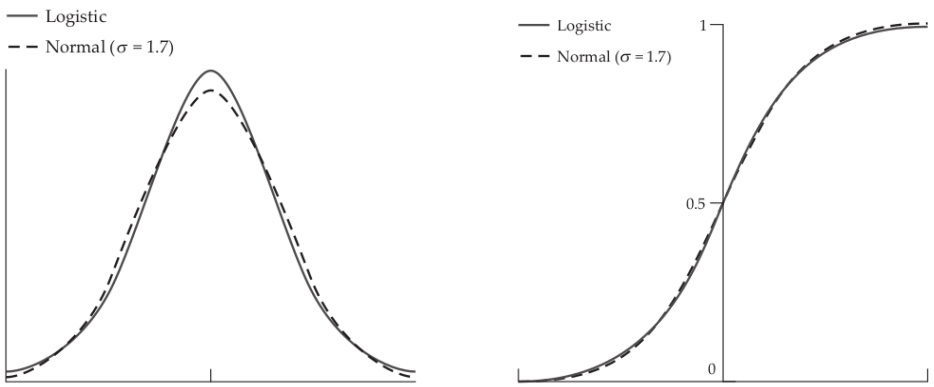
Importantly, the general form of Thurstone's theory primarily addressed pairwise comparisons of stimuli made by a single judge (Thurstone, 1927a, pp. 267). Thus, for practical application, Thurstone introduced five distinct cases derived from this general form, each defined by progressively simplifying assumptions. Table 1 summarizes these cases, highlighting key assumptions such as the distribution of discriminal processes, the similarity of discriminal dispersions across stimuli, the correlation between stimuli, and the number of judges performing the comparisons. For a comprehensive discussion of this progression, refer to Thurstone (1927a) and Bramley (2008, pp. 248-253).

3. Three critical issues in CJ literature

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

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 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 the assumptions of Case V, with one key difference: while Case V assumes a Normal distribution for the discriminal processes of 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, 2017, pp. 16) (see Figure 2).



(a) Probability density

(b) Cummulative probability

Figure 2: Probability density and cumulative probability of the logistic and Normal distributions. Extracted from Bramley (2008, pp. 254-255).

However, Case V was originally developed 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). As a result, its assumptions may not be suitable for applications beyond the psycho-physical contexts for which it was created. Thurstone himself cautioned that its use "should not be made without (an) experimental test" (Thurstone, 1927a, pp. 270), acknowledging that some assumptions could prove problematic in the presence of complex traits or heterogeneous stimuli, such as handwriting or English compositions (Thurstone, 1927b, pp. 374). Consequently, given that modern CJ applications frequently involve these types of traits and stimuli, two main assumptions of Case V may not consistently hold in theory or practice: the zero correlation and equal dispersion between stimuli.

The assumption of *zero correlation between stimuli* can be better understood through an example. For instance, when using pairwise comparisons to evaluate text quality, the assumption implies that a judge's perception of a trait in one text does not influence his perception of the same trait in another text. 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). This cancellation was mathematically demonstrated by Andrich (1978), using the BTL model under the assumption of additive biases in the discriminal processes. 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 and heterogeneous stimuli may introduce dependencies between stimuli. Research on text quality suggests that when judges evaluate these traits, they often rely on various intricate aspects of the stimuli to form their judgments (van Daal et al., 2016; Lesterhuis, 2018b; Chambers and Cunningham, 2022). In this context, it is not inconceivable that these aspects, being neither equally weighted nor opposing, may unevenly influence judges' perceptions, resulting in biases that resist cancellation. For example, this might occur when a judge assessing the argumentative quality of a text places disproportionate emphasis on grammatical accuracy, 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 judges' biases, supporting the idea that the different factors influencing pairwise comparisons may not always cancel out.

In the second scenario, the shared context or inherent connections created by the additional hierarchical structure may introduce dependencies between stimuli, a statistical phenomenon commonly known as clustering (Everitt and Skrondal, 2010). Nevertheless, despite recognizing such hierarchical structures in CJ data, the statistical handling of this extra source of dependency in the CJ literature has been inadequate. For instance, in cases where the CJ data included multiple samples of stimuli from the same individuals, researchers have often relied on (averaged) 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, has the significant limitation of ignoring the uncertainty associated with the scores (refer to section Section 3.2 for a detailed discussion of this issue).

In contrast, the assumption of *equal dispersion between stimuli* suggests that the variability in the perceived trait level of the stimuli is the same across stimuli. While Thurstone acknowledged that this assumption may be violated when "dealing with less conspicuous attributes or with less homogeneous stimuli" (Thurstone, 1927b, pp. 374), no study explicitly proposes that this assumption could also be violated due to the presence of an additional hierarchical (grouping) structure relevant to the texts. One such scenario might arise, for example, when comparing texts produced by university and secondary school students. In this case, university students may consistently (or more precisely) produce higher-quality texts, while secondary school students, who exhibit a broader range of writing abilities, would show greater variability in the quality of their texts. Although this example is somewhat contrived, it effectively illustrates how assuming equal dispersions across texts can overlook meaningful differences in the reliability of text quality across groups or individuals.

"however" paragraph

"to mitigate" paragraph

3.2. The disconnect between trait measurement and hypothesis testing

Building on the previous section, it is evident that the BTL model commonly functions as the trait's measurement model in CJ experiments (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 quality level of texts (the latent variable) (Laming, 2004; Pollitt, 2012; 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 the estimated BTL scores, or their transformations, to conduct additional analyses and tests, or to make decisions regarding the exclusion of certain data in these analyses and tests. The literature shows that these scores have been employed to calculate correlations with other assessment methods (Goossens and De Maeyer, 2018; Bouwer et al., 2023) or to test hypotheses related to the underlying traits 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). Additionally, the BTL scores have been used to detect biases in judges' ratings (Pollitt and Elliott, 2003; Pollitt, 2012), as well as to identify "misfit" judges and stimuli (Pollitt, 2012; van Daal et al., 2017; Goossens and De Maeyer, 2018), with considerations for their possible exclusion.

However, the statistical literature advises caution when using estimated scores for additional analyses and tests, as well as when eliminating data through ad hoc univariate procedures. 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). Moreover, excluding data using ad hoc univariate procedures can compound these issues by discarding potentially valuable information, further exacerbating the bias (Zimmerman, 1994; McElreath, 2020). Finally, the reduced precision in hypothesis tests diminishes their statistical power, 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: 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.

7. 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. 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.

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. 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 as-

sessor'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.

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 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.), 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. 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.