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

Jose Manuel Rivera Espejo

Tine van van Daal

Sven De De Maeyer

Steven Gillis

2024-11-25

Abstract

(to do)

# Introduction

In *comparative judgment* (CJ) studies, judges assess a specific trait or attribute across various stimuli by performing pairwise comparisons (Thurstone 1927b, 1927a). 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, Béguin, and Sijtsma 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, Furlong, and Bouwer 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, Kloots, and Gillis 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.

# 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.

Nevertheless, the theory posits that because 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. This emphasize that the outcome depends on the relative distance between stimuli, rather than their absolute positions on the trait continuum. 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*.

These concepts are more easily understood through an example. For instance, in the context of evaluating text quality, [Figure 1 (a)](#fig-discriminal_process) depicts the underlying discriminal process distributions for two written texts. The figure highlights differences in the texts’ positions along the quality trait continuum, represented by their modal discriminal processes, as well as differences in the variability of their perceived trait levels, represented by their discriminal dispersions. Furthermore, [Figure 1 (b)](#fig-discriminal_difference) displays the discriminal difference distribution for these texts. This figure shows 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 likely 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, 249–51).

Importantly, the general form of Thurstone’s theory primarily addressed pairwise comparisons of stimuli made by a single judge (Thurstone 1927a, 267). Thus, to facilitate its practical application, Thurstone developed five distinct cases derived from this general form. Each case progressively introduces additional simplifying assumptions related to different features of the theory, such as the distribution of discriminal processes, the uniformity of discriminal dispersions across stimuli, the correlation between stimuli, and the number of judges performing the comparisons. A summary of these cases is provided in [Table 1](#tbl-thurstone_cases). For a detailed examination of this progression, refer to Thurstone (1927a) and Bramley (2008, 248–53).

|  |
| --- |
| Table 1: Thurstones cases and asumptions |

# Three critical issues in CJ literature

## The Case V and the statistical analysis of CJ data

Despite its reliance on the largest number of simplifying assumptions (Bramley 2008, 253; Kelly, Richardson, and Isaacs 2022, 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 the stimuli, the BTL model uses the more mathematically tractable Logistic distribution (Andrich 1978; Bramley 2008, 254) (see [Table 1](#tbl-thurstone_cases)). 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 (van der Linden 2017a, 1:16) (see [Figure 2](#fig-logistic_vs_normal)).

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Probability density | |  |

|  |  |
| --- | --- |
| |  | | --- | | (b) Cummulative probability | |

Figure 2: Probability density and cumulative probability of the logistic and Normal distributions. Extracted from Bramley (2008, 254–55).

However, Case V was originally developed to provide a “rather coarse scaling” of traits (Thurstone 1927a, 269), prioritizing statistical simplicity over precision in trait measurement (Kelly, Richardson, and Isaacs 2022, 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, 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, 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, 268). This cancellation was mathematically demonstrated by Andrich (1978), 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 heterogeneous stimuli due to certain judges’ biases that resist cancellation. Research on text quality indicates that when judges evaluate such 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 could 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 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, Kloots, and Gillis 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 BTL scores (refer to section [Section 3.2](#sec-theory-issue2) 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 stimulus is the same across all stimuli. While Thurstone acknowledged that this assumption may be violated when “dealing with less conspicuous attributes or with less homogeneous stimuli” (Thurstone 1927b, 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.

But the psychometric and statistical literature strongly advises against ignoring additional traits that could confound the measurement of the trait of interest, overlooking clustering (grouping) structures, or mis-specifying the trait’s measurement model. First, when additional traits are ignored in favor of a unidimensional trait structure, such as when judges’ biases are disregarded, the association between stimuli does not vanish (van der Linden 2017b, 2:346). This mismatch can result in a potential overestimation of precision (reliability) in the measurement process (Hoyle 2023, 340–41), or worse, introduce bias into the trait’s measurement (Ackerman 1989). Second, when relevant hierarchical (or grouping) structures are present in the data, ignoring these structures can further exacerbate the inflation of measurement reliability (citation needed). Lastly, mis-specifying the measurement model by assuming, for example, equal discriminal dispersions across stimuli can overlook meaningful differences in the reliability of the trait. Taken together, these mis-specifications and oversights can lead to misleading conclusions about the trait measurement, ultimately casting doubts about the validity of the trait, as validity cannot exist without reliability (Perron and Gillespie 2015, 2).

“to mitigate” paragraph

## 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, Kloots, and Gillis 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, 25; Hoyle 2023, 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, 138) and Item Response Theory (IRT) (Fox 2010, chap. 6; van der Linden 2017a, vol. 1, 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.

