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 judgment (CJ), focusing on two prominent issues of traditional approaches. First, it critiques the heavy reliance on Thurstone's Case V assumptions and, by extension, the Bradley-Terry-Luce (BTL) model when analyzing CJ data. Specifically, the study raises concerns about the assumptions of equal discriminal dispersions and zero correlation between the stimuli. While these assumptions simplify the trait measurement model, they may fail to capture the complexity of CJ data, potentially leading to unreliable and inaccurate trait estimates. Second, the study highlights the apparent disconnect between CJ's trait measurement and hypothesis testing processes. Although separating these processes simplifies the analysis of CJ data, it may also undermine the reliability of various statistical results derived from these processes.

To address these issues, the study extends Thurstone's general form using a systematic and integrated approach based on Causal and Bayesian inference methods. This extension integrates core theoretical principles alongside key assessment design features relevant to CJ experiments, such as the selection of judges, stimuli, and comparisons. It then translates these elements into a probabilistic statistical model for analyzing dichotomous CJ data, overcoming the rigid assumptions of Case V and the BTL model.

Finally, the study emphasizes the relevance of this extension for contemporary empirical CJ research, particularly stressing the need for bespoke CJ models tailored to the experiments and data assumptions. It also lays the foundation for broader applications, encouraging researchers across the social sciences to adopt more robust and interpretable methodologies.

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 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 validity indicates the capacity of CJ scores to represent accurately the traits under measurement (Whitehouse, 2012; van Daal et al., 2016; Lesterhuis, 2018a; Bartholomew et al., 2018; Bouwer et al., 2023). Lastly, research on its practical applicability highlights CJ's 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, research in this domain remains unsystematic and fragmented, leaving several critical issues unresolved. This study identifies and discusses two prominent issues of traditional approaches that can undermine the reliability and validity of CJ's trait estimates (Perron and Gillespie, 2015). First, it critiques the heavy reliance on Thurstone's Case V assumptions (Thurstone, 1927a) and, by extension, the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959) when analyzing CJ data. Specifically, the study raises concerns about the assumptions of equal discriminal dispersions and zero correlation

*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)

between the stimuli. While these assumptions simplify the trait measurement model, they may fail to capture the complexity of CJ data, potentially leading to unreliable and inaccurate trait estimates. Second, the study highlights the disconnect between CJ's trait measurement and hypothesis testing processes. Although separating these processes simplifies the analysis of CJ data, it may also undermine the reliability of various statistical results derived from these processes.

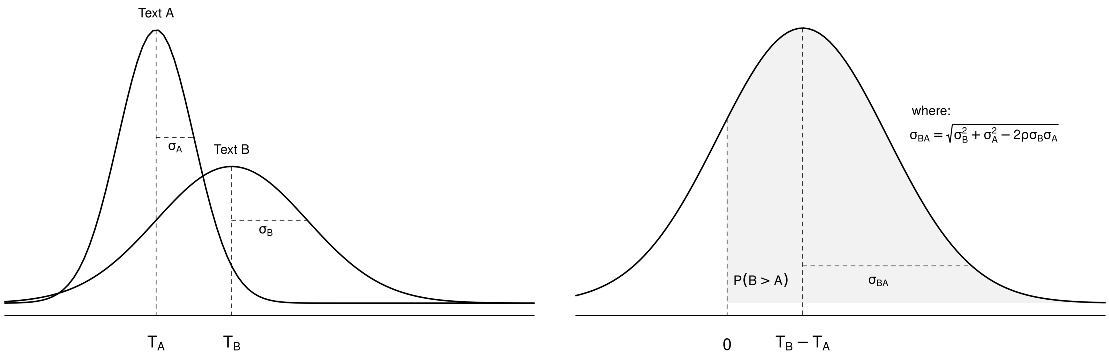
To address these issues, this study extends Thurstone's general form using a systematic and integrated approach based on Causal and Bayesian inference methods. In addition to improving statistical accuracy and strengthening measurement reliability and validity, the approach offers two key advantages. First, it clarifies the interactions among all actors and processes involved in CJ experiments. Second, it shifts the current comparative data analysis paradigm from passively accepting the BTL model assumptions to actively testing whether those assumptions fit the data under analysis.

