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

1. 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 accurately the traits under measurement (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).

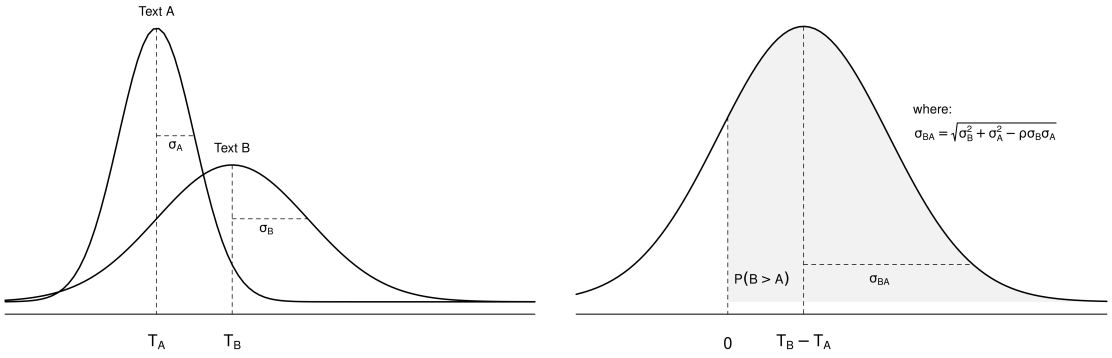
Nevertheless, despite the growing number of CJ studies, unsystematic and fragmented research approaches have left several critical issues unresolved. This study addresses two such issues that can compromise the reliability and validity of CJ's trait estimates. The first concerns the overreliance on Thurstone's Case V assumptions in the statistical analysis of CJ data. The second focuses on the apparent disconnect between CJ's approach to trait measurement and hypothesis testing.

The study is organized into six main sections. It begins with an overview of Thurstone's theory in Section 2, followed by a detailed examination of these issues in Section 3. In Section 4, the study develops a theoretical model for CJ that simultaneously addresses these issues. The model integrates Thurstone's core principles with key assessment design features relevant to CJ experiments, such as the selection of judges, stimuli, and comparisons. Section 5 translates these theoretical and practical elements into a statistical model that supports the analysis of pairwise comparison data. Finally, Section 6 discusses the concepts developed in the study, their implications, and avenues for future exploration, while Section 7 summarizes the findings.

2. 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 the 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, since 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.

3. The two critical issues in CJ literature

This section explores the two critical issues in CJ's theory that have the potential to compromise the reliability and validity of CJ's trait estimates. Section 3.1 inspects the overreliance on Thurstone's Case V assumptions in the statistical analysis of CJ data. Section 3.2 focuses on the apparent disconnect between CJ's approach to trait measurement and hypothesis testing.

3.1. 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 2, 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).

Table 1: Thurstones cases and their asumptions

Assumption	General form	Thurstone's					BTL model
		Case I	Case II	Case III	Case IV	Case V	
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

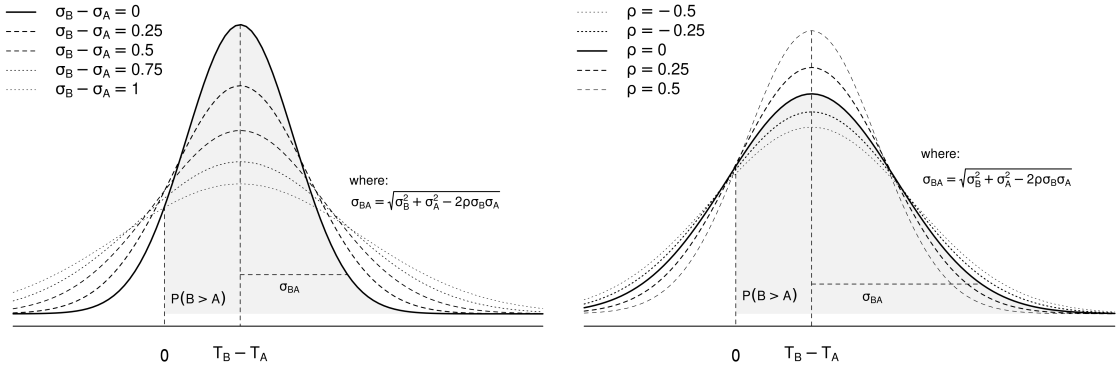
Notably, despite relying on the most extensive set of simplifying assumptions (Bramley, 2008, pp. 253; Kelly et al., 2022, pp. 677), Case V remains the most widely used case in the CJ literature. This popularity stems mainly from its simplified statistical representation in the Bradley-Terry-Luce (BTL) 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 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.

3.1.1. 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. A thought experiment can illustrate this idea. In this experiment, the researcher can observe the discriminal processes for the texts depicted in Figure 1a. Furthermore, the experiment assumes

that the discriminal dispersion for Text A remains constant and that the texts are uncorrelated ($\rho = 0$). In this scenario, Figure 2a 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, the thought experiment 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, when 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 the discriminal difference then 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 3.1.2.

