# Causes and effects in Dichotomous Comparative Judgments: an information-theoretical system with plausible mechanism

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**

Dichotomous Comparative Judgment (DCJ, Pollitt, 2012a, @Pollitt_2012b) requires judges to evaluate the relative manifestation of traits between pairs of stimuli, resulting in a dichotomous outcome indicating which stimulus exhibits the trait more strongly. Research has demonstrated DCJ's effectiveness and reliability in various domains (Pollitt, 2012b; Bartholomew et al., 2018; van Daal et al., 2019; Lesterhuis, 2018; Bartholomew and Williams, 2020; Boonen et al., 2020). Nevertheless, despite the method's widespread use, the literature lacks a transparent depiction of the DCJ system and the plausible mechanisms that generate the DCJ data. Particularly, there is no detailed explanation of how different assessment factors can potentially influence the observed DCJ data. This study aims to fill this gap by applying the framework of causal analysis and Directed Acyclic Graphs [DAG; Pearl (2009)]. Using this framework, the study will construct a scientific model to elucidate the causal assumptions and mechanisms inherent the system. This model will enable researchers to draw inferences about causal relationships from DCJ data. Subsequently, the study will translate this model into a probabilistic statistical model, aiming to derive statistical estimands for different targets of inference. The outcomes of this study will inform the planning of DCJ experiments and hold significance for researchers or analysts involved in education and assessment procedures who implement the DCJ methodology.

*Keywords:* comparative judgement, directed acycilc graph, causal analysis, probabilistic statistics

## 1. Introduction

In contemporary contexts, Thurstone's law of comparative judgment (1927) primarily refers to the method of *dichotomous* comparative judgment (DCJ, Pollitt, 2012a,b).

*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)

In DCJ, a judge assesses the relative manifestation of a *trait* within a pair of stimuli. This assessment results in a dichotomous value indicating which stimulus possesses a higher degree of the trait. After different judges perform multiple rounds of pairwise comparisons, an outcome vector is produced. This vector is modeled using the Bradley-Terry-Luce model (BTL, Bradley and Terry, 1952; Luce, 1959), which creates a score that corresponds with the trait of interest. This score is then used to rank the stimuli from lowest to highest or to evaluate the influence of certain variables on the stimuli's positions in the ranking.

DCJ has proven effective in assessing competencies and traits predominantly within the educational realm, as demonstrated by Pollitt (2012b), Jones (2015), van Daal et al. (2019), Bartholomew et al. (2018), Lesterhuis (2018), Bartholomew and Williams (2020), and Marshall et al. (2020). However, its application transcends education, as exemplified by Boonen et al. (2020). The methodology has also evolved to include multiple, as opposed to pairwise comparisons (Luce, 1959; Plackett, 1975), and to accommodate comparisons with ordinal outcomes (Tutz, 1986; Agresti, 1992). Overall, research suggests that DCJ offers an alternative and efficient approach to measurement and evaluation, characterized by its reliability and validity (Lesterhuis, 2018; van Daal, 2020; Marshall et al., 2020). Nevertheless, despite the method's widespread use, the literature does not offer a clear representation of the plausible mechanisms that generate DCJ data. Particularly, there is no depiction of the complexity and the underlying assumptions of the DCJ system, nor how different assessment factors can potentially influence the observed DCJ outcome.

According to Verhavert et al. (2019) and van Daal (2020), several assessment factors interact and influence the method's outcome. These factors include the number and characteristics of the stimuli, their *proximity* in terms of the assessed trait, the number of comparison per stimulus, and the pairing algorithm used. Furthermore, since the method relies on judges' assessments, the number and characteristics of judges, their *discrimination* abilities, and the number of comparisons per judge also play pivotal roles. Moreover, when the stimuli represent sub-units of higher-levels units, factors such as the number and characteristics of these units, along with their *proximity* in terms of the assessed trait, can significantly influence the outcome. For example, van Daal et al. (2019) assessed the academic writing skills of university students (units) using multiple argumentative essays (sub-units).

