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

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

This study revisits Thurstone's law of comparative judgments (CJ) by addressing two key limitations in traditional approaches. Firstly, it addresses the overreliance on the assumptions of Thurstone's Case V in the statistical analysis of CJ data. Secondly, it addresses the apparent disconnect between CJ's approach to trait measurement and hypothesis testing. We put forward a systematic approach based on causal analysis and Bayesian statistical methods, which results in a model that facilitates a more comprehensive understanding of the factors influencing CJ experiments while offering a robust statistical translation. The new model accommodates unequal dispersions and correlations between stimuli, enhancing the reliability and validity of CJ's trait estimation, thereby ensuring the accurate measurement and interpretation of comparative data. The paper highlights the relevance of this updated framework for modern empirical research, particularly in education and social sciences. This contribution advances current research methodologies, providing a robust foundation for future applications in diverse fields.

Keywords: causal inference, directed acyclic graphs, structural causal models, bayesian statistical methods, thurstonian model, comparative judgement, probability, statistical modeling

Introduction

In *comparative judgment* (CJ) studies, judges assess a specific trait or attribute across different stimuli by performing pairwise comparisons (Thurstone, 1927b,a). Each comparison produces a dichotomous outcome, indicating which stimulus is perceived to have a higher trait level. For example, when assessing writing quality, judges compare pairs of written texts (the stimuli) to

*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)

determine the relative writing quality each text exhibit (the trait) (Laming, 2004; Pollitt, 2012b; Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018a; Coertjens et al., 2017; Goossens and De Maeyer, 2018; Bouwer et al., 2023).

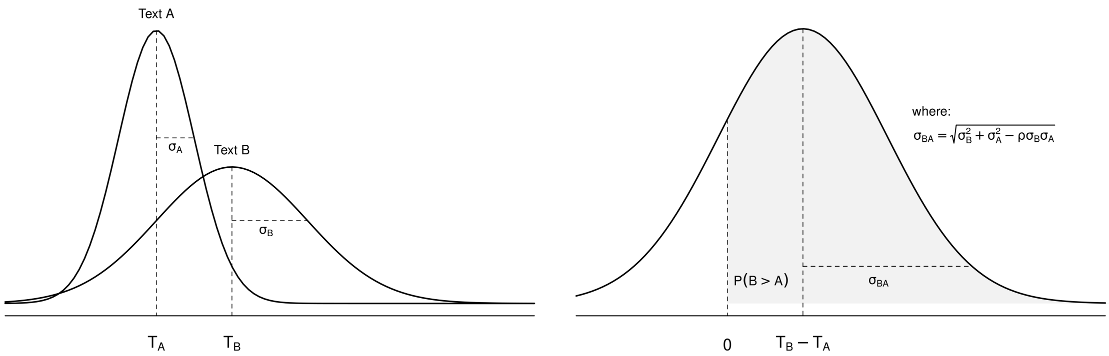
Numerous studies have documented the effectiveness of CJ in assessing traits and competencies over the past decade. These studies have highlighted three aspects of the method's effectiveness: its reliability, validity, and practical applicability. Research on reliability suggests that CJ requires a relatively modest number of pairwise comparisons (Verhavert et al., 2019; Crompvoets et al., 2022) to generate trait scores that are as precise and consistent as those generated by other assessment methods (Coertjens et al., 2017; Goossens and De Maeyer, 2018; Bouwer et al., 2023). In addition, the evidence suggests that the reliability and time efficiency of CJ are comparable, if not superior, to those of other assessment methods when employing adaptive comparison algorithms (Pollitt, 2012b; Verhavert et al., 2022; Mikhailiuk et al., 2021). Meanwhile, research on the validity of CJ scores indicates their capacity to represent the traits under measurement accurately (Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018a; Bartholomew et al., 2018; Bouwer et al., 2023). Moreover, research on CJ's practical applicability highlights its versatility across both educational and non-educational contexts (Kimbell, 2012; Jones and Inglis, 2015; Bartholomew et al., 2018; Jones et al., 2019; Marshall et al., 2020; Bartholomew and Williams, 2020; Boonen et al., 2020).

Nevertheless, despite the increasing number of CJ studies, the prevalence of unsystematic and fragmented research approaches has left several critical issues unaddressed. The present study primarily focuses on two issues: the overreliance on Thurstone's Case V assumptions in the statistical analysis of CJ data and the apparent disconnect between CJ's approach to trait measurement and hypothesis testing. The following sections begin with a brief overview of Thurstone's theory followed by a detailed examination of these issues. Subsequently, the study introduces a theoretical model for CJ that builds upon Thurstone's theory, alongside its statistical translation, designed to address the two concerns simultaneously.

Thurstone's theory

In its most general form, Thurstone's theory addresses pairwise comparisons wherein a single judge evaluates multiple stimuli (Thurstone, 1927a, pp. 267). The theory posits that two key factors determine the dichotomous outcome of these comparisons: the discriminal process of each stimulus and their discriminal difference. The *discriminal process* captures the psychological impact each stimulus exerts on the judge or, more simply, his perception of the stimulus trait. The theory assumes that the discriminal process for any given stimulus forms a Normal distribution along the trait continuum (Thurstone, 1927a, pp. 266). The mode (mean) of this distribution, known as the *modal discriminal process*, indicates the stimulus position on this continuum, while its dispersion, referred to as the *discriminal dispersion*, reflects variability in the perceived trait of the stimulus.

Figure 1a illustrates hypothetical discriminal processes along a quality trait continuum for two written texts. The figure indicates that the modal discriminal process for Text B is positioned further along the continuum than that of Text A ($T_B > T_A$), suggesting that Text B exhibits higher quality. Additionally, the figure highlights that Text B has a broader distribution compared to Text A, which arises from its larger discriminal dispersion ($\sigma_B > \sigma_A$).



(a) Discriminal processes

(b) Discriminal difference

Figure 1: Hypothetical discriminal processes and discriminant difference along a quality trait continuum for two written texts.

However, given that the individual discriminal processes of the stimuli are not directly observable, the theory introduces the *law of comparative judgment*. This law posits that in pairwise comparisons, a judge perceives the stimulus with a discriminal process positioned further along the trait continuum as possessing more of the trait (Bramley, 2008, pp. 251). This suggests that the relative distance between stimuli, rather than their absolute positions on the continuum, likely defines the outcome of pairwise comparisons. Indeed, the theory assumes that the difference between the underlying discriminal processes of the stimuli, referred to as the *discriminal difference*, determines the observed dichotomous outcome. Furthermore, the theory assumes that because the individual discriminal processes form a Normal distribution on the continuum, the discriminal difference will also conform to a Normal distribution (Andrich, 1978). In this distribution, the mode (mean) represents the relative separation between the stimuli, and its dispersion indicates the variability of that separation.

Figure 1b illustrates the distribution of the discriminal difference for the hypothetical texts depicted in Figure 1a. The figure indicates that the judge perceives Text B as having significantly higher quality than Text A. This conclusion is supported by two key observations: the positive difference between their modal discriminal processes ($T_B - T_A > 0$) and the probability area where the discriminal difference distinctly favors Text B over Text A, represented by the shaded gray area denoted as $P(B > A)$. As a result, the dichotomous outcome of this comparison is more likely to favor Text B over Text A.

The two critical issues in CJ literature

This section examines the two critical issues in the CJ literature that serve as the primary focus of the present study. The first is related to the overreliance on Thurstone's Case V assumptions in the statistical analysis of CJ data. The second concern with the apparent disconnect between CJ's approach to trait measurement and hypothesis testing.