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, introducing 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.

3.1.2. 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. A similar thought experiment, as the

one in Section 3.1.1, can illustrate this idea. The experiment now assumes that the discriminal dispersions for both texts remain constant. Figure 2b 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 lack of correlation (independence) between 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, creating a statistical phenomenon known as clustering (Everitt and Skrondal, 2010). Although the CJ literature acknowledges 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). This approach, however, can introduce additional statistical and measurement issues, which we discuss in greater detail in Section 3.2.

In any case, similar to Section 3.1.1, 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 model 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 overestimating the trait's reliability (Hoyle, 2023, pp. 482).

3.2. 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 the aforementioned issues.

4. Updating the theoretical model

4.1. 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

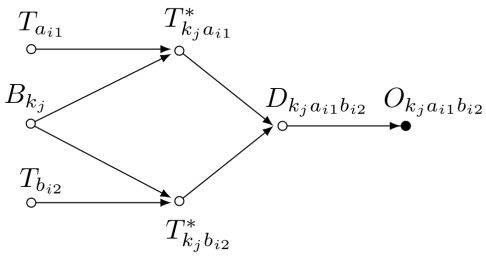


Figure 3

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

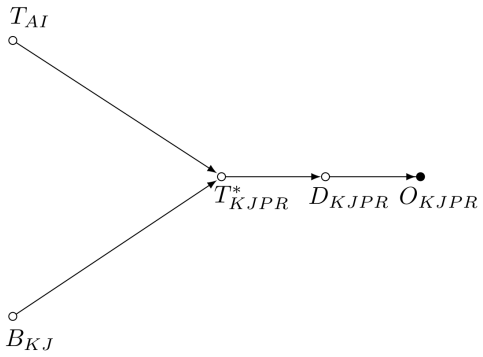


Figure 4

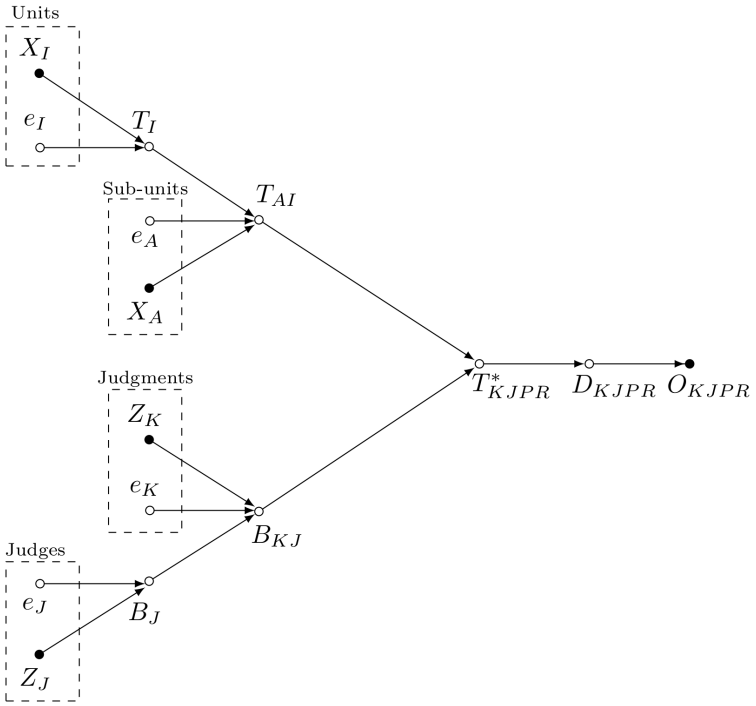


Figure 5

4.2. The sample-comparison model

Considering the sampling mechanism

Considering comparison mechanisms

5. From theory to statistics

6. Discussion

6.1. Findings

6.2. Limitations and further research

7. Conclusion

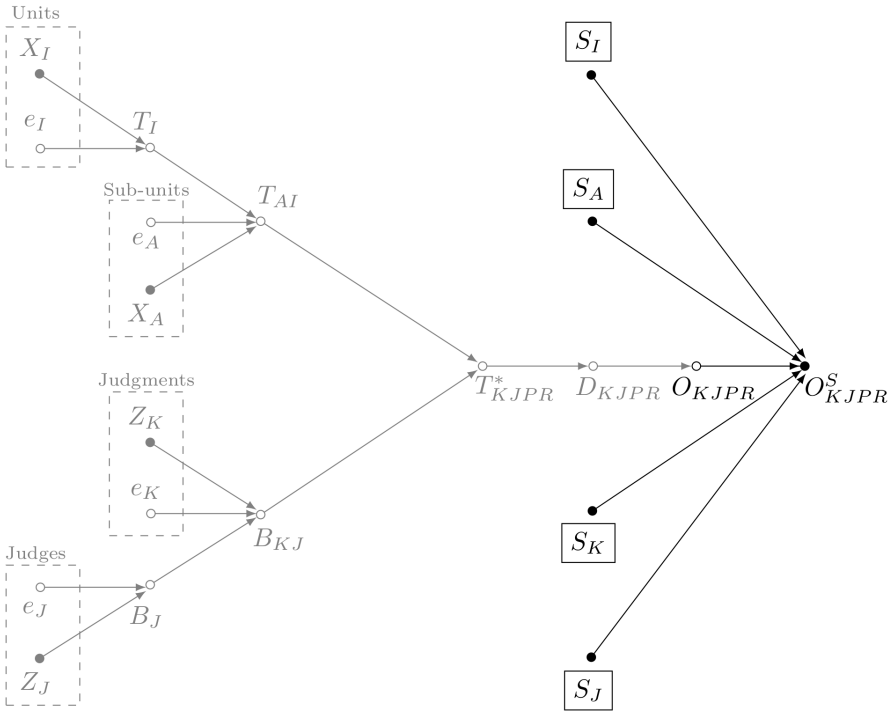


Figure 6

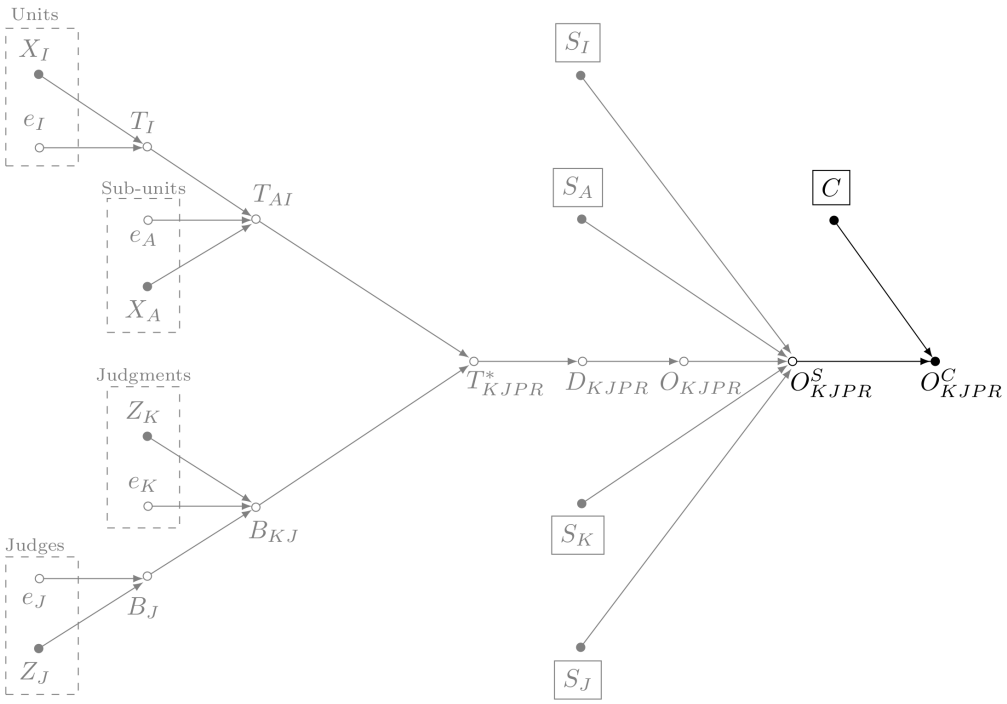


Figure 7

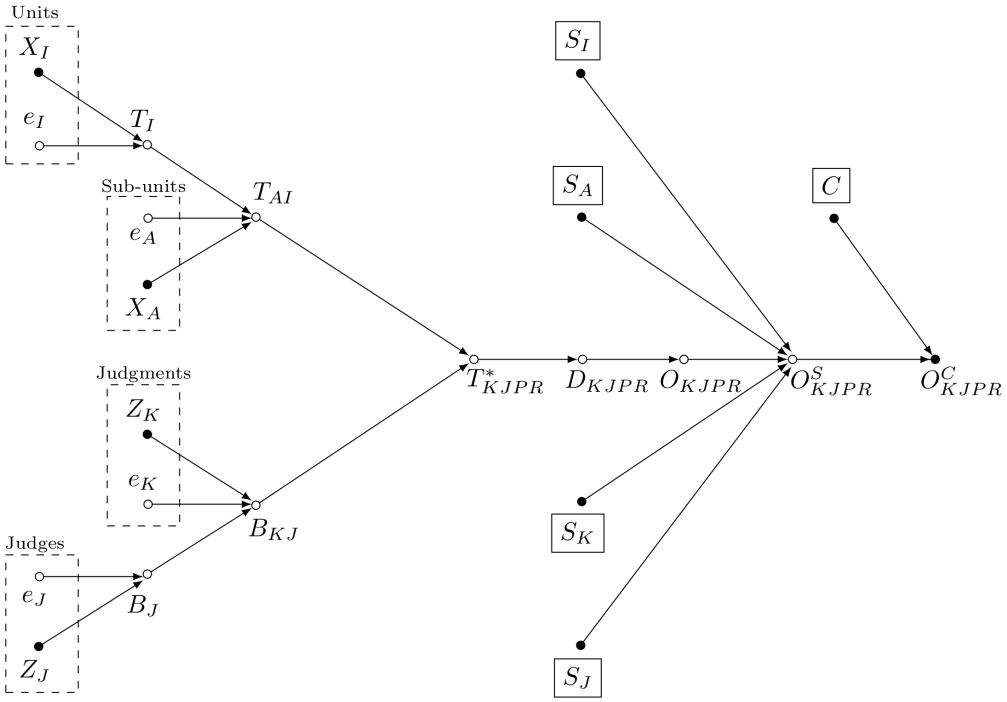


Figure 8

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.

