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

Jose Manuel Rivera Espejo

Tine van van Daal

Sven De De Maeyer

Steven Gillis

2024-08-02

Abstract

Dichotomous Comparative Judgment (DCJ) requires judges to compare pairs of stimuli to determine which one exhibits a higher degree of a specific trait. DCJ has proven effective and reliable across various fields (Pollitt 2012b; Jones 2015; van Daal et al. 2019; Bartholomew et al. 2018; Lesterhuis 2018; Bartholomew and Williams 2020; Marshall et al. 2020; Boonen, Kloots, and Gillis 2020). However, despite the method’s widespread use, existing literature lacks a clear explanation of the complexities and assumptions underpinning the DCJ system, as well as the plausible mechanisms through which DCJ data could be generated. This study addresses these issues by representing DCJ within the framework of causal inference. Specifically, utilizing the structural approach, the study develops a scientific model to clarify plausible causal assumptions and mechanisms inherent in the DCJ system. The study then translates this model into a probabilistic statistical model to estimate statistical relationships and infer causal effects within the system. This research provides a robust probabilistic foundation for the statistical analysis of DCJ data, building upon Thurstone’s law of comparative judgment (1927). Its findings offer valuable insights for researchers and analysts designing and implementing DCJ experiments.

# 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; Bouwer et al. 2023). Nevertheless, despite the method’s widespread use, existing literature lacks a clear representation of the plausible mechanisms through which DCJ data could be generated. Particularly, there is no depiction of the complexity and the assumptions underpinning 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 instance, van Daal et al. (2019) assessed 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; Bouwer et al. 2023), 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 framework of causal inference. Specifically, utilizing the structural approach, the study develops a scientific model to clarify plausible causal assumptions and mechanisms inherent in the DCJ system. The study then translates the scientific model into a probabilistic statistical model. This model aims to produce statistical estimates to draw inferences about plausible causal relationships within the DCJ system.

Ultimately, this study provides a robust causal and probabilistic foundation for the statistical analysis of DCJ data, building upon Thurstone’s law of comparative judgment (1927). Consequently, its findings offer valuable insights for researchers and analysts designing and implementing DCJ experiments.

# Theoretical framework

This section introduces fundamental concepts in causal inference, such as directed acyclic graphs (DAGs), structural causal models (SCMs), and the flow of association and causation in graphs. The section is not a comprehensive description of causal inference methods. Readers interested in deeper exploration should consult introductory papers like Pearl (2010), Rohrer (2018), Pearl (2019), and Cinelli, Forney, and Pearl (2020). They may also find introductory books such as Pearl and Mackenzie (2018), Neal (2020) and McElreath (2020) useful. For more advanced study, seminal intermediate papers like Neyman (1923), Rubin (1974), Spirtes, Glymour, and Scheines (1991), and Sekhon (2009), as well as books such as Pearl (2009), Morgan and Winship (2014) and Hernán and Robins (2020) are recommended.

## 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 notation for analyzing causes and counterfactuals (Pearl 2009).

Counterfactuals represent scenarios *contrary to fact*, where alternative *potential* outcomes resulting from a cause are neither observed nor observable (Neal 2020; Counterfactual 2024). According to Pearl and Mackenzie (2018), counterfactuals are the foundation of causal inference and occupy the highest level of cognitive abstraction in the ladder of causation, followed by intervention and association. Nevertheless, despite their abstract nature, counterfactuals enable the development of a *theory of the world* that explains why specific causes have specific effects and what occurs in their absence (Pearl and Mackenzie 2018). They achieve this by translating causal statements into counterfactual statements, that is, statements about “what would have happened in the world under different circumstances.”

Several approaches to causal inference and counterfactuals exist, but two are particularly prominent: the potential outcomes approach, also known as the Neyman-Rubin causal model (Neyman 1923; Rubin 1974), and the structural approach (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 employs the do-operator and structural causal models (SCM, Pearl 2009; Pearl, Glymour, and Jewell 2016). Despite these differences, both notations can be expressed in terms of the other, and both approaches provide methods for using experimental and observational data to estimate causal effects (Pearl 2010).