As a result, the study divides its content into six main sections. Section 2 provides an overview of Thurstone's theory. Section 3 discusses the identified issues in detail. Section 4 extends Thurstone's general form to address these challenges. The extension integrates core theoretical principles alongside 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 probabilistic statistical model to analyze dichotomous pairwise comparison data. Section 6 discusses the findings, limitations, and challenges and explores avenues for future research. Finally, Section 7 summarizes the study's conclusions.

2. Thurstone's theory

In its most general form, Thurstone's theory addresses pairwise comparisons wherein a single judge evaluates multiple stimuli (Thurstone, 1927a). 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). 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). This suggests that pairwise comparison outcomes depend on the relative distance between stimuli, not their absolute positions on the continuum. 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. Two key observations support this conclusion: 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

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

as $P(B > A)$. As a result, the dichotomous outcome of this comparison is more likely to favor Text B over Text A.

3. Two Prominent Issues in Traditional CJ Practice

Thurstone noted from the outset that his general form, described in Section 2, led to a *trait scaling problem*. Specifically, the model required estimating more "unknown" parameters than the number of available pairwise comparisons (Thurstone, 1927a, pp. 267). For instance, in a CJ experiment with five texts, the general form would require estimating 20 parameters: five modal discriminal processes, five discriminal dispersions, and 10 correlations –one per comparison (see Table 1). However, a single judge could only provide $\binom{5}{2} = 10$ unique comparisons, an insufficient data set to estimate the required parameters.

To address this issue and facilitate the practical implementation of the theory, Thurstone developed five cases derived from this general form, each progressively incorporating additional simplifying assumptions. 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).

However, Thurstone developed Case V prioritizing statistical simplicity over precise trait measurement, and without providing clear guidance on how to use its measurement scores for hypothesis testing. Specifically, he cautioned that its use "should not be made without (an) experimental test"

(Thurstone, 1927a, pp. 270), as the case imposes the most extensive set of simplifying assumptions (Bramley, 2008; Kelly et al., 2022) (see Table 1). Moreover, because Thurstone's primary goal was to provide a "rather coarse scaling" of traits (Thurstone, 1927a, pp. 269) and "allocate the compared stimuli on this continuum" (Thurstone, 1927a, pp. 269), he did not address how to use the resulting measurement scores for hypothesis testing. Thus, given these limitations, it is surprising that CJ research has predominantly relied on Case V to measure different traits, raising significant concerns about the reliability and validity of such measurements in contexts where the case's assumptions may not hold (Kelly et al., 2022; Andrich, 1978). Furthermore, although the CJ practice has attempted to address the issue of hypothesis testing by using the scores' point estimates or their transformations, a critical question remains: Is this approach suitable for addressing these inquiries?

Thus, this section discusses these two prominent issues. Specifically, Section 3.1 examines the heavy reliance on Thurstone's Case V assumptions in the statistical analysis of CJ data. Conversely, Section 3.2 focuses on the apparent disconnect between the approaches to trait measurement and hypothesis testing in CJ.

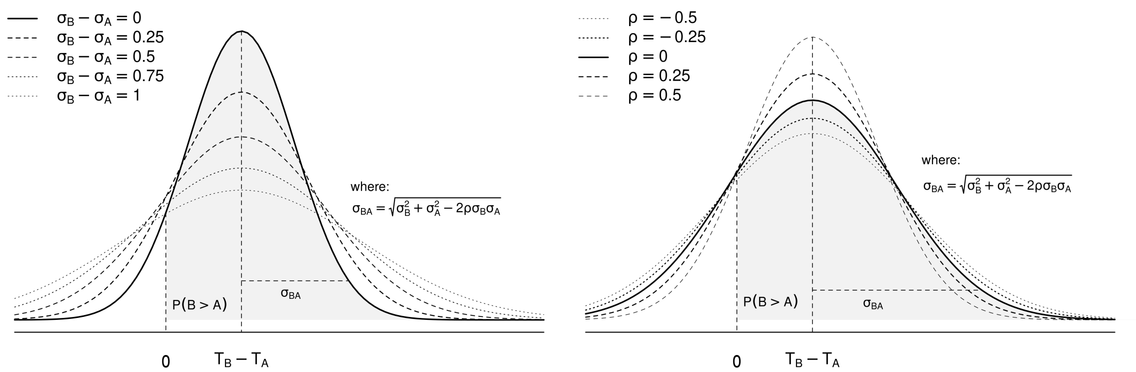
3.1. The Case V and the statistical analysis of CJ data

