Causes and effects in Dichotomous Comparative Judgments: an information-theoretical system of 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) 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 et al., 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.

Keywords: causal inference, probability, Thurstone, comparative judgement, directed acyclic graph, structural causal models, statistical modeling

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). 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

*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)

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; 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 et al., 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.

2. 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 et al. (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 et al. (1991), and Sekhon (2009), as well as books such as Pearl (2009), Morgan and Winship (2014) and Hernán and Robins (2020) are recommended.

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 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 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 employs the do-operator and structural causal models (SCM, Pearl, 2009; Pearl et al., 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).

However, 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 et al., 2018; Neal, 2020). DAGs serve as heuristics, effectively conveying the assumed causal structure of a system. A heuristic is a strategy that simplifies information to make decisions more quickly, efficiently, and sometimes more accurately than complex methods (Chow, 2015). Consequently, DAGs 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 DAGs are accurate (McElreath, 2020).

2.2. DAGs and SCMs

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 et al., 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 graphical and formal representation of the causal structure of a system (Neal, 2020). In this context, a *graph* denotes a collection of nodes connected by edges, where nodes represent random variables. The term *directed* indicates 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).

DAGs offer two key advantages for modeling causality. Firstly, DAGs represent causal relations in a nonparametric and fully interactive manner. This feature implies that feasible causal analysis strategies usually do not require specifying the type of data or the nature of the functional dependence among variables (Morgan and Winship, 2014). Secondly, regardless of complexity, DAGs can represent various causal structures using only five fundamental building blocks (Neal, 2020; McElreath, 2020). These building blocks can be depicted in three ways: the magnified representation, the standard representation, and the structural causal model form (SCM, Morgan and Winship, 2014).

The left panels of Figure 1 illustrate the *magnified* representation. These graphs depict the *endogenous* variables $V = \{X, Z, Y\}$ alongside the *exogenous* variables $E = \{e_X, e_Z, e_Y\}$. Endogenous variables are those whose causal mechanisms the investigator chooses to model (Neal, 2020). In contrast, exogenous variables represent *errors* or *disturbances* arising from omitted factors that the investigator chooses not to model explicitly (Pearl, 2009, 27,68). The graphs show endogenous variables as solid black circles to signify that they are observed random variables, while endogenous variables are depicted as open circles to signify their unobserved (latent) nature. Lastly, the arrows in the graphs reflect the expected direction of causal influences among these variables.

Often, the DAGs omit the exogenous variables for simplicity, resulting in the *standard* representation. However, including exogenous variables in a graph can be beneficial in some scenarios, as their presence can reveal potential issues related to conditioning and confounding (Cinelli et al., 2020), concepts explored in the following section. The standard representation is illustrated in the middle panels of Figure 1.

Lastly, the right panels of Figure 1 depict the SCM form of the five fundamental building blocks. SCMs are formal mathematical models that represent causal relationships within a system or population (Hitchcock, 2024). An SCM is defined by a set of endogenous variables V, a set of exogenous variables E, and a set of functions $F = \{f_X, f_Z, f_Y\}$ (Pearl, 2009; Neal, 2020). These functions, referred to as structural equations, specify each endogenous variable as nonparametric functions of other variables. Moreover, SCMs use the symbol ':=' to indicate the variables' asymmetrical causal dependence and the symbol '$\perp\!\!\!\perp$' to represent *d-separation*, which roughly equates to the concept of variable independence. The concepts of d-separation and causal (in)dependence are explored in the following section.

A careful examination of Figure 1 highlights the assumptions underlying these building blocks. Figures 1a, 1b, and SCM 1c depict two unconnected nodes, representing a scenario where variables X and Y are not causally related. Figures 1d, 1e, and SCM 1f illustrate two connected nodes, showing a scenario where a *parent* node X exerts a causal influence on a *child* node Y. Consequently, Y is considered a *descendant* of X. Figures 1g, 1h, and SCM 1i depict a *chain* or *pipe*, where X influences Z, and Z influences Y. In this configuration, X is a parent node of Z, and Z is a parent node of Y. Furthermore, the DAGs show that X is an *ancestor* of Y and that Z fully *mediates* the relationship between these variables. Figures 1j, 1k, and SCM 1l illustrate a *fork*, where variables X and Y are both influenced by Z. Here, Z is a parent node of X and Y. Finally, Figures 1m, 1n, SCM 1o depict a *collider*, also known as *inmorality*, where variables X and Y are concurrent causes of Z. In this configuration, X and Y are not causally related to each other but both influence Z. Additionally, in all SCMs, the errors are assumed to be mutually independent of each other and of all other variables in the graph, as evidenced by the pairwise relations $e_X \perp\!\!\!\perp e_Y$, $e_X \perp\!\!\!\perp e_Z$, and $e_Z \perp\!\!\!\perp e_Y$.