1. The Case V and the statistical analysis of CJ data

Thurstone noted from the outset that the general form of the theory, as outlined in Section **??**, gave rise to a problem of trait scaling. The model required estimating more "unknown" parameters than the available pairwise comparisons (Thurstone, 1927a, pp. 267). To address this issue and facilitate the practical implementation of the theory, he developed five cases derived from this general form, each case progressively incorporated additional simplifying assumptions into the model.

In Case I, Thurstone postulated that pairs of stimuli would maintain a constant correlation across all comparisons. In Case II, he allowed multiple judges to undertake comparisons instead of confining evaluations to a single judge. In Case III, he posited that there was no correlation between stimuli. In Case IV, he assumed that the stimuli exhibited similar dispersions. Finally, in Case V, he replaced this assumption with the condition that stimuli had equal discriminal dispersions. Table 1 summarizes the assumptions of the general form and the five cases. For a detailed discussion of these cases and their progression, refer to Thurstone (1927a) and Bramley (2008, pp. 248–253).

Notably, despite relying on the most extensive set of simplifying assumptions (Bramley, 2008, pp. 253; Kelly et al., 2022, pp. 677), Case V remains the most widely used case in the CJ literature. This popularity stems mainly from its simplified statistical representation in the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959). The BTL model mirrors the assumptions of Case V, with one notable distinction: whereas Case V assumes a Normal distribution for the

Table 1: Thurstones cases and their asumptions

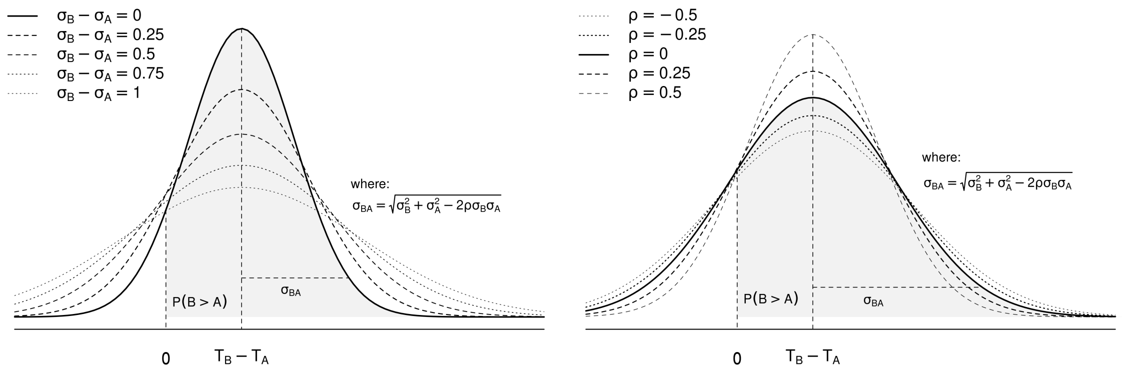
Assumption	**General form**	**Thurstone's** Case I	Case II	Case III	Case IV	Case V	**BTL model**
Discriminal process (distribution)	Normal	Normal	Normal	Normal	Normal	Normal	Logistic
Discriminal dispersion (between stimuli)	Different	Different	Different	Different	Similar	Equal	Equal
Correlation (between stimuli)	One per pair	Constant	Constant	Zero	Zero	Zero	Zero
How many judges compare?	Single	Single	Multiple	Multiple	Multiple	Multiple	Multiple

stimuli's discriminal processes, the BTL model uses the more mathematically tractable Logistic distribution (Andrich, 1978; Bramley, 2008, pp. 254) (see Table 1). This substitution has little impact on the model's estimation or interpretation, as the Normal and Logistic distributions exhibit analogous statistical properties, differing only by a scaling factor of approximately 1.7 (van der Linden, 2017a, pp. 16).

However, Thurstone originally developed Case V to provide a "rather coarse scaling" of traits (Thurstone, 1927a, pp. 269), prioritizing statistical simplicity over precision in trait measurement (Kelly et al., 2022, pp. 677). He explicitly warned against its untested application, stating that its use "should not be made without (an) experimental test" (Thurstone, 1927a, pp. 270). Furthermore, he acknowledged that some assumptions could prove problematic when researchers assess complex traits or heterogeneous stimuli (Thurstone, 1927b, pp. 376). Consequently, given that modern CJ applications frequently involve such traits and stimuli, two main assumptions of Case V and, by extension, of the BTL model may not consistently hold in theory or practice, namely the assumption of equal dispersion and zero correlation between stimuli.

1.1. The assumption of equal dispersions between stimuli

According to the theory, discrepancies in the discriminal dispersions of stimuli shape the distribution of the discriminal difference, exerting a direct influence on the outcome of pairwise comparisons. Figure 2a presents a thought experiment to illustrate this idea. In this experiment, a researcher can observe the discriminal processes for the texts depicted in Figure 1a. Furthermore, the figure assumes that the discriminal dispersion for Text A remains constant and that the texts are uncorrelated ($\rho = 0$). The figure reveals that an increase in the uncertainty associated with the perception of Text B in comparison to Text A, ($\sigma_B - \sigma_A$), broadens the distribution of their discriminal difference. This broadening affects the probability area where the discriminal difference distinctly favors Text B over Text A, expressed as $P(B > A)$, ultimately influencing the comparison outcome. Additionally, the figure reveals that when the discriminal dispersions of the texts are equal ($\sigma_B - \sigma_A = 0$), the discriminal difference is more likely to favor Text B over Text A (shaded gray area), compared to situations where their dispersions differ.



(a) Discrepancies in the dispersions of stimuli

(b) Correlation between stimuli

Figure 2: The effect of dispersion discrepancies and stimulus correlation on the distribution of the discriminal difference.

In experimental practice, however, this process occurs in reverse. Researchers first observe the comparison outcome and then use the BTL model to infer the discriminal difference between the stimuli and their respective discriminal processes (Thurstone, 1927b, pp. 373). Therefore, the outcome's ability to reflect the "true" differences between stimuli largely depends on the validity of the model's assumptions (Kohler et al., 2019, pp. 150), particularly the assumption of equal dispersions. For instance, when researchers observe a sample of outcomes favoring Text B over Text A and correctly assume equal dispersions between the texts, the BTL model estimates a discriminal difference distribution that accurately represents the "true" discriminal difference of the texts. This scenario is illustrated in Figure 2a, where the model's discriminal difference distribution aligns with the "true" distribution, represented by the thick continuous line corresponding to $\sigma_B - \sigma_A = 0$. The accuracy of these discriminal difference ensures reliable estimates for the texts' discriminal processes (citation needed?).

However, Thurstone argued that the assumption of equal dispersions may not be applicable when researchers assess complex traits or heterogeneous stimuli (Thurstone, 1927b, pp. 376), as these traits and stimuli can introduce judgment discrepancies due to their unique features (van Daal et al., 2016; Lesterhuis, 2018b; Chambers and Cunningham, 2022). Indeed, evidence of this violation may already be present in the CJ literature in the form of misfit statistics, which measure judgment discrepancies associated with specific stimuli (Pollitt, 2004, pp. 12; Goossens and De Maeyer, 2018, pp. 20). For example, labeling texts as “misfits” indicates that comparisons involving these texts result in more judgment discrepancies than those involving other texts (Pollitt, 2012a,b; van Daal et al., 2016; Goossens and De Maeyer, 2018). These discrepancies, in turn, suggest that the discriminal differences for “misfit” texts have broader distributions, indicating that their discriminal processes may also exhibit more variation than that of other texts. A similar line of reasoning applies to the concept of “misfit” judges, whose evaluations deviate substantially from the shared consensus due to the unique characteristics of the stimuli or the judges themselves. These “misfit” judges and their associated deviations can give rise to additional statistical and measurement issues, which we discuss in more detail in Section 1.2.