Although several studies have examined the individual impact of these factors on the method's reliability (Bramley, 2015; Pollitt, 2012b; Bramley and Vitello, 2019; Verhavert et al., 2019; Crompvoets et al., 2022; van Daal et al., 2017; Gijsen et al., 2021), none, to the best of the authors' knowledge, have provided a transparent depiction of the DCJ system and the mechanisms generating the DCJ outcome. This study aims to fill this gap by representing DCJ within the causal inference framework. Specifically, using the *structural approach* to causal inference (Wright, 1927; Pearl, 2009; Pearl et al., 2016), the study aims to construct a scientific model. This model will elucidate the underlying assumptions of the DCJ system, providing plausible mechanisms for how the DCJ outcome is generated. Next, using a minimal set of assumptions, the study will translate the scientific model into a probabilistic statistical model. This model will produce statistical estimates to draw inferences about plausible causal relationships within the DCJ system.

Ultimately, this research aims to extend the law of comparative judgment initially proposed by Thurstone (1927) and provide a sound probabilistic base for the statistical analysis of DCJ data. Consequently, this research holds significance for researchers and analysts involved in education and assessment procedures who implement or design DCJ experiments.

## 2. Preliminaries

### *2.1. The structural approach to causal inference*

In statistics, *causal inference* refers to the process of identifying the causes of a phenomenon and estimating their effects using data (Shaughnessy et al., 2010; Neal, 2020). Unlike classical statistical modeling, which focuses solely on summarizing data and inferring associations, causal inference provides a coherent mathematical representation for analyzing causes and counterfactuals (Pearl, 2009). Counterfactuals reflect scenarios *contrary to fact*, where alternative *potential* outcomes have not been observed and cannot be observed (Neal, 2020; Counterfactual, 2024). Using counterfactuals, researchers can construct a theory of the world that explains why specific causes have specific effects, and what happens in the absence of those causes. Consequently, counterfactuals represent the highest level of abstraction in causal inference (Pearl and Mackenzie, 2018).

Several approaches to constructing counterfactuals and applying causal inference exist, but two are particularly prominent: the potential outcomes approach and the structural approach. The potential outcomes approach, also known as the Neyman-Rubin causal model, was developed by Neyman et al. (1990) and Rubin (1974). In contrast, the structural approach was pioneered by Wright (1927), and later extended by Pearl (Pearl, 2009; Pearl et al., 2016). Both approaches employ rigorous mathematical notation to characterize causal inference, but they do so in different ways (Neal, 2020). The potential outcomes approach relies on counterfactual notation, whereas the structural approach utilizes the do-operator and Structural Causal Models (SCM, Pearl, 2009; Pearl et al., 2016). Despite these differences, both approaches provide methods for using experimental and observational data to estimate causal effects (Pearl, 2010).

However, the structural approach offers a significant advantage over the potential outcomes approach by enabling the graphical representation of systems through directed acyclic graphs (DAG, Gross et al., 2018; Neal, 2020). These graphical representations provides a transparent depiction of a system's complexity, revealing its underlying assumptions and the plausible mechanisms generating the system's outcome. This capability aligns well with the aims of this study. Consequently, the structural approach will be the primary method used to address the study's research goal.

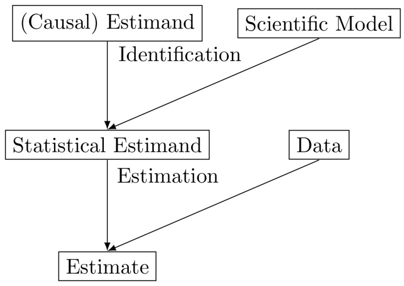


Figure 1: Identification-Estimation flowchart. Extracted from Neal (2020, 32)

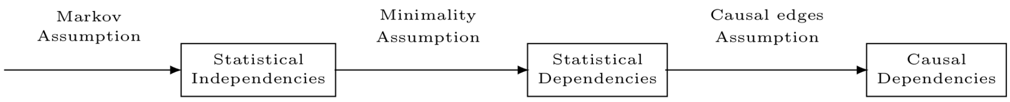


Figure 2: The flow of association and causation in graphs. Extracted from Neal (2020, 31)