Given the heuristic nature of DAGs, a motivating example can help clarify how to use the five fundamental building blocks to construct the causal structure of a system. Consider a research problem where the causal effect of a variable T on an outcome Y needs investigation. The problem suggests that a variable X potentially influences both T and Y. Beyond these relationships, the problem does not specify any further variables of interest. Such scenarios are not hard to imagine. For instance, T might represent different treatments that could affect the recovery from cancer Y, while X could denote the patient's age. Similarly, in the context of a DCJ study like the one described by Boonen et al. (2020), T could represent the duration of a child's cochlear implant use, which might influence the child's overall speech quality Y, with X indicating the child's hearing status. (not a bad example, but I prefer one using writing skills)

Figure 2 illustrates the plausible causal structure of the motivating example. A detailed examination of Figures 2a, 2b, and SCM 2c reveals the presence of at least four of the five fundamental building blocks. The figures display multiple descendants, as indicated by pairwise relations such as $X \to T$, $X \to Y$, and $T \to Y$. Additionally, the figures features multiple pairs of unconnected nodes, evident from the relations $e_T \perp\!\!\!\perp e_X$, $e_T \perp\!\!\!\perp e_Y$, and $e_X \perp\!\!\!\perp e_Y$. Finally, the figures illustrate the fork $X \to \{T, Y\}$, and two colliders with $\{X, e_T\} \to T$ and $\{X, T, e_Y\} \to Y$.

(a) Two unconnected nodes

(b) Two unconnected nodes

$$X := f_X(e_X)$$
$$Y := f_Y(e_Y)$$
$$e_X \perp\!\!\!\perp e_Y$$

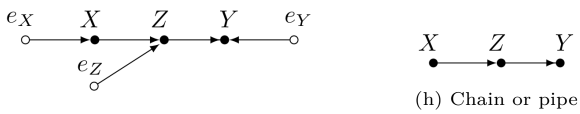
(c) Two unconnected nodes

(d) Two connected nodes or descendant

(e) Two connected nodes or descendant

$$X := f_X(e_X)$$
$$Y := f_Y(X, e_Y)$$
$$e_X \perp\!\!\!\perp e_Y$$

(f) Two connected nodes or descendant

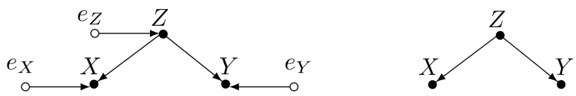


(g) Chain or pipe

(h) Chain or pipe

$$X := f_X(e_X)$$
$$Z := f_Z(X, e_Z)$$
$$Y := f_Y(Z, e_Y)$$
$$e_X \perp\!\!\!\perp e_Y$$
$$e_X \perp\!\!\!\perp e_Z$$
$$e_Z \perp\!\!\!\perp e_Y$$

(i) Chain or pipe

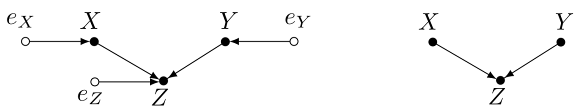


(j) Fork

(k) Fork

$$X := f_X(Z, e_X)$$
$$Z := f_Z(e_Z)$$
$$Y := f_Y(Z, e_Y)$$
$$e_X \perp\!\!\!\perp e_Y$$
$$e_X \perp\!\!\!\perp e_Z$$
$$e_Z \perp\!\!\!\perp e_Y$$

(l) Fork



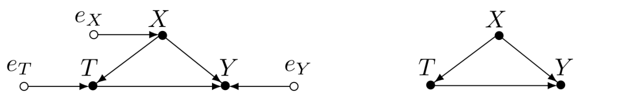
(m) Collider or inmorality

(n) Collider or inmorality

$$X := f_X(e_X)$$
$$Z := f_Z(X, Y, e_Z)$$
$$Y := f_Y(e_Y)$$
$$e_X \perp\!\!\!\perp e_Y$$
$$e_X \perp\!\!\!\perp e_Z$$
$$e_Z \perp\!\!\!\perp e_Y$$

(o) Collider or inmorality

Figure 1: The five fundamental building blocks of DAGs. **Note:** left panels show the magnified representation, middle panels show the standard representation, and the right panels show their corresponding SCM form.