8. Appendix

This section introduces fundamental statistical and causal inference concepts necessary for understanding the core theoretical principles described in this document. 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).

8.1. 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 associational quantity in Equation 1, where $E[\cdot]$ denotes the expected value.

$$E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 2] \tag{1}$$

Ideally, researchers then proceed to identify the estimands. *Identification* determines whether an estimator can accurately compute the estimand based solely on its assumptions, regardless of random variability (Schuessler and Selb, 2023, pp. 4). Identification is crucial because if an estimator cannot reliably compute an estimand using "infinite" and error-free data, it is impossible to gain meaningful insights about that estimand using finite data (Schuessler and Selb, 2023, pp. 5). An *estimator* refers to a method or function that transforms data into an estimate (Neal, 2020). *Estimates* are

numerical values that approximate the estimand derived through the process of *estimation*, which integrates data with an estimator (Everitt and Skrondal, 2010). The Identification-Estimation flowchart (McElreath, 2020; Neal, 2020), shown in Figure 9, visually represents the transition from estimands to estimates.

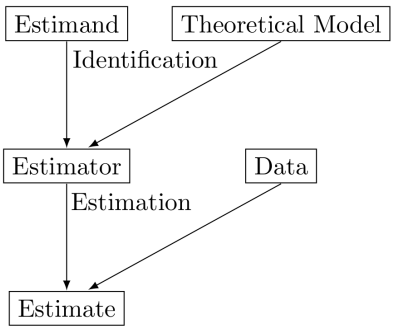


Figure 9: Identification-Estimation flowchart. Extracted and slightly modified from Neal (2020, pp. 32)

Consequently, although many estimators can approximate an estimand, researchers prioritize those with desirable identification properties to ensure accurate and reliable estimates. For instance, the Z-test is a widely used estimator for comparing group proportions, providing accurate estimates when its underlying assumptions are satisfied (Kanji, 2006). Moreover, researchers can interpret the Z-test estimates as causal, provided they collect the data 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 interpret associational estimates as causal. They achieve this by ensuring data, and by extension an estimator, satisfies several key identification 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).

8.2. 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 8.1, the approach begins by defining the *individual causal effect* (ICE) as the difference between each student's potential outcomes, as in Equation 2.

$$\tau_i = Y_i \mid do(T_i = 1) - Y_i \mid do(T_i = 2) \tag{2}$$

where $do(T_i = t)$ represents the intervention operator, $Y_i \mid do(T_i = 1)$ represents the potential outcome under intervention $T_i = 1$, and $Y_i \mid do(T_i = 1)$ represents the potential outcome under intervention $T_i = 2$. Here, an *intervention* involves assigning a constant value to the treatment variable for each student's potential outcomes. 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 this challenge, the structural approach extends the ICE to the *average causal effect* (ACE, Equation 3), representing the average difference between the students' observed potential outcomes and their counterfactual counterparts.

$$\begin{aligned} \tau &= E[\tau_i] \\ &= E[Y_i \mid do(T_i = 1)] - E[Y_i \mid do(T_i = 2)] \end{aligned} \tag{3}$$

Even though counterfactuals are unobservable, researchers can still identify the ACE from associational estimates by leveraging the structural approach. The approach identifies the ACE by statistically conditioning data 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. When such a set exists, then T and Y are *d-separated* by X ($T \perp Y \mid X$) (Pearl, 2009), and X satisfies the *backdoor criterion* (Neal, 2020, pp 37). Here, *conditioning* describes the process of restricting the focus to the subset of the population defined by the conditioning variable (Neal, 2020, pp. 32) (see Equation 1).

Conditioning on a sufficient adjustment set enables researchers to estimate the ACE, even when the data comes from an observational study. This process is feasible because such conditioning ensures that the ACE estimator satisfies several critical properties, including confounding elimination (Morgan and Winship, 2014). Naturally, the validity of claims about the causal 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) noted, drawing conclusions about the real world from observed data inevitably requires assumptions. This requirement holds true for both observational and experimental data.