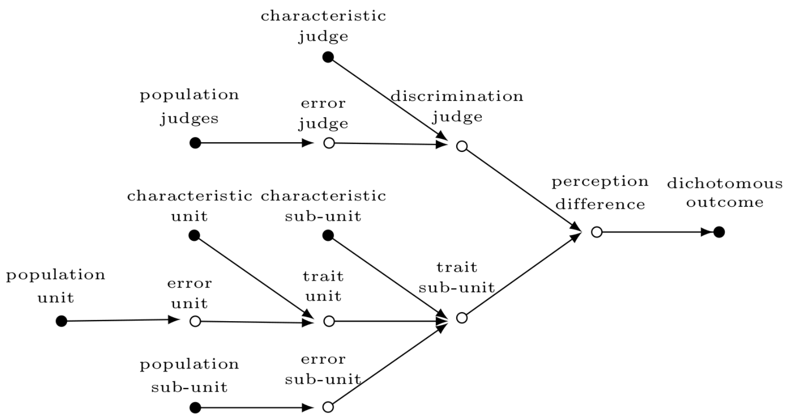


Figure 3: DCJ causal diagram, simplified description

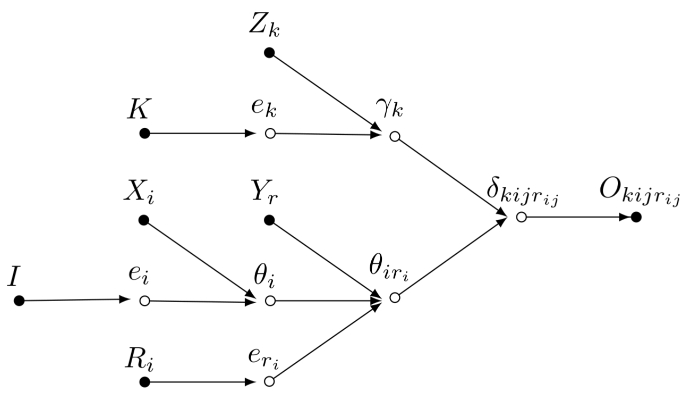


Figure 4: DCJ causal diagram, simplified mathematical description

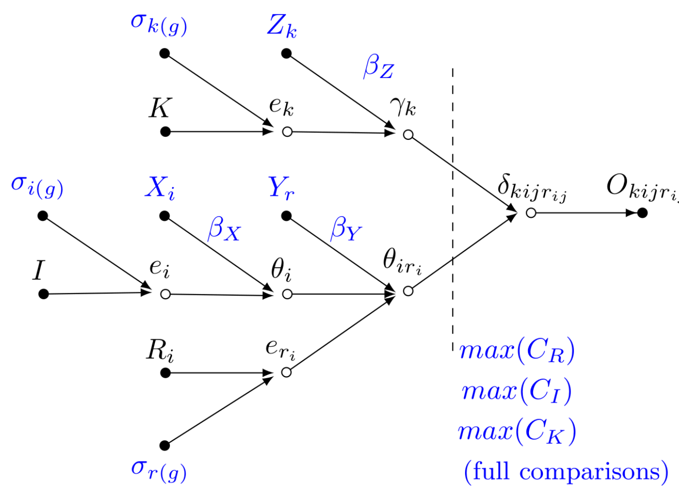


Figure 5: DCJ causal diagram, population mathematical description

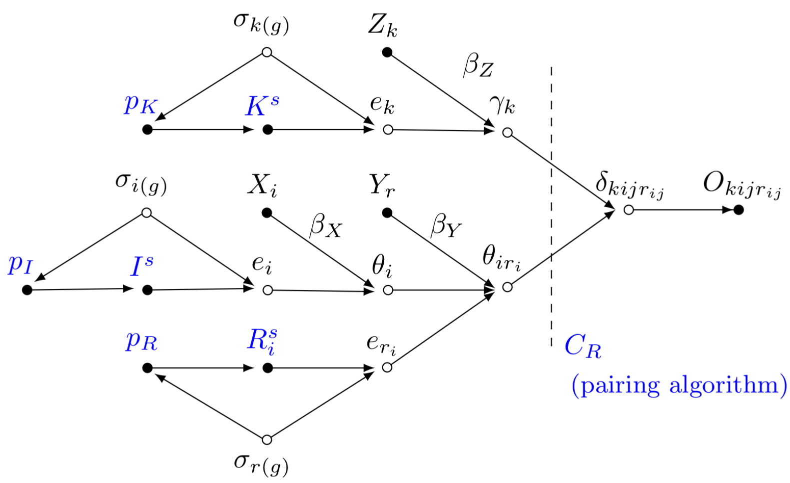


Figure 6: DCJ causal diagram, sample with comparisons mathematical description

### *2.2. Graphs, DAGs and SCMs*

### *2.3. The flow of association and causation*

## 3. Theoretical framework

### *3.1. A scientific model for the DCJ*

### *3.2. Probabilitics assumptions of the scientific model*

$$
\begin{aligned}
O_{kijr_{ij}} &:= f_O(\delta_{kijr_{ij}}) \\
\delta_{kijr_{ij}} &:= f_D(\gamma_k, \theta_{ir_i}) \\
\gamma_k &:= f_G(Z_k, e_k) \\
\theta_{ir_i} &:= f_R(\theta_i, Y_r, e_{r_i}) \\
\theta_i &:= f_T(X_i, e_i) \\
e_k &\perp\!\!\!\perp e_i \\
e_k &\perp\!\!\!\perp e_{r_i} \\
e_i &\perp\!\!\!\perp e_{r_i}
\end{aligned}
\tag{1}
$$

### 3.3. From the scientific to statistical model

$$
\begin{aligned}
O_{kijr_{ij}} &\sim \text{Bernoulli}\left[\, logit^{-1}\left(\delta_{kijr_{ij}}\right)\,\right] \\