Thus, model misspecification, in the form of an erroneous assumption of equal dispersions between stimuli, can give rise to significant statistical and measurement issues. For instance, the model may overestimate the degree to which the outcome accurately reflects the “true” discriminal differences between stimuli. This overestimation can result in researchers drawing spurious conclusions about these differences (McElreath, 2020, pp. 370) and, by extension, about the underlying discriminal processes of stimuli. Figure 2a also illustrates this issue when the model's discriminal difference distribution aligns with the thick continuous line for $\sigma_B - \sigma_A = 0$, while the “true” discriminal difference follows any discontinuous line where $\sigma_B - \sigma_A \neq 0$. Additionally, if researchers recognize that misfit statistics highlight these critical differences in dispersions, the conventional CJ practice of excluding stimuli based on these statistics (Pollitt, 2012a,b; van Daal et al., 2016; Goossens and De Maeyer, 2018) can unintentionally discard valuable information. Such exclusions can introduce bias into trait estimates (Zimmerman, 1994; McElreath, 2020, chap. 12). The direction and magnitude of these biases are often unpredictable, as they depend on which stimuli are excluded from the analysis.

1.2. The assumption of zero correlation between stimuli

The correlation, represented by the symbol ρ, measures how much a judge's perception of a specific trait in one stimulus depends on their perception of the same trait in another. As with the discriminal dispersions, this correlation shapes the distribution of the discriminal difference, directly impacting the outcomes of pairwise comparisons. Figure 2b presents a similar thought experiment as in Section 1.1 to illustrate this idea. The illustration now assumes that the discriminal dispersions for both texts remain constant. The figure reveals that as the correlation between the texts increases, the distribution of their discriminal difference becomes narrower. This narrowing affects the area under the curve where the discriminal difference distinctly favors Text B over Text A, denoted as

$P(B > A)$, thus influencing the comparison outcome. Furthermore, the figure shows that when two texts are independent or uncorrelated ($\rho = 0$), their discriminal difference is less likely to favor Text B over Text A (shaded gray area) compared to scenarios where the texts are highly correlated.

Off course, in experimental practice, researchers approach this process in reverse. They begin by observing the sample of outcomes favoring Text B over Text A and then use the BTL model to estimate the discriminal difference and the discriminal processes of the stimuli. Given that the BTL model assumes independent discriminal processes across comparisons, if this assumption holds, then the model estimates a discriminal difference distribution that accurately reflects the "true" discriminal difference of the texts. This scenario is also illustrated in Figure 2b when the discriminal difference distribution of the model aligns with the "true" distribution, represented by the thick continuous line corresponding to $\rho = 0$. Once more, the estimation accuracy of the discriminal difference ensures reliable estimates for the discriminal processes of the texts (citation needed?).

Notably, Thurstone attributed the independence of stimuli to the cancellation of potential judges' biases. He argued that this cancellation resulted from two opposing and equally weighted effects occurring during pairwise comparisons (Thurstone, 1927a, pp. 268). Andrich (1978) provided a mathematical demonstration of this cancellation using the BTL model under the assumption of discriminal processes with additive biases. However, it is easy to imagine at least two scenarios in which the zero correlation assumption is almost certainly invalid: when the pairwise comparison involves multidimensional, complex traits with heterogeneous stimuli and when an additional hierarchical structure is relevant to the stimuli.

In the first scenario, the intricate aspects of multidimensional, complex traits may introduce dependencies between the stimuli due to certain judges' biases that resist cancellation. Research on text quality suggests that when judges evaluate these traits, they often rely on various intricate characteristics of the stimuli to form their judgments (van Daal et al., 2016; Lesterhuis, 2018b; Chambers and Cunningham, 2022). These additional relevant characteristics, which are unlikely to be equally weighted or opposing, can exert an uneven influence on judges' perceptions, creating biases in their judgments and, ultimately, introducing dependencies between stimuli (van der Linden, 2017b, pp. 346). For example, this could occur when a judge assessing the argumentative quality of a text places more weight on its grammatical accuracy than other judges, thereby favoring texts with fewer errors but weaker arguments. While direct evidence for this particular scenario is lacking, studies such as Pollitt and Elliott (2003) demonstrate the presence of such biases, supporting the notion that the factors influencing pairwise comparisons may not always cancel out.

In the second scenario, the shared context or inherent connections created by additional hierarchical structures may further introduce dependencies between stimuli, a statistical phenomenon commonly known as clustering (Everitt and Skrondal, 2010). Despite the CJ literature acknowledging the existence of such hierarchical structures, the statistical handling of this additional source of dependence between stimuli has been inadequate. For instance, when CJ data incorporates multiple samples of stimuli from the same individuals, researchers frequently rely on (average) estimated BTL scores to conduct subsequent analyses and tests at the individual hierarchical level (Bramley and Vitello, 2019; Boonen et al., 2020; Bouwer et al., 2023; van Daal et al., 2017; Jones et al., 2019; Gijsen et al., 2021). However, this approach can introduce additional statistical and measurement issues, which we discuss in greater detail in Section 2.

In any case, similar to Section 1.1, model misspecification due to an erroneous assumption of zero correlation between stimuli can lead to significant statistical and measurement issues. For instance, the model may over- or underestimate how accurately the outcome reflects the "true" discriminal differences between stimuli. Such inaccuracies can result in spurious inferences about these differences and, by extension, about the stimuli's discriminal processes. This scenario is also illustrated by Figure 2b, when the model's discriminal difference distribution aligns with the thick continuous line for $\rho = 0$, while the "true" discriminal difference follows any discontinuous line where $\rho \neq 0$.

The misspecification may arise from neglecting additional relevant traits, excluding judges based on misfit statistics, or ignoring hierarchical (grouping) structures. Neglecting relevant traits, such as judges' biases, can cause dimensional mismatches in the BTL model, artificially inflating the trait's reliability (Hoyle, 2023, pp. 341) or, worse, introducing bias into the trait's estimates (Ackerman, 1989). Excluding judges based on misfit statistics risks discarding valuable information, which may further bias the trait's estimates (Zimmerman, 1994; McElreath, 2020, chap. 12). Finally, ignoring hierarchical structures may reduce the precision of model parameter estimates, potentially amplifying the overestimation of the trait's reliability (Hoyle, 2023, pp. 482).

2. The disconnect between trait measurement and hypothesis testing

Building on the previous section, it is clear that, despite its limitations, the BTL model is commonly used as a measurement model in CJ assessments. A measurement model specifies how manifest variables contribute to the estimation of latent variables (Everitt and Skrondal, 2010). For example, when evaluating writing quality, researchers use the BTL model to process the dichotomous outcomes resulting from the pairwise comparisons (the manifest variables) to estimate scores that reflect the underlying level of writing quality (the latent variable) (Laming, 2004; Pollitt, 2012b; Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018a; Coertjens et al., 2017; Goossens and De Maeyer, 2018; Bouwer et al., 2023).

Researchers then typically use these estimated BTL scores, or their transformations, to conduct additional analyses or hypothesis tests. For example, these scores have been used to identify 'misfit' judges and stimuli (Pollitt, 2012b; van Daal et al., 2016; Goossens and De Maeyer, 2018), detect biases in judges' ratings (Pollitt and Elliott, 2003; Pollitt, 2012b), calculate correlations with other assessment methods (Goossens and De Maeyer, 2018; Bouwer et al., 2023), or test hypotheses related to the underlying trait of interest (Bramley and Vitello, 2019; Boonen et al., 2020; Bouwer et al., 2023; van Daal et al., 2017; Jones et al., 2019; Gijsen et al., 2021).