## The diverse assessment design features and their role on reliability and validity

# An updated theoretical and statistical model for CJ

## The theoretical model

## From theory to statistics

# Discussion

## Findings

## Limitations and further research

# 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.

# Appendix

# References

Ackerman, T. 1989. “Unidimensional IRT Calibration of Compensatory and Noncompensatory Multidimensional Items.” *Applied Psychological Measurement* 13 (2): 113–27. <https://doi.org/10.1177/014662168901300201>.

Andrich, D. 1978. “Relationships Between the Thurstone and Rasch Approaches to Item Scaling.” *Applied Psychological Measurement* 2 (3): 451–62. <https://doi.org/10.1177/014662167800200319>.

Bartholomew, S., L. Nadelson, W. Goodridge, and E. Reeve. 2018. “Adaptive Comparative Judgment as a Tool for Assessing Open-Ended Design Problems and Model Eliciting Activities.” *Educational Assessment* 23 (2): 85–101. <https://doi.org/10.1080/10627197.2018.1444986>.

Bartholomew, S., and P. Williams. 2020. “STEM Skill Assessment: An Application of Adaptive Comparative Judgment.” In *Integrated Approaches to STEM Education. Advances in STEM Education*, edited by J. Anderson and Y. Li, 331–49. Springer. <https://doi.org/10.1007/978-3-030-52229-2_18>.

Boonen, N., H. Kloots, and S. Gillis. 2020. “Rating the Overall Speech Quality of Hearing-Impaired Children by Means of Comparative Judgements.” *Journal of Communication Disorders* 83: 1675–87. <https://doi.org/10.1016/j.jcomdis.2019.105969>.

Bouwer, R., M. Lesterhuis, F. De Smedt, H. Van Keer, and S. De Maeyer. 2023. “Comparative Approaches to the Assessment of Writing: Reliability and Validity of Benchmark Rating and Comparative Judgement.” *Journal of Writing Research* 15 (3): 497–518. <https://doi.org/10.17239/jowr-2024.15.03.03>.

Bradley, R., and M. Terry. 1952. “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.” *Biometrika* 39 (3-4): 324–45. <https://doi.org/10.2307/2334029>.

Bramley, T. 2008. “Paired Comparison Methods.” In *Techniques for Monitoring the Comparability of Examination Standards*, edited by P. Newton, J. Baird, H. Goldsteing, H. Patrick, and P. Tymms, 246--300. GOV.UK. <https://assets.publishing.service.gov.uk/media/5a80d75940f0b62305b8d734/2007-comparability-exam-standards-i-chapter7.pdf>.

Bramley, T., and S. Vitello. 2019. “The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement.” *Assessment in Education: Principles, Policy and Practice* 71 (9): 1–25. <https://doi.org/10.1080/0969594X.2017.1418734>.

Chambers, L., and E. Cunningham. 2022. “Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?” *Frontiers in Education*. <https://doi.org/10.3389/feduc.2022.802392>.

Coertjens, L., M Lesterhuis, S. Verhavert, R. Van Gasse, and S. De Maeyer. 2017. “Teksten Beoordelen Met Criterialijsten of via Paarsgewijze Vergelijking: Een Afweging van Betrouwbaarheid En Tijdsinvestering.” *Pedagogische Studien* 94: 283–303. <https://repository.uantwerpen.be/docman/irua/e71ea9/147930.pdf>.

Crompvoets, E., A. Béguin, and K. Sijtsma. 2022. “On the Bias and Stability of the Results of Comparative Judgment.” *Frontiers in Education* 6. <https://doi.org/10.3389/feduc.2021.788202>.

Everitt, B., and A. Skrondal. 2010. *The Cambridge Dictionary of Statistics*. Cambridge University Press.

Fox, J. P. 2010. *Bayesian Item Response Modeling, Theory and Applications*. Statistics for Social and Behavioral Sciences. Springer.

Gijsen, M., T. van Daal, Marije Lesterhuis, David Gijbels, and Sven De Maeyer. 2021. “The Complexity of Comparative Judgments in Assessing Argumentative Writing: An Eye Tracking Study.” *Frontiers in Education* 5. <https://doi.org/10.3389/feduc.2020.582800>.

Goossens, M., and S. De Maeyer. 2018. “How to Obtain Efficient High Reliabilities in Assessing Texts: Rubrics Vs Comparative Judgement.” In *Technology Enhanced Assessment*, edited by E. Ras and A. Guerrero Roldán, 13–25. Springer International Publishing. <https://doi.org/10.1007/978-3-319-97807-9_2>.

Hoyle, R. (eds.). 2023. *Handbook of Structural Equation Modeling*. Guilford Press.

Jones, I., M. Bisson, C. Gilmore, and M. Inglis. 2019. “Measuring Conceptual Understanding in Randomised Controlled Trials: Can Comparative Judgement Help?” *British Educational Research Journal* 45 (3): 662–80. <https://doi.org/10.1002/berj.3519>.