\delta_{kijr_{ij}} &= \gamma_k(\theta_{ir_i} - \theta_{jr_j}) \\
\gamma_k &= logit^{-1}\left[\beta_Z Z_k + e_k\right] \\
\theta_{ir_i} &= \theta_i + \beta_Y Y_r + e_{r_i} \\
\theta_i &= \beta_X X_i + e_i \\
e_k &\sim \text{Normal}(0, \sigma_{k(g)}) \\
e_i &\sim \text{Normal}(0, \sigma_{i(g)}) \\
e_{r_i} &\sim \text{Normal}(0, \sigma_{r(g)})
\end{aligned}
\tag{2}
$$

for identification purposes we can set $\frac{1}{G}\sum_{g=1}^{G}\sigma_{k(g)} = 0.02$, $\frac{1}{G}\sum_{g=1}^{G}\sigma_{i(g)} = 1$, and $\frac{1}{G}\sum_{g=1}^{G}\sigma_{r(g)} = 1$. A special case of this would be to assume that the data comes from the same population, in that case, $\sigma_{k(g)} = \sigma_k = 0.02$, $\sigma_{i(g)} = \sigma_i = 1$

### 3.4. Let's talk about Thurstone

Thurstone's comparative judgment Thurstone (1927) is based on the formula:

$$X_{AB} = \frac{S_A - S_B}{\sigma_{AB}}$$

where $X_{AB}$ defines the comparative judgment outcome, $S_A$ and $S_B$ are the modal discriminal processes, $\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2 + 2\rho\sigma_A\sigma_B}$, with $\sigma_A$ and $\sigma_B$ being the dispersion of discriminal processes $A$ and $B$, respectively, and $\rho$ the correlation between discriminal processes.