However, the statistical literature advises caution when using estimated scores for additional analyses and tests. A key consideration is that BTL scores are parameter estimates that inherently carry uncertainty. Ignoring this uncertainty can bias the analysis and reduce the precision of hypothesis tests. Notably, the direction and magnitude of such biases are often unpredictable. Results may be attenuated, exaggerated, or remain unaffected depending on the degree of uncertainty in the scores and the actual effects being tested (Kline, 2023, pp. 25; Hoyle, 2023, pp. 137). Finally, the reduced precision in hypothesis tests diminishes their statistical power, increasing the likelihood of committing type-I or type-II errors (McElreath, 2020).

In aggregate, researchers' inadequate handling of violations to the assumptions of equal dispersion and zero correlation between stimuli, coupled with the apparent disconnect between CJ's approach to trait measurement and hypothesis testing, can potentially compromise the reliability of the trait estimates and, by extension, their validity (Perron and Gillespie, 2015, pp. 2). Consequently, adopting a more systematic and integrated approach to handling these assumptions and examining the factors influencing CJ experiments could offer several statistical and measurement benefits, including the ability to address these issues.

Updating the theoretical and statistical model

This section uses the structural approach to causal inference (Pearl, 2009; Pearl et al., 2016) to articulate a theoretical model that captures the core principles of Thurstone's theory. The model also incorporates various assessment design features relevant to CJ experiments, such as the selection of judges, stimuli, and comparisons. Finally, the section employs Bayesian inference methods to transform these theoretical and practical elements into a statistical model that facilitates the analysis of pairwise comparison data. See Section **??** for an overview of the statistical and causal inference concepts required for the development of this section.

3. The theoretical model

The theoretical model uses structural causal models (SCMs) and directed acyclic graphs (DAGs) (Pearl, 2009; Pearl et al., 2016; Gross et al., 2018; Neal, 2020) to formally and graphically represent the assumed causal structure of the CJ system. First, *the population model* is created to represent a conceptual population of CJ experiments. The model then integrates various assessment design features relevant to CJ experiments, leading to the development of *the sample-comparison model*.

3.1. The population model

Assuming population data or more commonly known as census data, we ...

The (latent) discriminal difference of the stimuli directly determines the (manifest) outcome of the pairwise comparisons

The (latent) "perceived" discriminal processes for the stimuli directly determines their discriminal difference

The (latent) "true" discriminal processes for the stimuli and the judges' biases directly determines their (latent) "perceived" discriminal processes

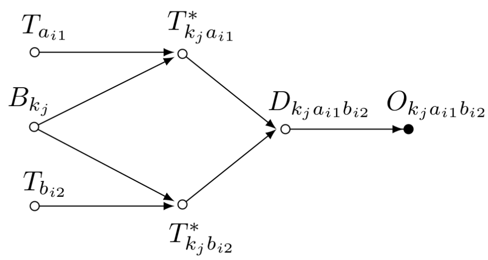


Figure 3

without loosing generality, the (latent) “perceived” and “true” discriminal processes for the stimuli can be depicted in a vector for each judge, as in

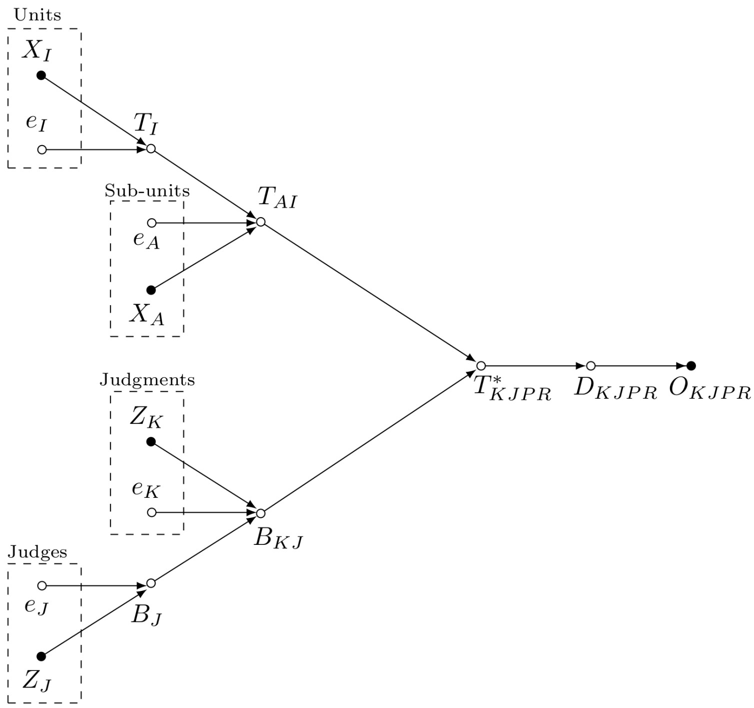


Figure 4

3.2. The sample-comparison model

Considering the sampling mechanism

Considering comparison mechanisms

4. From theory to statistics

Bayesian inference procedures offer three key advantages. First, they are well-suited to handling complex and overparameterized models, enabling researchers to estimate models where the number of parameters exceeds the number of observations for estimation (Baker, 1998; Kim and Cohen, 1999). Second, they allow researchers to incorporate prior information, which helps constrain parameters within specified bounds. This capability addresses challenges such as non-convergence or improper parameter estimation that often arise in complex models when analyzed with frequentist methods (Martin and McDonald, 1975; Seaman III et al., 2011). Finally, Bayesian methods are particularly effective at drawing inferences from small sample sizes, where relying on the asymptotic properties of frequentist approaches may not be justified (Baldwin and Fellingham, 2013; Lambert et al., 2006; Depaoli, 2014).