For instance, If researchers cannot conduct the randomized experiments described in Section 8.1 and must instead rely on observational data, they can still identify the ACE as long as an observed variable X, such as the socio-economic status of the school, satisfies the backdoor criterion. Under these circumstances, researchers first identify the *conditional average causal effect* (CACE, Equation 4)

$$CACE_t = E[Y_i \mid T_i = t, X] \tag{4}$$

From the CACE, researchers can identify the ACE from associational quantities as in Equation 5. This identification process is commonly known as the *backdoor adjustment*. Here, $E_X[\cdot]$ represents the marginal expected value over X (Morgan and Winship, 2014).

$$\begin{aligned} \tau &= E[Y_i \mid do(T_i = 1)] - E[Y_i \mid do(T_i = 2)] \\ &= E_X[CACE_1 - CACE_2] \\ &= E_X\left[E[Y_i \mid T_i = 1, X] - E[Y_i \mid T_i = 2, X]\right] \end{aligned} \tag{5}$$

Notably, the approach extends the ACE identification for a continuous variable T as in Equation 6,

ensuring broad applicability across different causal scenarios (Neal, 2020, pp. 45)

$$\begin{aligned} \tau &= E[Y_i \mid do(T_i = t)] \\ &= dE_X \left[E[Y_i \mid T_i = t, X] \right] / dt \end{aligned} \tag{6}$$

8.2.1. Diving into the specifics

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). Essentially, these tools serve as *conceptual (theoretical) models* on which identification analysis rests (Schuessler and Selb, 2023, pp. 4). Specifically, they guide researchers in determining which statistical models can identify an estimand (ACE, CACE, or other), assuming the depicted causal structure is correct (McElreath, 2020), thus enabling valid causal inference. Figure 9 shows the role of theoretical models in the inference process.

SCMs and DAGs support causal inference through two key advantages. First, regardless of complexity, they can represent various causal structures using only five fundamental building blocks (Neal, 2020; McElreath, 2020). This feature allows researchers to decompose complex structures into manageable components, facilitating their analysis (McElreath, 2020). Second, they depict causal relationships in a non-parametric and fully interactive way. This flexibility enables feasible ACE identification strategies without defining the variables' data types, the functional form between them, or their parameters (Pearl et al., 2016, pp. 35).

Section 8.2.1.1 and Section 8.2.1.2 elaborate on the first advantage, while Section 8.2.1.2 and Section 8.2.1.3 do so for the second.

8.2.1.1. The five fundamental block for SCMs and DAGs.

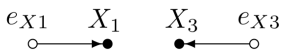
Figures 10, 11, 12, 13, and 14 display the five fundamental building blocks for SCMs and DAGs. The left panels of the figures show the formal mathematical models, represented by the SCMs, defined in terms of a set of *endogenous* variables $X = \{X_1, X_2, X_3\}$, a set of *exogenous* variables $E = \{e_{X1}, e_{X2}, e_{X3}\}$, and a set of functions $F = \{f_{X1}, f_{X2}, f_{X3}\}$ (Pearl, 2009; Cinelli et al., 2020). 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 ' $\perp$ ' to represent *d-separation*, a concept akin to (conditional) independence.

Notably, every SCM has an associated DAG (Pearl et al., 2016; Cinelli et al., 2020). The right panels of the figures display these DAGs. A DAG is a graph consisting of nodes connected by edges, where the nodes represent random variables. The term *directed* means that the edges extend from one node to another, with arrows indicating the direction of causal influence. The term *acyclic* implies 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). Although the *standard representation* of DAGs typically omits exogenous variables for simplicity, the *magnified representation* depicted in the figures offers one key advantage: including exogenous variables can help researchers highlight potential issues related to conditioning and confounding (Cinelli et al., 2020).

$$X_1 := f_{X1}(e_{X1})$$
$$X_3 := f_{X3}(e_{X3})$$
$$e_{X1} \perp e_{X3}$$

(a) SCM

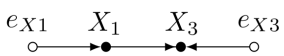


(b) DAG

Figure 10: Two unconnected nodes

$$X_1 := f_{X1}(e_{X1})$$
$$X_3 := f_{X3}(X_1, e_{X3})$$
$$e_{X1} \perp e_{X3}$$

(a) SCM



(b) DAG

Figure 11: Two connected nodes or descendant

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

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

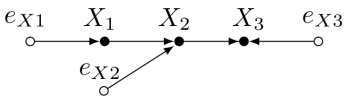
$$X_3 := f_{X3}(X_2, e_{X3})$$

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

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

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

(a) SCM



(b) DAG

Figure 12: Chain or mediator

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

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

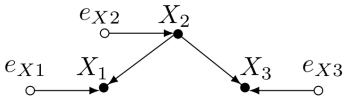
$$X_3 := f_{X3}(X_2, e_{X3})$$

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

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

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

(a) SCM



(b) DAG

Figure 13: Fork or confounder

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

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

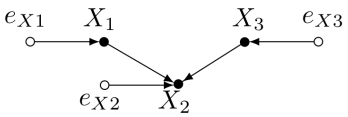
$$X_3 := f_{X3}(e_{X3})$$

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

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

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

(a) SCM



(b) DAG

Figure 14: Collider or inmorality

A careful examination of these building blocks highlights the theoretical assumptions underlying their observed variables. SCM 10a and DAG 10b depict two unconnected nodes, representing a scenario where variables X_1 and X_3 are independent or not causally related. SCM 11a and DAG 11b 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 there is a *direct path* connecting them. SCM 12a and DAG 12b 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. SCM 13a and DAG 13b 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, SCM 14a and DAG 14b 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, all building blocks assume the errors are independent of each other and from all other variables in the graph, as evidenced by the pairwise relations $e_{X1} \perp e_{X2}$, $e_{X1} \perp e_{X3}$, and $e_{X2} \perp e_{X3}$.