The theory identifies five cases:

- **Case 1:** only constant $\rho$ (not $\rho_{ij}$)
- **Case 2:** $X_{ij}$ becomes $X_{kij}$ with $k = 1, \ldots, K$ judges (replication, not duplication)
- **Case 3:** $\rho = 0$, then $\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2}$
- **Case 4:** $\sigma_B = \sigma_A + d$, then $\lim_{d \leq 0.1\sigma_A} \sigma_{AB} = (\sigma_A + \sigma_B)/\sqrt{2}$
- **Case 5:** $\sigma_B = \sigma_A$, then $\sigma_{AB} = \sqrt{2}\sigma$

Now using the DAG and statistical notation

$$
\begin{aligned}
O_{kijr_{ij}} &:= f_O(\delta_{kijr_{ij}}) \\
\delta_{kijr_{ij}} &= \gamma_k(\theta_{ir_i} - \theta_{jr_j}) \\
\gamma_k &= f_G(Z_k, e_k) \\
\theta_{ir_i} &= \theta_i + \beta_Y Y_r + e_{r_i} \\
\theta_i &= \beta_X X_i + e_i \\
e_k &\sim \text{Normal}(0, \sigma_{k(g)}) \\
e_i &\sim \text{Normal}(0, \sigma_{i(g)}) \\
e_{r_i} &\sim \text{Normal}(0, \sigma_{r(g)})
\end{aligned}
\tag{3}
$$

The theory identifies five cases:

- **Case 1:** only constant $\rho \approx \sigma_i$
- **Case 2:** now judges are separated by using $\gamma_k$
- **Case 3:** $\rho \approx \sigma_{e_i} = 0$ (no nesting of texts on students), then $\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2}$
- **Case 4:** $\sigma_B = \sigma_A + d$, then $\lim_{d \leq 0.1\sigma_A} \sigma_{AB} = (\sigma_A + \sigma_B)/\sqrt{2}$
- **Case 5:** $\sigma_B = \sigma_A$, then $\sigma_{AB} = \sqrt{2}\sigma$

But now can we see other scenarios than just those 5 cases?

- consider different $\rho \approx \sum_{p=1}^{P} \sigma_p$, depending on $P$ nesting structures
- we can now investigate $\gamma_k$
- we can assume $\sigma_B \neq \sigma_A$, no need for results on the limit

## 4. Discussion

### *4.1. Findings*

### *4.2. Limitations and further research*

## 5. Conclusion

### Declarations

**Funding:** The project was founded through the Research Fund of the University of Antwerp (BOF).

**Conflict of interests:** The authors declare no conflict of interest.

**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/.

**Authors' contributions:** *Conceptualization:* S.G., S.DM., T.vD., and J.M.R.E; *Methodology:* S.DM., T.vD., and J.M.R.E; *Software:* J.M.R.E.; *Validation:* J.M.R.E.; *Formal Analysis:* J.M.R.E.; *Investigation:* J.M.R.E; *Resources:* S.G., S.DM., and T.vD.; *Data curation:* J.M.R.E.; *Writing - original draft:* J.M.R.E.; *Writing - review & editing:* S.G., S.DM., and T.vD.; *Visualization:* J.M.R.E.; *Supervision:* S.G. and S.DM.; *Project administration:* S.G. and S.DM.; *Funding acquisition:* S.G. and S.DM.

# 6. Appendix

## *6.1. Additional definitions*

**Counterfactual:**

## *6.2. Why do we need to estimate judges' abilities?*

*References*


Agresti, A., 1992. Analysis of ordinal paired comparison data. Journal of the Royal Statistical Society 41, 287–297. URL: https://www.jstor.org/stable/2347562, doi:10.2307/2347562.

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.

Bradley, R., Terry, M., 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39, 324–345. URL: http://www.jstor.com/stable/2334029, doi:10.2307/2334029.

Bramley, T., 2015. Investigating the reliability of adaptive comparative judgment. URL: http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf. cambridge Assessment Research Report.

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.

Counterfactual, 2024. Merriam-webster.com dictionary. URL: https://www.merriam-webster.com/dictionary/hacker. retrieved July 23, 2024.