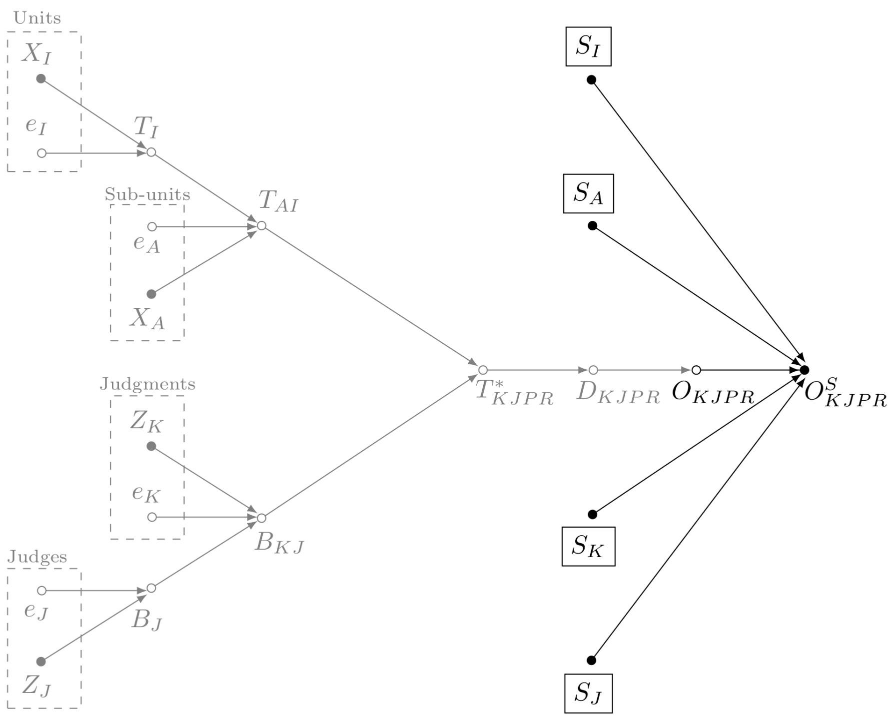


Figure 5

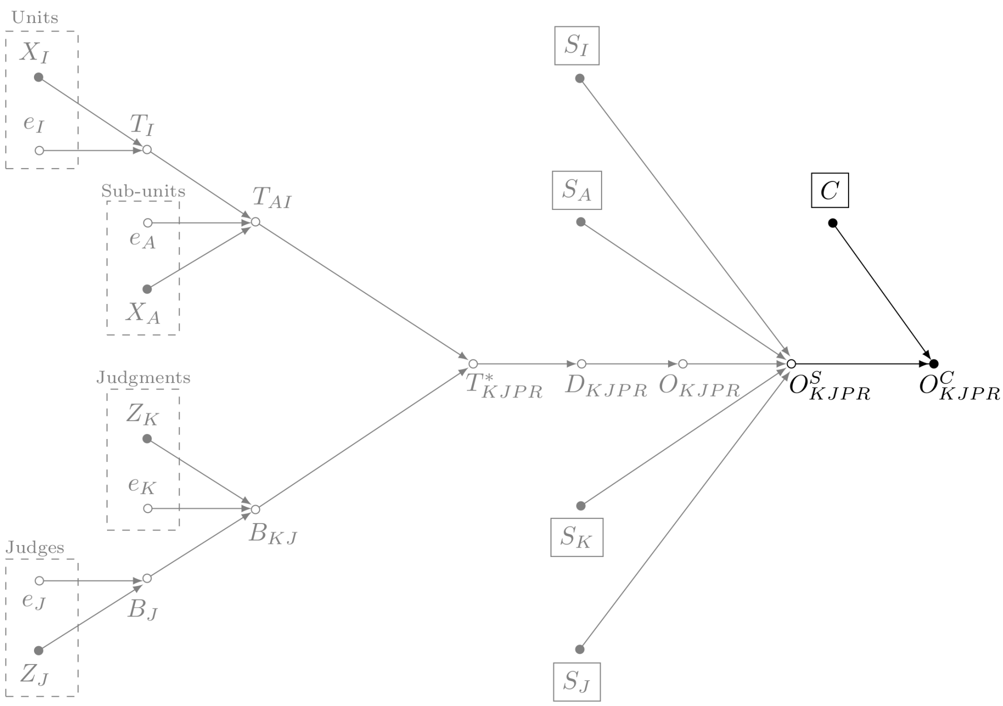


Figure 6

Discussion

5. Findings

6. Limitations and further research

Conclusion

Declarations

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

Financial interests: The authors have no relevant financial interest to disclose.

Non-financial interests: The authors have no relevant non-financial interest to disclose.

Ethics approval: The University of Antwerp Research Ethics Committee has confirmed that no ethical approval is required.

Consent to participate: Not applicable

Consent for publication: All authors have read and agreed to the published version of the manuscript.

Availability of data and materials: No data was utilized in this study.

Code availability: All the code utilized in this research is available in the digital document located at: https://jriveraespejo.github.io/paper2_manuscript/.

AI-assisted technologies in the writing process: The authors utilized a range of AI-based language tools throughout the preparation of this work. They occasionally employed the tools to refine phrasing and optimize wording, ensuring appropriate language use and enhancing the manuscript's clarity and coherence. The authors take full responsibility for the final content of the publication.

CRediT authorship contribution statement: *Conceptualization:* S.G., S.DM., T.vD., and J.M.R.E; *Methodology:* S.DM., T.vD., and J.M.R.E; *Software:* J.M.R.E.; *Validation:* J.M.R.E.; *Formal Analysis:* J.M.R.E.; *Investigation:* J.M.R.E; *Resources:* S.G., S.DM., and T.vD.; *Data curation:* J.M.R.E.; *Writing - original draft:* J.M.R.E.; *Writing - review and editing:* S.G., S.DM., and T.vD.; *Visualization:* J.M.R.E.; *Supervision:* S.G. and S.DM.; *Project administration:* S.G. and S.DM.; *Funding acquisition:* S.G. and S.DM.

Appendix

This section introduces fundamental statistical and causal inference concepts necessary for understanding the core theoretical principles described in Section **??**. It does not, however, offer a comprehensive overview of causal inference methods. Readers seeking more in-depth understanding may wish to explore introductory papers such as Pearl (2010), Rohrer (2018), Pearl (2019), and Cinelli et al. (2020). They may also find it helpful to consult introductory books like Pearl and Mackenzie (2018), Neal (2020), and McElreath (2020). For more advanced study, readers may refer to seminal intermediate papers such as 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).

7. Empirical research and randomized experiments

Empirical research uses evidence from observation and experimentation to address real-world challenges. In this context, researchers typically formulate their research questions as *estimands* or *targets of inference*, i.e., the specific quantities they seek to determine (Everitt and Skrondal, 2010). For instance, researchers might be interested in answering the following question: "To what extent do different teaching methods (T) influence students' ability to produce high-quality written texts (Y)?" To investigate this, researchers could randomly assign students to two groups, each exposed to a different teaching method ($T_i = \{1, 2\}$). Then, they would perform pairwise comparisons, generating a dichotomous outcome ($Y_i = \{0, 1\}$) showing which student exhibits more of the ability. In this scenario, the research question can be rephrased as the estimand, "*On average*, is there a difference in the ability to produce high-quality written texts between the two groups of students?" and this estimand can be mathematically represented by the random quantity $E[Y_i|T_i = 1] - E[Y_i|T_i = 2]$, where $E[\cdot]$ denotes the expected value.

Researchers would then proceed to identify the estimands. *Identification* refers to the process of accurately computing an estimand using an estimator. An *estimator* is a method or function that transforms data into an estimate (Neal, 2020). *Estimates* are numerical values that approximate the estimand and are derived through *estimation*, which refers to the process of integrating data with an estimator (Everitt and Skrondal, 2010). The Identification-Estimation flowchart (McElreath, 2020; Neal, 2020) in Figure 7 provides a visual representation of the process of transitioning from estimands to estimates.

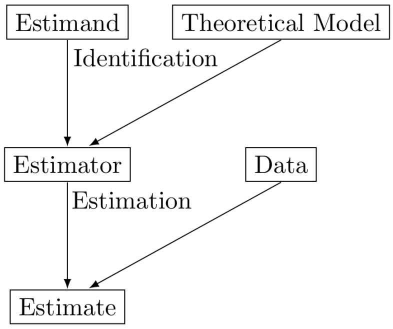


Figure 7: Identification-Estimation flowchart. Extracted and slightly modified from Neal (2020, pp. 32)

While numerous methods can approximate an estimand, researchers prioritize estimators with desirable properties that ensure the accuracy of estimates. For instance, the Z-test is an estimator known for its effectiveness in comparing groups' proportions, yielding accurate estimates when the underlying assumptions of the statistic are met (Kanji, 2006). If this is the case, the Z-test is expressed as a signal-to-noise statistic $Z = (\hat{p}_1 - \hat{p}_2)/\hat{s}_p$. The signal is defined as the difference between the groups' sample proportions, $\hat{p}_1 = \sum_{i=1}^{n_1} Y_i/n_1$ and $\hat{p}_2 = \sum_{i=1}^{n_2} Y_i/n_2$, analogous to $E[Y_i|T_i = 1]$ and $E[Y_i|T_i = 2]$, respectively. The noise, represented by $\hat{s}_p$, is defined as the unpooled sample variability observed between the two groups.

However, researchers often seek to uncover the mechanisms underlying specific data and establish causal relationships rather than simply estimate associations. In the example, researchers can interpret the associational estimate represented by the Z-statistic as causal. This interpretation relies on the data meeting the assumptions of the Z-test and being collected through a randomized experiment.