As previously discussed, Case V remains the most widely used model in CJ literature. This preference primarily stems from the BTL model, which provides a simplified statistical representation of the case. The BTL model mirrors the assumptions of Case V, with one notable distinction: while Case V assumes a Normal distribution for the stimuli' discriminal processes, the BTL model uses the more mathematically tractable Logistic distribution (Andrich, 1978; Bramley, 2008) (see Table 1). However, this substitution has minimal impact on the model's estimation or interpretation because the discriminal process scale is arbitrary up to a non-monotonic transformation (van der Linden, 2017a; McElreath, 2021). Furthermore, this limited impact is supported by the fact that the Normal and Logistic distributions exhibit analogous statistical properties, differing only by a scaling factor of approximately 1.7.

However, Thurstone acknowledged that some assumptions of Case V could be problematic when researchers assess complex traits or heterogeneous stimuli (Thurstone, 1927b). Thus, given that modern CJ applications often involve such traits and stimuli, two key assumptions of Case V, and by extension, the BTL model, may not always hold in theory or practice. These assumptions are the 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, directly influencing the outcome of pairwise comparisons. A thought experiment can help illustrate this idea. In it, researchers observe the discriminal processes for two texts, A and B, assuming that the dispersion for Text A remains constant and that the two texts are uncorrelated ($\rho = 0$). Figure 2a demonstrates that an increase in the uncertainty associated with the perception of Text B relative 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, as in the BTL model ($\sigma_B - \sigma_A = 0$), the discriminal difference distribution is more narrow than when the dispersions differ. As a result, the discriminal difference is more likely to favor Text B over Text A, as it is represented by the shaded gray area.



(a) Discriminal Difference distribution under varying discrepancies in stimuli dispersions

(b) Discriminal Difference distribution under varying levels of correlation between stimuli

Figure 2: The effect of dispersion discrepancies and stimuli 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 stimuli and their respective discriminal processes (Thurstone, 1927b). Consequently, the outcome's ability to reflect *true* differences between stimuli largely depends on the validity of the model's assumptions (Kohler et al., 2019), in this case, the assumption of equal dispersions. For instance, when the assumption accurately captures the complexity of the data, the BTL model estimates a discriminal difference distribution that accurately represents the *true* discriminal difference between the texts. This scenario is illustrated in Figure 2a, when the model's discriminal difference distribution aligns with the *true* discriminal difference distribution, represented by the thick continuous line corresponding to $\sigma_B - \sigma_A = 0$. The accuracy of this discriminal difference then ensures reliable estimates for the texts' discriminal processes.

Notably, while assuming equal dispersions simplifies the trait measurement model, evidence from the CJ literature suggests that this assumption may fail to capture the complexity of modern CJ data. In particular, the assumption may not hold when researchers assess complex traits or heterogeneous stimuli (Thurstone, 1927b; Bramley, 2008; Kelly et al., 2022), as these traits and stimuli can introduce judgment discrepancies due to their unique characteristics (van Daal et al., 2016; Lesterhuis, 2018b; Chambers and Cunningham, 2022). Indeed, the CJ literature may already provide indications of such discrepancies, particularly in the form of "misfit" statistics. For instance, *misfit texts* are those whose comparisons result in more judgment discrepancies than those involving other texts (Pollitt, 2004, 2012b,a; Goossens and De Maeyer, 2018). These misfit texts may exhibit larger discriminal dispersions or represent outliers –texts with distinctive characteristics that deviate markedly from the rest of the sample (Grubbs, 1969). In either case, comparing misfit texts results in more judgment discrepancies, yet the BTL model neither accounts for these cases nor offers any means to address them.

Significant statistical and measurement issues can arise when the assumption of equal dispersions between stimuli does not hold. Specifically, the BTL model may overestimate the trait's reliability, that is, the degree to which the outcome accurately reflects the *true* discriminal differences between stimuli. This overestimation, in turn, results in spurious conclusions about these differences (McElreath, 2020) and, by extension, about the underlying discriminal processes of stimuli. Figure 2a also illustrates this scenario 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 $\sigma_B - \sigma_A \neq 0$. Furthermore, if researchers acknowledge that misfit statistics may represent outlying observations, the common CJ practice of excluding stimuli based on these statistics (Pollitt, 2012a,b; van Daal et al., 2016; Goossens and De Maeyer, 2018), may unintentionally discard valuable information, introducing bias into the trait estimates (Zimmerman, 1994; McElreath, 2020, chap. 12). The direction and magnitude of these biases remain unpredictable, as they depend on which stimuli researchers exclude from the analysis.