Researchers can then use these building blocks to represent the scenario outlined in Section 8.2. SCM 15a and DAG 15b depict the plausible causal structure for this example. In this context, the variable X (socio-economic status of the school) is thought to be a confounder in the relationship between the teaching method T and the outcome Y. In this scenario, the figures display multiple descendant relationships such as $X \to T$, $X \to Y$, and $T \to Y$. They also highlight unconnected node pairs, evident from the relationships $e_T \perp e_X$, $e_T \perp e_Y$, and $e_X \perp e_Y$. Additional, the figures show one fork, $X \to \{T, Y\}$, and two colliders: $\{X, e_T\} \to T$ and $\{X, T, e_Y\} \to Y$.

8.2.1.2. The probabilistic implications of these blocks.

Beyond their graphical capabilities, SCMs and DAGs can encode the probabilistic information embedded within a causal structure. They achieve this encoding by relying on three fundamental assumptions: the local Markov, the minimality, the causal edges assumption. The *local Markov assumption* encodes probabilistic independencies between variables by declaring that nodes in a graph are independent of all its non-descendants, given its parents (Neal, 2020, pp. 20). Meanwhile, the *minimality assumption* encodes probabilistic dependencies among variables by stating that every pair of adjacent nodes exhibits a dependency (Neal, 2020, pp. 21). Finally, the *causal edges assumption* encodes causal relationships between variables by declaring that each parent node acts

$$X := f_X(e_X)$$

$$T := f_T(X, e_T)$$

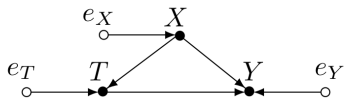
$$Y := f_Y(T, X, e_Y)$$

$$e_T \perp e_X$$

$$e_T \perp e_Y$$

$$e_X \perp e_Y$$

(a) SCM



(b) DAG

Figure 15: Plausible causal structure the scenario outlined in Section 8.2.

as a direct cause of its children (Neal, 2020, pp. 22). Figure 16 illustrates how these assumptions influence the statistical and causal interpretations of graphs.

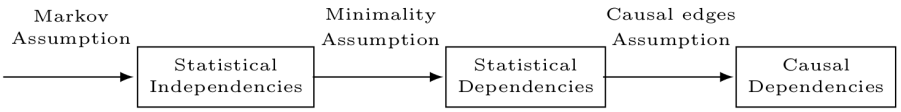


Figure 16: The flow of association and causation in graphs. Extracted and slightly modified from Neal (2020, pp. 31)

A notable implication of the assumptions underlying the probabilistic encoding is that any conceptual model described by an SCM and DAG can represent the joint distribution of variables more efficiently (Pearl et al., 2016, pp. 29). This expression takes the form of a product of conditional probability distributions (CPDs) of the type $P(child \mid parents)$. This property is formally known as the *Bayesian Network factorization* (BNF, Equation 7) (Pearl et al., 2016, pp. 29; Neal, 2020, pp. 21). In this expression, $pa(X_i)$ denotes the set of variables that are the parents of X_i.

$$\begin{aligned} P(X_1, X_2, \ldots, X_P) &= X_1 \cdot \prod_{p=2}^{P} P(X_i \mid X_{i-1}, \ldots, X_1) \quad \text{(by chain rule)} \\ &= X_1 \cdot \prod_{p=2}^{P} P(X_i \mid pa(X_i)) \qquad\qquad \text{(by BNF)} \end{aligned} \tag{7}$$

This encoding enables researchers with conceptual (theoretical) knowledge in the form of an SCM and DAG to predict patterns of (in)dependencies in the data. As highlighted by Pearl et al. (2016, pp. 35), these predictions depend solely on the structure of these conceptual models without requiring the quantitative details of the equations or the distributions of the errors. Moreover,

once researchers observe empirical data, the patterns of (in)dependencies in the data can provide significant insights into the validity of the proposed conceptual model.

The five fundamental building blocks described in Section 8.2.1.1 clearly illustrate which (in)dependencies can SMCs and DAGs predict. For instance, applying the BNF to the causal structure shown in the SCM 10a and DAG 10b enables researchers to express the joint probability distribution of the observed variables as $P(X_1, X_3) = P(X_1)P(X_3)$, supporting the theoretical assumption that the observed variables X_1 and X_3 are unconditionally independent ($X_1 \perp X_3$) (Neal, 2020, pp. 24). Conversely, when X_3 is unconditionally dependent on X_1 ($X_1 \not\perp X_3$), as depicted in the SCM 11a and DAG 11b, the BNF express their joint probability distribution as $P(X_1, X_3) = P(X_3 \mid X_1)P(X_1)$. Notably, these descriptions demonstrate the clear correspondence between the structural equations illustrated in Section 8.2.1.1 and the CPDs.