Randomized experiments are widely recognized as the gold standard in evidence-based science (Hariton and Locascio, 2018; Hansson, 2014). This recognition stems from their ability to enable researchers to interpret associational estimates as causal. They achieve this by ensuring data, and by extension an estimator, satisfies several key properties, such as common support, no interference, and consistency (Morgan and Winship, 2014; Neal, 2020). The most critical property, however, is the elimination of confounding. *Confounding* occurs when an external variable X simultaneously influences the outcome Y and the variable of interest T, resulting in spurious associations (Everitt and Skrondal, 2010). Randomization addresses this issue by decoupling the association between the intervention allocation T from any other variable X (Morgan and Winship, 2014; Neal, 2020).

Nevertheless, researchers often face constraints that limit their ability to conduct randomized experiments. These constraints include ethical concerns, such as the assignment of individuals to potentially harmful interventions, and practical limitations, such as the infeasibility of, for example, assigning individuals to genetic modifications or physical impairments (Neal, 2020). In these cases, causal inference offers a valuable alternative for generating causal estimates and understanding the mechanisms underlying specific data. In addition, the framework can provide significant theoretical insights that can enhance the design of experimental and observational studies (McElreath, 2020).

8. Identification under causal inference

Unlike classical statistical modeling, which focuses primarily on summarizing data and inferring associations, the *causal inference* framework is designed to identify causes and estimate their effects using data (Shaughnessy et al., 2010; Neal, 2020). The framework uses rigorous mathematical techniques to address the *fundamental problem of causality* (Pearl, 2009; Pearl et al., 2016; Morgan and Winship, 2014). This problem revolves around the question, "What would have happened 'in the world' under different circumstances?" This question introduces the concept of counterfactuals, which are instrumental in defining and identifying causal effects.

Counterfactuals are hypothetical scenarios that are *contrary to fact*, where alternative outcomes resulting from a given cause are neither observed nor observable (Neal, 2020; Counterfactual, 2024). The structural approach to causal inference (Pearl, 2009; Pearl et al., 2016) provides a formal framework for defining counterfactuals. For instance, in the scenario described in Section 7, the approach begins by defining the *individual causal effect* (ICE) as the difference between each student's potential outcomes: $\tau_i = Y_i | do(T_i = 1) - Y_i | do(T_i = 2)$. Here, $do(T_i = t)$ represents the intervention operator, $Y_i | do(T_i = 1)$ represents the potential outcome under intervention $T_i = 1$, and $Y_i | do(T_i = 1)$ represents the potential outcome under intervention $T_i = 2$. Note that if a student is assigned to intervention $T_i = 1$, the potential outcome under $T_i = 2$ becomes a counterfactual, as it is no longer observed nor observable. To address the challenge of unobserved counterfactuals, the approach extends the ICE to the *average causal effect* (ACE): $\tau = E[\tau_i] = E[Y_i | do(T_i = 1)] - E[Y_i | do(T_i = 2)]$, representing the average difference between observed potential outcomes and their counterfactual counterparts.

Even when data originates from an observational study, researchers can identify the ACE from associational estimates using the structural approach. They achieve this by performing statistical conditioning on a *sufficient adjustment set* of variables X (Pearl, 2009; Pearl et al., 2016; Morgan and Winship, 2014). This *sufficient* set (potentially empty) must block all non-causal paths between

T to Y without opening new ones, ensuring the ACE estimator satisfies several key properties, including confounding elimination. If such a set exists, then T and Y are *d-separated* by X (Pearl, 2009), meaning researchers can estimate the ACE from associational random quantities (Morgan and Winship, 2014). Naturally, the validity of claims about the effects of T on Y now hinges on the assumption that X serves as a sufficient adjustment set. However, as Kohler et al. (2019, pp. 150) observed, drawing conclusions about the real world from observed data inevitably requires assumptions. This holds true for both observational and experimental data.

For instance, if researchers are unable to conduct the randomized experiments described in Section 7 and instead rely on observational data, they can still estimate the ACE provided a variable X, such as the socio-economic status of the school, blocks all non-causal paths between the teaching method T to the outcome Y without opening any new ones. Under these conditions, the ACE can be estimated from associational quantities as $\tau = E[Y_i|do(T_i = 1)] - E[Y_i|do(T_i = 2)] = E_X\left[E[Y_i|T_i = 1, X] - E[Y_i|T_i = 2, X]\right]$, where $E_X[\cdot]$ represents the marginal expected value over X (Morgan and Winship, 2014). Notably, the approach extends the ACE for a continuous variable T as $\tau = E[Y_i|do(T_i = t)] = dE_X\left[E[Y_i|T_i = t, X]\right]/dt$, ensuring broad applicability across different causal scenarios (Neal, 2020, pp. 45).

9. SCMs and DAGs

The structural approach to causal inference uses SCMs and DAGs to formally and graphically represent the presumed causal structure underlying the ACE (Pearl, 2009; Pearl et al., 2016; Gross et al., 2018; Neal, 2020). In essence, SCMs and DAGs act as *conceptual models* that guide researchers in determining which statistical models can yield valid causal inferences, assuming the depicted causal structure of the models is correct (McElreath, 2020). Notably, every SCM has an associated DAG (Cinelli et al., 2020). Figure 7 provides a visual representation of the role of theoretical models in the inference process.

SCMs and DAGs offer two key advantages for modeling causal structures. First, they enable the representation of causal relations in a non-parametric and fully interactive manner. This feature allows for feasible ACE identification strategies without specifying the data type or the nature of the functional dependence among variables (Morgan and Winship, 2014). Second, regardless of complexity, they can represent a wide range of causal structures using only five fundamental building blocks (Neal, 2020; McElreath, 2020). This feature allows researchers to decompose complex structures, facilitating their analysis by focusing on the causal assumptions associated with each building block (McElreath, 2020).

Figures 8, 9, 10, 11, and 12 display these five fundamental building blocks. The left panels of the figures show the formal mathematical models, represented by the SCMs, defined in terms of a set of *endogenous* variables $V = \{X_1, X_2, X_3\}$, a set of *exogenous* variables $E = \{e_{X1}, e_{X2}, e_{X3}\}$, and a set of functions $F = \{f_{X1}, f_{X2}, f_{X3}\}$ (Pearl, 2009; Cinelli et al., 2020; Neal, 2020). Endogenous variables are those whose causal mechanisms a researcher 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, pp. 27,68). Lastly, the functions, referred to as *structural equations*, express the endogenous variables as non-parametric functions of other variables. These functions use the symbol ':=' to denote the asymmetrical causal dependence of the variables and the symbol '$\perp\!\!\!\perp$' to represent *d-separation*, a concept akin to (conditional) independence.

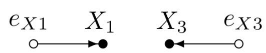
The right panels of the figures display the complementary DAGs. A DAG consists of nodes connected by edges, where nodes represent random variables. The term *directed* means the edges extend from one node to another, with arrows indicating the direction of causal influence. The term *acyclic* signifies that the causal influences do not form loops, ensuring the influences do not cycle back on themselves (McElreath, 2020). DAGs represent observed variables as solid black circles, while they use open circles for unobserved (latent) variables (Morgan and Winship, 2014). Finally, the arrows in the graphs show the expected direction of causal influences among these variables. While DAGs often omit exogenous variables for simplicity (the *standard* representation), including these variables is advantageous (the *magnified* representation shown in the figures), as their inclusion helps to identify potential issues related to conditioning and confounding (Cinelli et al., 2020).