Nevertheless, the structural approach offers an additional key advantage over the potential outcomes approach: it enables the graphical representation of any system through directed acyclic graphs (DAG, Gross, Yellen, and Anderson 2018; Neal 2020). DAGs are heuristics that can effectively convey the assumed causal structure of a system. They do not represent detailed statistical models but allow researchers to deduce which statistical models can provide valid causal inferences, assuming the causal structure depicted in the DAG is accurate (McElreath 2020).

## Directed acyclic graphs (DAG)

Graph theory is a branch of mathematics focused on the study of graphs. Graphs are mathematical structures modeling pairwise relations between objects. They can represent physical relations, such as electrical circuits and roadways, and less tangible structures, such as ecosystems and sociological relations. Graphs have proven useful in various fields, including computer science, operations research, and the natural and social sciences (Gross, Yellen, and Anderson 2018).

In statistics, one application incorporating concepts from graph theory is causal inference. Specifically, the structural approach to causal inference uses directed acyclic graphs (DAG) to provide a formal and graphical representation of the causal structure of a system (Neal 2020). In this context, a *graph* is a collection of nodes connected by edges, where nodes represent random variables. The term *directed* indicates that the edges of the graph extend from one node to another, with arrows showing the direction of causal influence. Moreover, the term *acyclic* indicates the causal influences do not form a loop, meaning the influences do not cycle back on themselves (McElreath 2020).

One key advantage of DAGs is that, regardless of complexity, they can represent various causal structures using only five fundamental building blocks (Neal 2020; McElreath 2020; Morgan and Winship 2014). [Figure 1](#fig-dags) illustrates these blocks. [Figure 1 (a)](#fig-dag_bb1) depicts two unconnected nodes, representing an scenario where variables and are not causally related. [Figure 1 (b)](#fig-dag_bb2) shows two connected nodes, illustrating a scenario where a parent node exerts a causal influence on a child node . Consequently, is considered a *descendant* of . [Figure 1 (c)](#fig-dag_bb3) depicts a *chain* (or *pipe*), where influences , and influences . In this configuration, is a parent node of , and a parent node of . Furthermore, the DAG shows that is an *ancestor* of and that the relationship between these variables is entirely *mediated* by . [Figure 1 (d)](#fig-dag_bb4) illustrates a *fork*, where variables and are both influenced by . In this scenario is a parent node of both and . Finally, [Figure 1 (e)](#fig-dag_bb5) depicts a *collider*, also known as *inmorality*, where variables and are concurrent causes of . In this configuration, and are not causally related to each other but both influence .

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Two unconnected nodes | |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | (b) Two connected nodes or descendant | |  | |  | | --- | | (c) Chain or pipe | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | | --- | | (d) Fork | |  |

|  |  |
| --- | --- |
| |  | | --- | | (e) Collider or inmorality | |

Figure 1: DAG’s fundamental building blocks.

Given the heuristic nature of DAGs, a motivating example can help illustrate the use of the five fundamental building blocks to construct a system’s causal structure. This example can also help to clarify some additional conventions for drawing DAGs.

Consider a research problem where the causal effect of a variable on an outcome needs to be investigated. Additionally, the problem suggests that a variable potentially influences both and . Such scenarios are not hard to imagine. For instance, might represent different treatments that could affect the recovery from cancer , while could denote the patient’s age. Similarly, in the context of a DCJ study like the one described by Boonen, Kloots, and Gillis (2020), could represent the duration of a child’s cochlear implant use, which might influence the child’s overall speech quality , with indicating the child’s hearing status. Note that these descriptions do not specify the type of data or the functional forms of the variables, highlighting the heuristic nature of DAGs. (not a bad example, but I prefer one using writing skills)