3.1.2. The assumption of zero correlation between stimuli

The correlation ρ measures how much the judges' 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 depicted in Section 3.1.1, can illustrate this idea. The experiment only 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 narrow compared to scenarios where the texts are positively correlated. As a result, the discriminal difference is less likely to favor Text B over Text A, as it is represented by the shaded gray area.

Thurstone assumed that stimuli were uncorrelated because judges' biases, arising from two opposing and equally weighted effects occurring during the pairwise comparisons, canceled each other out (Thurstone, 1927a). Andrich (1978) provided a mathematical demonstration of this cancellation using the BTL model under the assumption of discriminal processes with additive biases. However, evidence from the CJ literature indicates that the assumption of zero correlation does not hold in practice in at least two scenarios: when intricate aspects of multidimensional, complex traits or heterogeneous stimuli influence judges' perceptions or when additional hierarchical structures are relevant to the stimuli.

In the first scenario, research on text quality suggests that when judges evaluate multidimensional, complex traits or heterogeneous stimuli, 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, when assessed, are unlikely to be equally weighted or opposing. As a result, they may exert an uneven influence on judges' perceptions, creating biases that resist cancellation. 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. Furthermore, since the discriminal difference of the stimuli becomes an observable outcome only through the judges' perceptions, these biases could introduce dependencies between the stimuli (van der Linden, 2017b). While direct evidence for this particular scenario is lacking, studies such as Pollitt and Elliott (2003) and van Daal et al. (2016) 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 introduced by additional hierarchical structures may create dependencies between stimuli, a statistical phenomenon known as clustering (Everitt and Skrondal, 2010). For instance, when multiple texts are produced by the same individual, these texts are likely to share common features, such as writing style or even their general quality. Although the CJ literature acknowledges the existence of such hierarchical structures, the statistical approaches to account for this additional source of dependence have been insufficient. For instance, when CJ data incorporates multiple samples of stimuli from the same individuals, researchers frequently rely on (averaged) point estimates of the BTL scores to conduct subsequent analyses and tests at the individual 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 3.2.

Thus, erroneously assuming zero correlation between stimuli can also lead to significant statistical and measurement issues. Specifically, neglecting judges' biases or relevant hierarchical structures can create dimensional mismatches in the model, leading to the over- or underestimation of trait reliability (Ackerman, 1989; Hoyle, 2023). These inaccuracies can result in spurious conclusions about the discriminal differences (McElreath, 2020) and, by extension, the underlying discriminal processes of the stimuli. This issue is illustrated in Figure 2b when the discriminal difference distribution of the BTL scores follows the thick continuous line ($\rho = 0$), while the *true* discriminal difference follows any discontinuous line where $\rho \neq 0$.

Finally, as discussed in the previous section, removing *misfit* judges based risks discarding valuable information and even introduce bias into the trait estimates (Zimmerman, 1994; McElreath, 2020, chap. 12). The direction and magnitude of these biases remain unpredictable because they depend on which judges researchers exclude from the analysis. *Misfit judges* are those whose evaluations deviate substantially from the shared consensus due to the unique characteristics of either the stimuli or the judges themselves (Pollitt, 2012a,b; van Daal et al., 2016; Goossens and De Maeyer, 2018).

3.2. The disconnect between trait measurement and hypothesis testing

Building on the previous section, it is clear that, researchers typically rely on the BTL model to measure a trait and place the compared stimuli along its continuum (Thurstone, 1927a). Additionally, the CJ literature shows that researchers frequently use point estimates of BTL scores or their transformations to conduct further analyses or hypothesis tests. For example, researchers have used these scores 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 (Casalicchio et al., 2015; 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).

Nevertheless, while separating the trait measurement and hypothesis testing processes simplifies the analysis of CJ data, the statistical literature cautions against relying solely on the point estimates of BTL scores to conduct further analyses or hypothesis tests, as this practice can undermine the resulting statistical conclusions. A key consideration is that BTL scores are parameter estimates that inherently carry uncertainty (measurement error). Ignoring this uncertainty can bias the analysis and reduce the precision of hypothesis tests. 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 (McElreath, 2020; Kline, 2023; Hoyle, 2023). Furthermore, the reduced precision in hypothesis tests diminishes their statistical power, increasing the likelihood of committing type-I or type-II errors (McElreath, 2020). Figure 3 illustrates these issues, demonstrating how neglecting measurement error (σ_T) by relying only on outcome averages can reduce the precision of a predictor's estimated effect.