$$X_1 := f_{X1}(e_{X1})$$

$$X_3 := f_{X3}(e_{X3})$$

$$e_{X1} \perp\!\!\!\perp e_{X3}$$

(a) SCM



(b) DAG

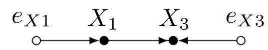
Figure 8: Two unconnected nodes

$$X_1 := f_{X1}(e_{X1})$$

$$X_3 := f_{X3}(X_1, e_{X3})$$

$$e_{X1} \perp\!\!\!\perp e_{X3}$$

(a) SCM



(b) DAG

Figure 9: Two connected nodes or descendant

$$X_1 := f_{X1}(e_{X1})$$

$$X_2 := f_{X2}(X_1, e_{X2})$$

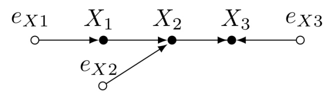
$$X_3 := f_{X3}(X_2, e_{X3})$$

$$e_{X1} \perp\!\!\!\perp e_{X2}$$

$$e_{X1} \perp\!\!\!\perp e_{X3}$$

$$e_{X2} \perp\!\!\!\perp e_{X3}$$

(a) SCM



(b) DAG

Figure 10: Chain or mediator

$$X_1 := f_{X1}(X_2, e_{X1})$$

$$X_2 := f_{X2}(e_{X2})$$

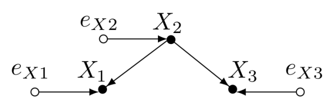
$$X_3 := f_{X3}(X_2, e_{X3})$$

$$e_{X1} \perp\!\!\!\perp e_{X2}$$

$$e_{X1} \perp\!\!\!\perp e_{X3}$$

$$e_{X2} \perp\!\!\!\perp e_{X3}$$

(a) SCM



(b) DAG

Figure 11: Fork or confounder

$$X_1 := f_{X1}(e_{X1})$$

$$X_2 := f_{X2}(X_1, X_3, e_{X2})$$

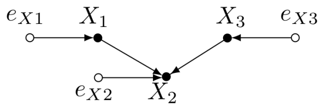
$$X_3 := f_{X3}(e_{X3})$$

$$e_{X1} \perp\!\!\!\perp e_{X2}$$

$$e_{X1} \perp\!\!\!\perp e_{X3}$$

$$e_{X2} \perp\!\!\!\perp e_{X3}$$

(a) SCM



(b) DAG

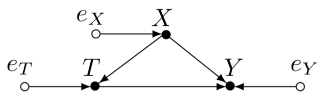
Figure 12: Collider or inmorality

A careful examination of the figures highlights the assumptions underlying these building blocks. Figures 8a and 8b depict two unconnected nodes, representing a scenario where variables X_1 and X_3 are not causally related. Figures 9a and 9b illustrate two connected nodes, representing a scenario where a *parent* node X_1 exerts a causal influence on a *child* node X_3. In this setup, X_3 is considered a *descendant* of X_1. Additionally, X_1 and X_3 are described as *adjacent* because an edge directly connects them. Figures 10a and 10b depict a *chain*, where X_1 influences X_2, and X_2 influences X_3. In this configuration, X_1 is a parent node of X_2, which is a parent node of X_3. This structure creates a *directed path* between X_1 and X_3. Consequently, X_1 is an *ancestor* of X_3, and X_2 fully *mediates* the relationship between the two. Figures 11a and 11b illustrate a *fork*, where variables X_1 and X_3 are both influenced by X_2. Here, X_2 is a parent node that *confounds* the relationship between X_1 and X_3. Finally, Figures 12a and 12b show a *collider*, where variables X_1 and X_3 are concurrent causes of X_2. In this configuration, X_1 and X_3 are not causally related to each other but both influence X_2 (an *inmorality*). Notably, in all SCMs, the errors are assumed to be independent of each other and from all other variables in the graph, as evidenced by the pairwise relations $e_{X1} \perp\!\!\!\perp e_{X2}$, $e_{X1} \perp\!\!\!\perp e_{X3}$, and $e_{X2} \perp\!\!\!\perp e_{X3}$.

Researchers can use these building blocks to represent the scenario outlined in Section 8. Figures 13a and 13b depict the plausible causal structure for this example, assuming that the variable X (socio-economic status of the school) acts as a confounder in the relationship between the teaching method T and the outcome Y. The figures show several descendant, such as $X \to T$, $X \to Y$, and $T \to Y$, and also highlight multiple pairs of unconnected nodes, evident from the relationships $e_T \perp\!\!\!\perp e_X$, $e_T \perp\!\!\!\perp e_Y$, and $e_X \perp\!\!\!\perp e_Y$. Additional, the figures depict one fork, $X \to \{T, Y\}$, and two colliders: $\{X, e_T\} \rightarrow T$ and $\{X, T, e_Y\} \rightarrow Y$.

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

(a) SCM



(b) DAG

Figure 13: Plausible causal structure the scenario outlined in Section 8.

10. The probabilistic implications of DAGs

After completing the identification process outlined in Section 8, researchers estimate the ACE using Bayesian inference methods. Bayesian inference calculates the probability of a set of parameters θ, enabling researchers to use their distributions to represent the ACE. The approach begins by defining two probability distributions: the likelihood of the data, $P(X_1, X_2, \dots, X_n|\theta)$, and the prior distribution, $P(\theta)$ (Everitt and Skrondal, 2010). Here, X_n represents a random variable, and for simplicity, θ denotes a parameter space of dimension one. After observing empirical data, researchers update these priors to posterior distributions using Bayes' rule (Jeffreys, 1998):

$$P(\theta|X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n|\theta) \cdot P(\theta)}{P(X_1, X_2, \dots, X_n)} \tag{1}$$

Researchers can further simplify the posterior updating process in Equation 1 into two steps: integrating new empirical data, which is defined by the likelihood, with the parameter update, which is determined by the priors, as demonstrated in equation Equation 2. This simplification is possible because the denominator on the right-hand side of Equation 1 acts as a normalizing constant, independent of θ.

$$P(\theta|X_1, X_2, \dots, X_n) \propto P(X_1, X_2, \dots, X_n|\theta) \cdot P(\theta) \tag{2}$$

Temporarily setting aside the definition of prior distributions, it is important to note that the posterior updating process depends on the assumptions underlying the likelihood of the data. However, as the number of random variables, n, increases, this joint distribution becomes quickly intractable (Neal, 2020). This is evident from Equation 3, where the likelihood distribution is re-expressed in terms of chained conditional distributions.

$$P(X_1, X_2, \dots, X_n|\theta) = P(X_1|\theta) \prod_{i=2}^{n} P(X_i|X_{i-1}, \dots, X_1, \theta) \tag{3}$$

To address the complexity of the likelihood, researchers simplify the distribution by assuming that variables exhibit specific local (in)dependencies. These assumptions improve model tractability and streamline the estimation process. Researchers formalize these local (in)dependencies using directed acyclic graphs (DAGs). DAGs represent the probabilistic implications of these assumptions by relying on three key principles: the local Markov assumption, the minimality assumption, and the causal edges assumption. The *local Markov assumption* defines the statistical independencies implied by a DAG. It states that a variable is independent of all its non-descendants, given its parents (Neal, 2020). The *minimality assumption* identifies the statistical dependencies implied by a DAG. It asserts that every pair of adjacent variables are dependent (Neal, 2020). Finally, the *causal edges assumption* establishes causal relationships between parents and their children in a DAG by stating that every parent is a direct cause of their children (Neal, 2020). Figure Figure 14 illustrates the flow of association and causation in graphs.

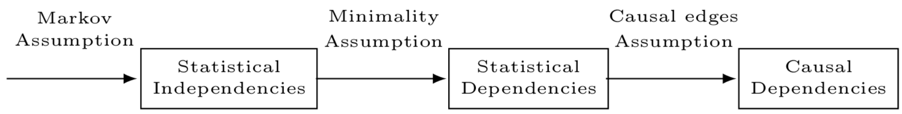


Figure 14: The flow of association and causation in graphs. Extracted and slightly modified from Neal (2020, pp. 31)

This lead to the *Bayesian Network factorization* of Equation 3.

$$P(X_1, X_2, \dots, X_n|\theta) = P(X_1|\theta) \prod_{i=2}^{n} P(X_i|pa(X_i), \theta) \tag{4}$$

References

Ackerman, T., 1989. Unidimensional irt calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement 13, 113–127. doi:10.1177/014662168901300201.