[Figure 2](#fig-example) presents two graphs illustrating the plausible causal structure of the motivating example. [Figure 2 (a)](#fig-dag_example1) shows the *simplified* DAG. In this graph, the variables of interest or *endogenous* variables are depicted as solid black circles, indicating that they are observed variables. The arrows in the graph reflect the expected direction of causal influences among the variables. In contrast, [Figure 2 (b)](#fig-dag_example2) depicts the *magnified* version of the DAG. This graph includes, in addition to the endogenous variables, the *exogenous* variables . These exogenous variables are commonly referred to as *disturbances* or *errors*, because they represent factors not explicitly modeled. They are depicted as open circles, indicating their unobserved nature. These exogenous variables are often omitted for simplicity, resulting in the simplified DAG in [Figure 2 (a)](#fig-dag_example1). However, including them in the graph can be advantageous in certain scenarios, as they can help to highlight potential issues related to conditioning and confounding (Cinelli, Forney, and Pearl 2020). These concepts are introduced in the next section.

Furthermore, a detailed examination of both figures reveals the presence of at least four of the five fundamental building blocks. [Figure 2 (a)](#fig-dag_example1) displays multiple descendants, evident in pairwise relations such as , , and . It also illustrates a fork with . Similarly, [Figure 2 (b)](#fig-dag_example2) features multiple two-unconnected nodes, evident in the pairwise relations , , and , as well as colliders such as and . The symbol denotes a concept know as *d-separation*, which roughly implies the independence of the variables. This concept is introduced in the next section.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Simplied DAG | |  |

|  |  |
| --- | --- |
| |  | | --- | | (b) Magnified DAG | |

Figure 2: DAGs for a plausible causal structure in a system.

## The flow of association and causation in graphs

|  |
| --- |
| Figure 3: The flow of association and causation in graphs. Extracted from Neal (2020, 31) |

## But where does it all fit?

|  |
| --- |
| Figure 4: Identification-Estimation flowchart. Extracted from Neal (2020, 32) |

# Theory

## A scientific model for the DCJ

|  |
| --- |
| Figure 5: DCJ causal diagram, simplified description |

|  |
| --- |
| Figure 6: DCJ causal diagram, simplified mathematical description |

|  |
| --- |
| Figure 7: DCJ causal diagram, population mathematical description |

|  |
| --- |
| Figure 8: DCJ causal diagram, sample with comparisons mathematical description |

## Probabilitics assumptions of the scientific model

## From the scientific to statistical model

## Let’s talk about Thurstone

# 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:** Author XX serve on advisory broad of Company Y but receives no compensation this role.

**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

## Why do we need to estimate judges’ abilities?

## Latent variables as a mean of imputation

## Other comparative scenarios

## 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>.

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. 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>.

Cinelli, C., A. Forney, and J. Pearl. 2020. “A Crash Course in Good and Bad Controls.” *SSRN*, September. <https://doi.org/10.2139/ssrn.3689437>.

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 ) .

Hernán, M., and J. Robins. 2020. *Causal Inference: What If*. 1st ed. Chapman; Hall/CRC. <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book>.

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>.

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

Morgan, S., and C. Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Analytical Methods for Social Research. Cambridge University Press.

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. 1923. “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>.

———. 2019. “The Seven Tools of Causal Inference, with Reflections on Machine Learning.” *Communications of the ACM* 62 (3): 54–60. <https://doi.org/10.1177/0962280215586010>.

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>.

Rohrer, J. 2018. “Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data.” *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42. <https://doi.org/10.1177/2515245917745629>.

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>.

Sekhon, J. 2009. “The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods.” In *The Oxford Handbook of Political Methodology*, edited by J Box-Steffensmeier, H. Brady, and D. Collier, 271–99. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199286546.003.0011>.

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>.

Spirtes, P., C. Glymour, and R. Scheines. 1991. “From Probability to Causality.” *Philosophical Studies* 64 (1): 1–36. <https://www.jstor.org/stable/4320244>.

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>.