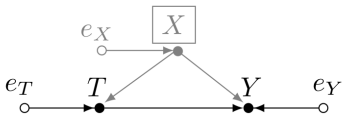
Beyond the insights gained from two-node structures, researchers can uncover more nuanced patterns of(in)dependencies from chains, forks, and colliders. These (in)dependencies apply to any data set generated by a causal model with those structures, regardless of the specific functions attached to the SCM (Pearl et al., 2016, pp. 36). For instance, applying the BNF to the chain structure depicted in the SCM 12a and DAG 12b allow researchers to represent the joint distribution for the observed variables as $P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$. This expression implies that X_1 and X_3 are unconditionally dependent ($X_1 \not\perp X_3$), but conditionally independent when controlling for X_2 ($X_1 \perp X_3 \,|\, X_2$). Moreover, in the fork structure shown in the SCM 13a and DAG 13b, researchers can express the joint distribution of the observed variables as $P(X_1, X_2, X_3) = P(X_1 \mid X_2)P(X_2)P(X_3 \mid X_2)$. Similar to the chain structure, this expression allows researchers to further infer that X_1 and X_3 are unconditionally dependent ($X_1 \not\perp X_3$), but conditionally independent when controlling for X_2 ($X_1 \perp X_3 \mid X_2$). Finally, researchers analyzing the collider structure illustrated in the SCM 14a and DAG 14b can express the joint distribution of the observed variables as $P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1, X_3)P(X_3)$. This representation allows researchers to infer that X_1 and X_3 are unconditionally independent ($X_1 \perp X_3$), but conditionally dependent when controlling for X_2 ($X_1 \not\perp X_3 \mid X_2$). The authors Pearl et al. (2016, pp. 37, 40, 41) and Neal (2020, pp. 25–26) provide the mathematical proofs for these conclusions.

Using these additional probabilistic insights, researchers can revisit the scenario in Section 8.2. In this context, applying the BNF to the SCM 17a structure, enables the representation of the joint probability distribution of the observed variables as $P(Y, T, X) = P(Y \mid T, X)P(T \mid X)P(X)$.

From this expression, researchers can infer that the outcome Y is unconditionally dependent on the teaching method T ($Y \not\perp\!\!\!\perp T$). This dependency arises from two key structures: a direct causal path from the teaching method T to the outcome Y, represented by the two-connected-nodes structure $T \to Y$ (black path in DAG 17b), and a confounding non-causal path from the teaching method T to the outcome Y through the socio-economic status of the school X, represented by the fork structure $T \leftarrow X \to Y$ (gray path in DAG 17b).

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

(a) SCM



(b) Conditioned DAG

Figure 17: Plausible causal structure the scenario outlined in Section 8.2.

8.2.1.3. From probability to causality.

The structural approach to causal inference translate the probabilistic insights from Section 8.2.1.3 into actionable strategies seeking to identify the ACE from associational quantities. It achieves this by relying on the *modularity assumption*, which posits that intervening on a node alters only the causal mechanism of that node, leaving others unchanged (Neal, 2020, pp. 34).

The modularity assumption underpins the concepts of manipulated graphs and Truncated Factorization, which are essential for representing interventions $P(Y_i \mid do(T_i = t))$ within SCMs and DAGs. *Manipulated graphs* simulate physical interventions by removing specific edges from a DAG, while preserving the remaining structure unchanged (Neal, 2020, pp. 34). In parallel, *Truncated Factorization* (TF) achieves a similar simulation by removing specific functions from the conceptual model and replacing them with constants, while keeping the rest of the structure unchanged (Pearl, 2010). The probabilistic implications of this factorization are formalized in Equation 8, where S represents the subset of variables X_p directly influenced by the intervention, while an example illustrating these concepts follows below.

$$P(X_1, X_2, \ldots, X_P \mid do(S)) = \begin{cases} \prod P(X_p \mid pa(X_p)) & \text{if } p \notin S \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

Using the TF, researchers can define the *backdoor adjustment* to identify the ACE. This adjustment states that if a variable $X_p \in S$ serves as a *sufficient adjustment set* for the effect of X_a on X_b, then the ACE can be identified using Equation 9. The sufficient adjustment set (potentially empty) must block all non-causal paths between X_a and X_b without introducing new paths. If such a set exists, then X_a and X_b are *d-separated* by X_p ($X_a \perp X_b \mid X_p$) (Pearl, 2009), and X_p satisfies the *backdoor criterion* (Neal, 2020, pp. 37).