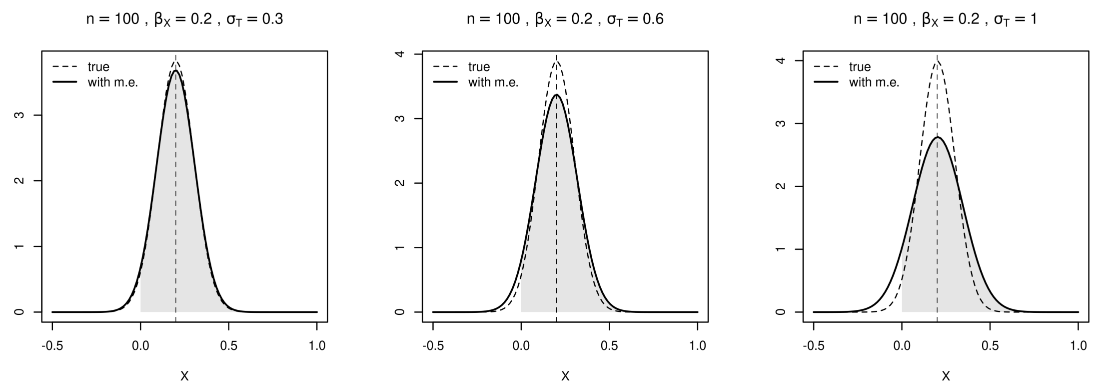


Figure 3: The effect of outcome uncertainty σ_T on the estimation of an effect β_X linked to the predictor. The example assumes a sample size $n = 100$ and an uncertainty that increases from left to right.

In aggregate, the heavy reliance on Thurstone's Case V assumptions in the statistical analysis of comparative data can compromise the reliability of trait estimates. This overreliance may also undermine their validity (Perron and Gillespie, 2015), particularly when coupled with the disconnect between the trait measurement and hypothesis testing processes. However, the structural approach to causal inference can address these issues by offering a systematic and integrated framework that strengthens measurement reliability and validity while enhancing the statistical accuracy of hypothesis tests.

4. Extending Thurstone's general form

The *structural approach* to causal inference provides a formal framework for identifying causes and estimating their effects using data. The approach 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 a system, such as the one found in CJ experiments. Essentially, SCMs and DAGs function as *conceptual models* on which identification analysis rests. *Identification analysis* helps researchers to determine whether an estimator can accurately compute an estimand based solely on its (causal) assumptions, regardless of random variability (Schuessler and Selb, 2023). Here, *estimands* represent the specific quantities researchers aim to determine (Everitt and Skrondal, 2010). *Estimators* denote the methods or functions that transform data into an estimate, while *estimates* are the numerical values approximating the estimand (Neal, 2020; Everitt and Skrondal, 2010).

A motivating example that will appear in the rest of the document clarifies these concepts. In this example, researchers aim to determine: "To what extent do different teaching methods influence students' ability to produce high-quality written texts?" To investigate this, a researcher designs a CJ experiment by randomly assigning students (individuals) to two groups, each receiving a different teaching method. Judges then compare pairs of students' written texts (stimuli) to produce a dichotomous outcome reflecting the relative quality of each text (trait). Based on this setup, researchers can reformulate the research question as the estimand: "*On average*, is there a difference in the ability to produce high-quality written texts between the two groups of students?". Following current CJ practices, researchers rely on estimates from the BTL model, or its transformations, to approximate this estimand.

However, Section 3 presents compelling evidence that Thurstone's Case V, and by extension the BTL model, suffers from several statistical and measurement limitations. These limitations hinder the model's ability to identify various estimands relevant to CJ inquiries, including the one described in the example. Identification is crucial because it is a necessary condition for ensuring consistent estimators. *Consistency* refers to the property of an estimator whose estimates converge to the "true" value of the estimand as the data size approaches infinity (Everitt and Skrondal, 2010). Without identification, consistency cannot be achieved, even with "infinite" and error-free data. Thus, deriving meaningful insights from finite data becomes impossible (Schuessler and Selb, 2023).