Crompvoets, E.A.V., Béguin, A.A., Sijtsma, K., 2022. On the bias and stability of the results of comparative judgment. Frontiers in Education 6. URL: url{https://www.frontiersin.org/articles/10.3389/feduc.2021.788202}, doi:10.3389/feduc.2021.788202.

Gijsen, M., van Daal, T., Lesterhuis, M., Gijbels, D., De Maeyer, S., 2021. The complexity of comparative judgments in assessing argumentative writing: An eye tracking study. Frontiers in Education 5. URL: url{https://www.frontiersin.org/articles/10.3389/feduc.2020.582800}, doi:10.3389/feduc.2020.582800.

Gross, J., Yellen, J., Anderson, M., 2018. Graph Theory and Its Applications. Textbooks in Mathematics, Chapman and Hall/CRC. doi:https://doi.org/10.1201/9780429425134. 3rd edition.

Jones, I., 2015. The problem of assessing problem solving: can comparative judgement help? Educational Studies in Mathematics 89, 337–355. doi:10.1007/s10649-015-9607-1.

Lesterhuis, M., 2018. The validity of comparative judgement for assessing text quality: An assessor's perspective. Ph.D. thesis. University of Antwerp.

Luce, R., 1959. On the possible psychophysical laws. The Psychologcal Review 66, 482–499. doi:10.1037/h0043178.

Marshall, N., Shaw, K., Hunter, J., Jones, I., 2020. Assessment by comparative judgement: An application to secondary statistics and english in new zealand. New Zealand Journal of Educational Studies 55, 49–71. doi:10.1007/s40841-020-00163-3.

Neal, B., 2020. Introduction to causal inference from a machine learning perspective. URL: https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf. last accessed 30 April 2024.

Neyman, J., Dabrowska, D., Speed, T., 1990. On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science 5, 465–472. URL: http://www.jstor.org/stable/2245382.

Pearl, J., 2009. Causality: Models, Reasoning and Inference. Cambrige University Press.

Pearl, J., 2010. An introduction to causal inference. The international journal of biostatistics 6, 855–859. URL: https://www.degruyter.com/document/doi/10.2202/1557-4679.1203/html, doi:10.2202/1557-4679.1203.

Pearl, J., Glymour, M., Jewell, N., 2016. Causal Inference in Statistics: A Primer. John Wiley & Sons, Inc.

Pearl, J., Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect. 1st ed., Basic Books, Inc.

Plackett, R., 1975. The analysis of permutations. Journal of the Royal Statistical Society 24, 193–202. URL: https://www.jstor.org/stable/2346567, doi:10.2307/2346567.

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.

Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701. doi:10.1037/h0037350.

Shaughnessy, J., Zechmeister, E., Zechmeister, J., 2010. Research Methods in Psychology. McGraw-Hill. URL: https://web.archive.org/web/20141015135541/http://www.mhhe.com/socscience/psychology/shaugh/ch01_concepts.html. retrieved July 23, 2024.

Thurstone, L., 1927. A law of comparative judgment. Psychological Review 34, 482–499. doi:10.1037/h0070288.

Tutz, G., 1986. Bradley-terry-luce model with an ordered response. Journal of Mathemathical Psychology 30, 306–316. doi:10.1016/0022-2496(86)90034-9.

van Daal, T., 2020. Making a choice is not easy?!: Unravelling the task difficulty of comparative judgement to assess student work. Ph.D. thesis. University of Antwerp.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., De Maeyer, S., 2019. Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. Assessment in Education: Principles, Policy & Practice 26, 59–74. doi:10.1080/0969594X.2016.1253542.

van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M., Donche, V., De Maeyer, S., 2017. The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. Frontiers in Education 2. URL: https://www.frontiersin.org/articles/10.3389/feduc.2017.00044, doi:10.3389/feduc.2017.00044.

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.

Wright, S., 1927. Correlation and causation. Correlation and causation 20, 557–585. URL: https://books.google.co.ao/books/about/Journal_of_Agricultural_Research.html?hl=pt-PT&id=lNNdIV_qpwIC.