$$P(X_a \mid do(X_b = x)) = \sum_{Xp} P(X_a \mid X_b = x, X_p) P(X_p) \tag{9}$$

Ultimately, the backdoor adjustment enables researchers to express the ACE as:

$$\begin{aligned} \tau &= E[X_a \mid do(X_b = 1)] - E[X_a \mid do(X_b = 2)] \\ &= E_{Xp}\left[E[X_a \mid do(X_b = 1), X_p] - E[X_a \mid do(X_b = 2), X_p]\right] \\ &= \sum_{Xp} X_a \cdot P(X_a \mid X_b = 1, X_p) \cdot P(X_p) - \sum_{Xp} X_a \cdot P(X_a \mid X_b = 2, X_p) \cdot P(X_p) \end{aligned} \tag{10}$$

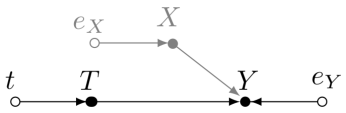
With these new insights, researchers revisiting the scenario in Section 8.2.1.2 can infer that the socio-economic status of the school, X, satisfies the backdoor criterion, assuming the causal structure depicted by the SCM 17a and DAG 17b is correct. This means that X serves as a sufficient adjustment set, as it effectively blocks all confounding non-causal paths introduced by the fork structure. Nevertheless, since Y remains dependent on T even after conditioning ($Y \not\perp T \mid X$), this dependency can only be attributed to the direct causal effect $T \to Y$.

Researchers can then apply the *backdoor adjustment* to identify the ACE of T on Y. They achieve this by first identifying the CACE of T on Y by conditioning on X, and then marginalizing this effect over X to obtain the ACE. This process is expressed in Equation 11 (see Section 8.2). Notably, for the purpose of identification, the conditioned DAG 17b is equivalent to the manipulated DAG 18b, because X satisfies the backdoor criterion.

$$
\begin{aligned}
\tau &= E[Y_i \mid do(T_i = 1)] - E[Y_i \mid do(T_i = 2)] \\
&= E_X\left[E[Y_i \mid T_i = 1, X] - E[Y_i \mid T_i = 2, X]\right] \\
&= \sum_X Y_i \cdot P(Y_i \mid T_i = 1, X) \cdot P(X) - \sum_X Y_i \cdot P(Y_i \mid T_i = 2, X) \cdot P(X)
\end{aligned} \tag{11}
$$

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

(a) SCM



(b) Manipulated DAG

Figure 18: Plausible causal structure the scenario outlined in Section 8.2.1.2.

8.2.1.4. The estimation process.

Ultimately, researchers can use Bayesian inference methods to estimate the ACE. The approach begins by defining two probability distributions: the likelihood of the data, $P(X_1, X_2, \ldots, X_P|\theta)$, and the prior distribution, $P(\theta)$ (Everitt and Skrondal, 2010), where X_P represents a random variable, and θ represents a one-dimensional parameter space for simplicity. After observing empirical data, researchers can update the priors to posterior distributions using Bayes' rule in Equation 12:

$$
P(\theta \mid X_1, X_2, \ldots, X_P) = \frac{P(X_1, X_2, \ldots, X_P \mid \theta) \cdot P(\theta)}{P(X_1, X_2, \ldots, X_P)} \tag{12}
$$

Given that the denominator on the right-hand side of Equation 12 serves as a normalizing constant independent of the parameter θ, researchers can simplify the posterior updating process into three steps. First, they integrate new empirical data through the likelihood. Second, they update the parameters' priors to a posterior distribution according to Equation 13. Ultimately, they normalize these results to obtain a valid probability distribution.

$$
P(\theta \mid X_1, X_2, \ldots, X_P) \propto P(X_1, X_2, \ldots, X_P \mid \theta) \cdot P(\theta) \tag{13}
$$

Temporarily setting aside the definition of prior distributions $P(\theta)$, note that the posterior updating process depends heavily on the assumptions underlying the likelihood of the data. However, as the number of random variables, P, increases, this joint distribution quickly becomes intractable (Neal, 2020). This intractability is evident from Equation 14, where the likelihood distribution is expressed by multiple chained CPDs.

$$P(X_1, X_2, \dots, X_P \mid \theta) = P(X_1 \mid \theta) \prod_{p=2}^{P} P(X_i \mid X_{i-1}, \dots, X_1, \theta) \tag{14}$$

Nevertheless, researchers can manage the complexity of the likelihood by assuming that variables exhibit specific local (in)dependencies. These assumptions enhance model tractability and facilitate the estimation process. As described in Section 8.2.1.2, researchers can use SCMs and DAGs to formalize these local (in)dependencies, which leads to the BNF of the likelihood (Equation 15). From this point, any probabilistic programming language can model this simpler structure and generate the parameter's posterior distribution using Equation 12.

$$P(X_1, X_2, \dots, X_P \mid \theta) = P(X_1 \mid \theta) \prod_{p=2}^{P} P(X_i \mid pa(X_i), \theta) \tag{15}$$

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.

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.

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.

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.

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.

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.

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-investigating-comparability-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.

Schuessler, J., Selb, P., 2023. Graphical causal models for survey inference. Sociological Methods and Research 0. doi:10.1177/00491241231176851.

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.