(a) Magnified representation (b) Standard representation

$$
\begin{aligned}
X &:= f_X(e_X) \\
T &:= f_Z(X, e_T) \\
Y &:= f_Y(T, X, e_Y) \\
e_T &\perp\!\!\!\perp e_X \\
e_T &\perp\!\!\!\perp e_Y \\
e_X &\perp\!\!\!\perp e_Y
\end{aligned}
$$

(c) Structural causal model

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

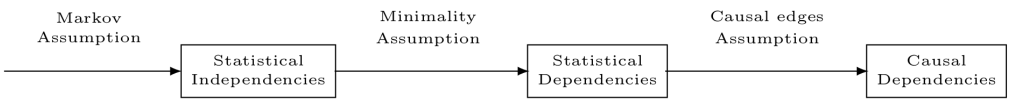


Figure 3: The flow of association and causation in graphs. Extracted from Neal (2020, 31)

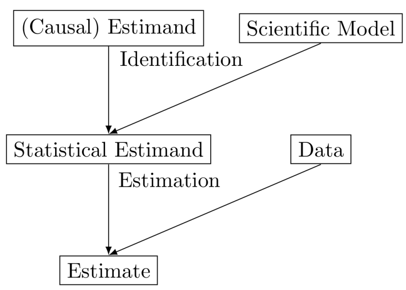


Figure 4: Identification-Estimation flowchart. Extracted from Neal (2020, 32)

