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

Jose Manuel Rivera Espejo

Tine van van Daal

Sven De De Maeyer

Steven Gillis

2024-07-23

Abstract

Dichotomous Comparative Judgment 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, Kloots, and Gillis 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.

# Introduction

In contemporary contexts, Thurstone’s law of comparative judgment (1927) primarily refers to the method of *dichotomous* comparative judgment (DCJ, Pollitt 2012a, 2012b). 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, Kloots, and Gillis (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, Béguin, and Sijtsma 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, Glymour, and Jewell 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.

# Preliminaries

## 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, Zechmeister, and Zechmeister 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, Dabrowska, and Speed (1990) and Rubin (1974). In contrast, the structural approach was pioneered by Wright (1927), and later extended by Pearl (Pearl 2009; Pearl, Glymour, and Jewell 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, Glymour, and Jewell 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, Yellen, and Anderson 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) |

## Graphs, DAGs and SCMs

## The flow of association and causation

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

# Theoretical framework

## A scientific model for the DCJ

|  |
| --- |
| 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 |

## Probabilitics assumptions of the scientific model

## From the scientific to statistical model

for identification purposes we can set , , and . A special case of this would be to assume that the data comes from the same population, in that case, ,

## Let’s talk about Thurstone

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

where defines the comparative judgment outcome, and are the modal discriminal processes, , with and being the dispersion of discriminal processes and , respectively, and the correlation between discriminal processes.

The theory identifies five cases:

* **Case 1:** only constant (not )
* **Case 2:** becomes with judges (replication, not duplication)
* **Case 3:** , then
* **Case 4:** , then
* **Case 5:** , then

Now using the DAG and statistical notation

The theory identifies five cases:

* **Case 1:** only constant
* **Case 2:** now judges are separated by using
* **Case 3:** (no nesting of texts on students), then
* **Case 4:** , then
* **Case 5:** , then

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

* consider different , depending on nesting structures
* we can now investigate
* we can assume , no need for results on the limit

# Discussion

## Findings

## Limitations and further research

# 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.

# Appendix

## Additional definitions

**Counterfactual:**

## 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 (2): 287–97. <https://doi.org/10.2307/2347562>.

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>.

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. 2015. “Investigating the Reliability of Adaptive Comparative Judgment.” <http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.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>.

Counterfactual. 2024. “Merriam-Webster.com Dictionary.” <https://www.merriam-webster.com/dictionary/hacker>.

Crompvoets, Elise A. V., Anton A. Béguin, and Klaas 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>.

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>.

Gross, J., J. Yellen, and M. Anderson. 2018. *Graph Theory and Its Applications*. Textbooks in Mathematics. Chapman; Hall/CRC. https://doi.org/[https://doi.org/10.1201/9780429425134](https://doi.org/10.1201/9780429425134 ) .

Jones, I. 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>.

Lesterhuis, M. 2018. “The Validity of Comparative Judgement for Assessing Text Quality: An Assessor’s Perspective.” PhD thesis, University of Antwerp.

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>.

Neal, B. 2020. “Introduction to Causal Inference from a Machine Learning Perspective.” <https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf>.

Neyman, J., D. Dabrowska, and T. Speed. 1990. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” *Statistical Science* 5 (4): 465–72. <http://www.jstor.org/stable/2245382>.

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

———. 2010. “An Introduction to Causal Inference.” *The International Journal of Biostatistics* 6 (2): 855–59. <https://doi.org/10.2202/1557-4679.1203>.

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

Pearl, J., and D. Mackenzie. 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 (2): 193–202. <https://doi.org/10.2307/2346567>.

Pollitt, A. 2012a. “Comparative Judgement for Assessment.” *International Journal of Technology and Design Education* 22 (2): 157--170. <https://doi.org/10.1007/s10798-011-9189-x>.

———. 2012b. “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>.

Rubin, D. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66 (5): 688–701. <https://doi.org/10.1037/h0037350>.

Shaughnessy, J., E. Zechmeister, and J. Zechmeister. 2010. *Research Methods in Psychology*. McGraw-Hill. <https://web.archive.org/web/20141015135541/http://www.mhhe.com/socscience/psychology/shaugh/ch01_concepts.html>.

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

Tutz, G. 1986. “Bradley-Terry-Luce Model with an Ordered Response.” *Journal of Mathemathical Psychology* 30 (3): 306–16. <https://doi.org/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.” PhD thesis, University of Antwerp.

van Daal, T., M. Lesterhuis, L. Coertjens, V. Donche, and S. De Maeyer. 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 (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 Sven 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>.

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>.

Wright, S. 1927. “Correlation and Causation.” *Correlation and Causation* 20 (7): 557–85. <https://books.google.co.ao/books/about/Journal_of_Agricultural_Research.html?hl=pt-PT&id=lNNdIV_qpwIC>.