Jones, I., and M. Inglis. 2015. “The Problem of Assessing Problem Solving: Can Comparative Judgement Help?” *Educational Studies in Mathematics* 89 (3): 337–55. <https://doi.org/10.1007/s10649-015-9607-1>.

Kelly, K., M. Richardson, and T. Isaacs. 2022. “Critiquing the Rationales for Using Comparative Judgement: A Call for Clarity.” *Assessment in Education: Principles, Policy & Practice* 29 (6): 674–88. <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: 135–55. <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 (2): 89–96. <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.” PhD thesis, University of Antwerp. <https://hdl.handle.net/10067/1548280151162165141>.

———. 2018b. “When Teachers Compare Argumentative Texts: Decisions Informed by Multiple Complex Aspects of Text Quality.” *L1-Educational Studies in Language and Literature* 18 (1): 1–22. <https://doi.org/10.17239/L1ESLL-2018.18.01.02>.

Luce, R. 1959. “On the Possible Psychophysical Laws.” *The Psychologcal Review* 66 (2): 482–99. <https://doi.org/10.1037/h0043178>.

Marshall, N., K Shaw, J. Hunter, and I. Jones. 2020. “Assessment by Comparative Judgement: An Application to Secondary Statistics and English in New Zealand.” *New Zealand Journal of Educational Studies* 55: 49–71. <https://doi.org/10.1007/s40841-020-00163-3>.

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

Mikhailiuk, A., C. Wilmot, M. Perez-Ortiz, D. Yue, and R. Mantiuk. 2021. “Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization.” In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2559–66. <https://doi.org/10.1109/ICPR48806.2021.9412676>.

Perron, B., and D. Gillespie. 2015. “Reliability and Measurement Error.” In *Key Concepts in Measurement*. Pocket Guides to Social Work Research Methods. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199855483.003.0004>.

Pollitt, A. 2012. “The Method of Adaptive Comparative Judgement.” *Assessment in Education: Principles, Policy and Practice* 19 (3): 281--300. <https://doi.org/10.1080/0969594X.2012.665354>.

Pollitt, A., and G. Elliott. 2003. “Finding a Proper Role for Human Judgement in the Examination System.” University of Cambridge Local Examinations Syndicate. <https://www.cambridgeassessment.org.uk/Images/109707-monitoring-and-investigating-comparability-a-proper-role-for-human-judgement.pdf>.

Thurstone, L. 1927a. “A Law of Comparative Judgment.” *Psychological Review* 34 (4): 482–99. <https://doi.org/10.1037/h0070288>.

———. 1927b. “Psychophysical Analysis.” *American Journal of Psychology*, no. 38: 368–89. <https://brocku.ca/MeadProject/Thurstone/Thurstone_1927g.html>.

van Daal, T., M. Lesterhuis, L. Coertjens, V. Donche, and S. De Maeyer. 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 (1): 59–74. <https://doi.org/10.1080/0969594X.2016.1253542>.

van Daal, T., M. Lesterhuis, L. Coertjens, MT. van de Kamp, V. Donche, and S. De Maeyer. 2017. “The Complexity of Assessing Student Work Using Comparative Judgment: The Moderating Role of Decision Accuracy.” *Frontiers in Education* 2. <https://doi.org/10.3389/feduc.2017.00044>.

van der Linden, W., ed. 2017a. *Handbook of Item Response Theory: Models*. Vol. 1. Statistics in the Social and Behavioral Sciences Series. CRC Press.

———, ed. 2017b. *Handbook of Item Response Theory: Statistical Tools*. Vol. 2. Statistics in the Social and Behavioral Sciences Series. CRC Press.

Verhavert, S., R. Bouwer, V. Donche, and S. De Maeyer. 2019. “A Meta-Analysis on the Reliability of Comparative Judgement.” *Assessment in Education: Principles, Policy and Practice* 26 (5): 541–62. <https://doi.org/10.1080/0969594X.2019.1602027>.

Verhavert, S., A. Furlong, and R. Bouwer. 2022. “The Accuracy and Efficiency of a Reference-Based Adaptive Selection Algorithm for Comparative Judgment.” *Frontiers in Education* 6. <https://doi.org/10.3389/feduc.2021.785919>.

Whitehouse, C. 2012. “Testing the Validity of Judgements about Geography Essays Using the Adaptive Comparative Judgement Method.” Centre for Education Research & Policy. <https://filestore.aqa.org.uk/content/research/CERP_RP_CW_24102012_0.pdf?download=1>.

Zimmerman, D. 1994. “A Note on the Influence of Outliers on Parametric and Nonparametric Tests.” *The Journal of General Psychology* 121 (4): 391–401. <https://doi.org/10.1080/00221309.1994.9921213>.