2.3. *The flow of association and causation in graphs*

2.4. *But where does it all fit?*

3. Theory

3.1. *A scientific model for the DCJ*

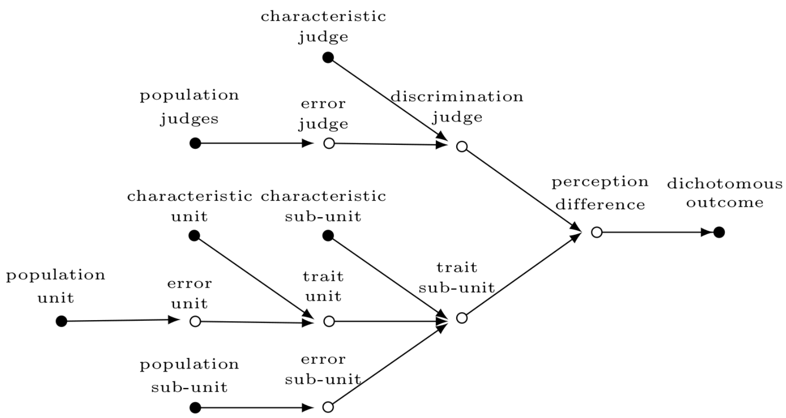


Figure 5: DCJ causal diagram, simplified description

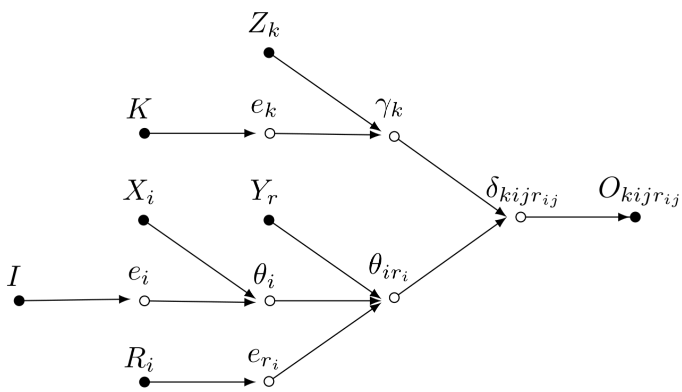


Figure 6: DCJ causal diagram, simplified mathematical description

3.2. *Probabilitics assumptions of the scientific model*

3.3. *From the scientific to statistical model*

3.4. *Let's talk about Thurstone*

4. Discussion

4.1. *Findings*

4.2. *Limitations and further research*

5. Conclusion

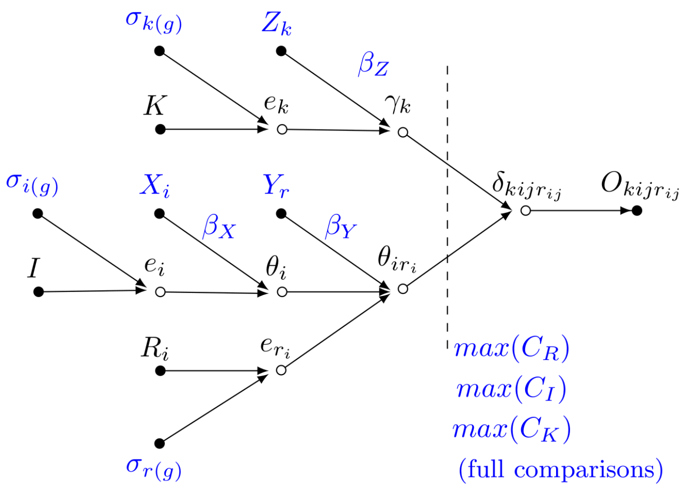


Figure 7: DCJ causal diagram, population mathematical description

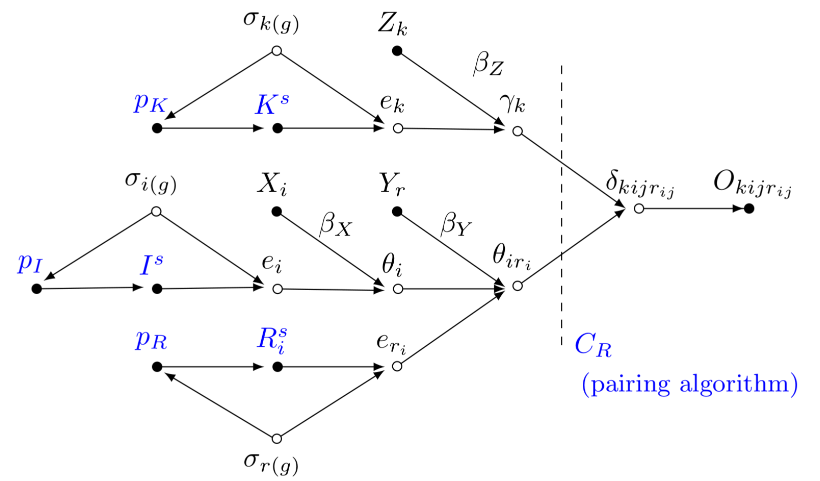


Figure 8: DCJ causal diagram, sample with comparisons mathematical description

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.

6. Appendix

6.1. Why do we need to estimate judges' abilities?

6.2. Latent variables as a mean of imputation

6.3. Other comparative scenarios

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.

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., De Maeyer, S., 2023. Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. Journal of Writing Research 15, 497–518. URL: https://www.jowr.org/index.php/jowr/article/view/867, doi:10.17239/jowr-2024.15.03.03.

Bradley, R., Terry, M., 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39, 324–345. 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.

Chow, S., 2015. Many meanings of 'heuristic'. The British Journal for the Philosophy of Science 66, 977–1016. doi:10.1093/bjps/axu028.

Cinelli, C., Forney, A., Pearl, J., 2020. A crash course in good and bad controls. SSRN URL: https://ssrn.com/abstract=3689437, doi:10.2139/ssrn.3689437.

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.

Hernán, M., Robins, J., 2020. Causal Inference: What If. 1 ed., Chapman and Hall/CRC. URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book. last accessed 31 July 2024.

Hitchcock, C., 2024. Causal Models, in: Zalta, E.N., Nodelman, U. (Eds.), The Stanford Encyclopedia of Philosophy. Summer 2024 ed.. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2023/entries/causal-models/.

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.

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

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

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., 1923. 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. translated by Dabrowska, D. and Speed, T. (1990).

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., 2019. The seven tools of causal inference, with reflections on machine learning. Communications of the ACM 62, 54–60. doi:10.1177/0962280215586010.

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.

Rohrer, J., 2018. Thinking clearly about correlations and causation: Graphical causal models for observational data. Advances in Methods and Practices in Psychological Science 1, 27–42. doi:10.1177/2515245917745629.

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

Sekhon, J., 2009. The neyman-rubin model of causal inference and estimation via matching methods, in: Box-Steffensmeier, J., Brady, H., Collier, D. (Eds.), The Oxford Handbook of Political Methodology. Oxford University Press, pp. 271–299. doi:10.1093/oxfordhb/9780199286546.003.0011.

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.

Spirtes, P., Glymour, C., Scheines, R., 1991. From probability to causality. Philosophical Studies 64, 1–36. URL: https://www.jstor.org/stable/4320244.

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.