Andrich, D., 1978. Relationships between the thurstone and rasch approaches to item scaling. Applied Psychological Measurement 2, 451–462. doi:10.1177/014662167800200319.

Baker, F., 1998. An investigation of the item parameter recovery characteristics of a gibbs sampling procedure. Applied Psychological Measurement 22, 153–169. doi:10.1177/01466216980222005.

Baldwin, S., Fellingham, G., 2013. Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Journal of Psychological Methods 18, 151–164. doi:10.1037/a0030642.

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. 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. doi:10.2307/2334029.

Bramley, T., 2008. Paired comparison methods, in: Newton, P., Baird, J., Goldsteing, H., Patrick, H., Tymms, P. (Eds.), Techniques for monitoring the comparability of examination standards. GOV.UK., pp. 246—300. URL: https://assets.publishing.service.gov.uk/media/5a80d75940f0b62305b8d734/2007-comparability-exam-standards-i-chapter7.pdf.

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.

Chambers, L., Cunningham, E., 2022. Exploring the validity of comparative judgement: Do judges attend to construct-irrelevant features? Frontiers in Education doi:10.3389/feduc.2022.802392.

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.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., De Maeyer, S., 2017. Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering. Pedagogische Studien 94, 283–303. URL: https://repository.uantwerpen.be/docman/irua/e71ea9/147930.pdf.

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

Crompvoets, E., Béguin, A., Sijtsma, K., 2022. On the bias and stability of the results of comparative judgment. Frontiers in Education 6. doi:10.3389/feduc.2021.788202.

Depaoli, S., 2014. The impact of inaccurate "informative" priors for growth parameters in bayesian growth mixture modeling. Journal of Structural Equation Modeling 21, 239–252. doi:10.1080/10705511.2014.882686.

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

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. doi:10.3389/feduc.2020.582800.

Goossens, M., De Maeyer, S., 2018. How to obtain efficient high reliabilities in assessing texts: Rubrics vs comparative judgement, in: Ras, E., Guerrero Roldán, A. (Eds.), Technology Enhanced Assessment, Springer International Publishing. pp. 13–25. doi:10.1007/978-3-319-97807-9_2.

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.

Hansson, S., 2014. Why and for what are clinical trials the gold standard? Scandinavian Journal of Public Health 42, 41–48. doi:10.1177/1403494813516712. pMID: 24553853.

Hariton, E., Locascio, J., 2018. Randomised controlled trials – the gold standard for effectiveness research. BJOG: An International Journal of Obstetrics & Gynaecology 125, 1716–1716. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.15199, doi:10.1111/1471-0528.15199.

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.

Hoyle, R.e., 2023. Handbook of Structural Equation Modeling. Guilford Press.

Jeffreys, H., 1998. Theory of probability. Oxford University Press.

Jones, I., Bisson, M., Gilmore, C., Inglis, M., 2019. Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? British Educational Research Journal 45, 662–680. doi:10.1002/berj.3519.

Jones, I., Inglis, M., 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.

Kanji, G., 2006. 100 Statistical Tests. Introduction to statistics, SAGE Publications.

Kelly, K., Richardson, M., Isaacs, T., 2022. Critiquing the rationales for using comparative judgement: a call for clarity. Assessment in Education: Principles, Policy & Practice 29, 674–688. doi:10.1080/0969594X.2022.2147901.

Kim, S., Cohen, A., 1999. Accuracy of parameter estimation in gibbs sampling under the two-parameter logistic model. URL: https://eric.ed.gov/?id=ED430012. annual Meeting of the American Educational Research Association.

Kimbell, R., 2012. Evolving project e-scape for national assessment. International Journal of Technology and Design Education 22, 135–155. doi:10.1007/s10798-011-9190-4.

Kline, R., 2023. Principles and Practice of Structural Equation Modeling. Methodology in the Social Sciences, Guilford Press.

Kohler, U., Kreuter, F., Stuart, E., 2019. Nonprobability sampling and causal analysis. Annual Review of Statistics and Its Application 6, 149–172. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-030718-104951, doi:https://doi.org/10.1146/annurev-statistics-030718-104951.

Lambert, P., Sutton, A., Burton, P., Abrams, K., Jones, D., 2006. How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs. Journal of Statistics in Medicine 24, 2401–2428. doi:10.1002/sim.2112.

Laming, D., 2004. Marking university examinations: Some lessons from psychophysics. Psychology Learning & Teaching 3, 89–96. doi:10.2304/plat.2003.3.2.89.

Lesterhuis, M., 2018a. The validity of comparative judgement for assessing text quality: An assessor's perspective. Ph.D. thesis. University of Antwerp. URL: https://hdl.handle.net/10067/1548280151162165141.

Lesterhuis, M., 2018b. When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality. L1-Educational Studies in Language and Literature 18, 1–22. doi:10.17239/L1ESLL-2018.18.01.02.

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.

Martin, J., McDonald, R., 1975. Bayesian estimation in unrestricted factor analysis: A treatment for heywood cases. Psychometrika , 505–517doi:10.1007/BF02291552.

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

Mikhailiuk, A., Wilmot, C., Perez-Ortiz, M., Yue, D., Mantiuk, R., 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization, in: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2559–2566. doi:10.1109/ICPR48806.2021.9412676.

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.

Perron, B., Gillespie, D., 2015. Reliability and Measurement Error, in: Key Concepts in Measurement. Oxford University Press. Pocket guides to social work research methods. chapter 4. doi:10.1093/acprof:oso/9780199855483.003.0004.

Pollitt, A., 2004. Let's stop marking exams, in: Proceedings of the IAEA Conference, University of Cambridge Local Examinations Syndicate, Philadelphia. URL: https://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf.

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.

Pollitt, A., Elliott, G., 2003. Finding a proper role for human judgement in the examination system. URL: https://www.cambridgeassessment.org.uk/Images/109707-monitoring-and-investigatingcomparability-a-proper-role-for-human-judgement.pdf. research & Evaluation Division.

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.

Seaman III, J., Seaman Jr., J., Stamey, J., 2011. Hidden dangers of specifying noninformative priors. The American Statistician 66, 77–84. doi:10.1080/00031305.2012.695938.

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

Thurstone, L., 1927b. Psychophysical analysis. American Journal of Psychology , 368–89URL: https://brocku.ca/MeadProject/Thurstone/Thurstone_1927g.html. last accessed 20 december 2024.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., De Maeyer, S., 2016. Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. Assessment in Education: Principles, Policy & Practice 26, 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. doi:10.3389/feduc.2017.00044.

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

van der Linden, W. (Ed.), 2017b. Handbook of Item Response Theory: Statistical Tools. volume 2 of *Statistics in the Social and Behavioral Sciences Series*. CRC Press.

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.

Verhavert, S., Furlong, A., Bouwer, R., 2022. The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. Frontiers in Education 6. doi:10.3389/feduc.2021.785919.

Whitehouse, C., 2012. Testing the validity of judgements about geography essays using the adaptive comparative judgement method. URL: https://filestore.aqa.org.uk/content/research/CERP_RP_CW_24102012_0.pdf?download=1. aQA Education.

Zimmerman, D., 1994. A note on the influence of outliers on parametric and nonparametric tests. The Journal of General Psychology 121, 391–401. doi:10.1080/00221309.1994.9921213.