Fortunately, SCMs and DAGs support identification analysis through two key advantages[1]. First, regardless of complexity, they can represent various causal structures using only five fundamental building blocks (Neal, 2020; McElreath, 2024). This feature allows researchers to decompose complex structures into manageable components, facilitating their analysis. Second, they depict causal relationships in a non-parametric way. This flexibility enables feasible identification strategies without requiring specification of the types of variables, the functional forms relating them, or the parameters of those functional forms (Pearl et al., 2016).

Thus, this section addresses the issues identified in Section 3 by extending Thurstone's general form using the structural approach to Causal inference. Specifically, it combines the core theoretical principles outlined in Section 2 with key assessment design features relevant to CJ experiments, such as the selection of judges, stimuli, and comparisons. In addition to improving statistical accuracy and strengthening measurement reliability and validity, the approach offers two key advantages. First, it clarifies the interactions among all actors and processes involved in CJ experiments. Second, it shifts the current comparative data analysis paradigm from passively accepting the model assumptions to actively testing whether those assumptions fit the data under analysis.

Accordingly, Section 4.1 incorporates the theoretical principles into what we refer to as the *conceptual-population model*. This model assumes an idealized scenario where researchers have access to a *conceptual population* of comparative data, that is, data representing all repeated judgments made by every available judge for each pair of stimuli produced by each pair of individuals in the population. Conversely, Section 4.2 integrates the assessment design features into what we refer to as the *sample-comparison model*. This model assumes a more realistic scenario where researchers only have access to a sample of judges, individuals, stimuli, and comparisons from the conceptual population.

[1]These topics are beyond the scope of this study, thus, readers seeking a more profound understanding can refer to introductory papers such as Pearl (2010), Rohrer (2018), Pearl (2019), and Cinelli et al. (2020), and introductory books like Pearl and Mackenzie (2018), Neal (2020), and McElreath (2020) are useful. For more advanced study, seminal papers such as Neyman (1923), Rubin (1974), Spirtes et al. (1991), and Sekhon (2009), along with books such as Pearl (2009), Morgan and Winship (2014), and Hernán and Robins (2020), are recommended.

4.1. The conceptual-population model

In the conceptual-population model, the idealized scenario of a *conceptual population* of comparative data enables the integration of Thurstone's theoretical principles and provides a foundation for proposing innovations aimed at addressing some of the issues discussed in Section 3.

4.1.1. Integrating the first theoretical principles

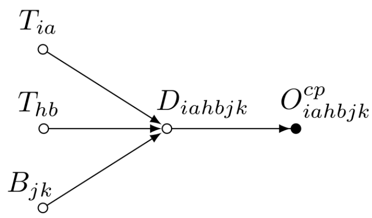
Before incorporating the first theoretical principles of Thurstone's theory, it is essential to further define SCMs. SCMs are formal mathematical models characterized by a set of *endogenous* variables V, a set of *exogenous* variables E, and a set of functions F (Pearl, 2009; Pearl et al., 2016; Cinelli 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). Lastly, the functions, referred to as *structural equations*, express the endogenous variables as non-parametric functions of other endogenous and exogenous variables. These functions use the symbol ':=' to denote the asymmetrical causal dependence between variables and the symbol ' $\perp$ ' to represent *d-separation*, a concept akin to (conditional) independence.

SCM 4a presents the first theoretical principles embedded in the conceptual-population model, which evaluates the impact of different teaching methods on students' writing ability. This SCM outlines the relationship between the conceptual-population outcome (O^{cp}_{iahbjk}) and several related variables. The subscripts i and h identify the students who authored the texts (i.e., the individuals). The indices a and b represent the texts under comparison (i.e., the stimuli). The index j indicates the judge conducting the comparison, while the index k captures *repeated measures designs* (Lawson, 2015, pp. 366-376), accounting for experimental conditions where a judge compares the same pair of stimuli multiple times. Thus, the indexing system supports comparisons between different texts written by the same student ($i = h$; $a \neq b$) and between texts written by distinct students ($i \neq h$; where $a = b$ is permitted), each judged once or repeatedly by all judges ($j = 1, \dots, n_J$; $k = 1, \dots, n_K$; where $n_J > 1$ and $n_K \geq 1$). However, it excludes cases where a judge compares a student's text to itself, whether once or multiple times ($i = h$; $a = b$; $j = 1, \dots, n_J$; $k = 1, \dots, n_K$; where $n_J > 1$ and $n_K \geq 1$), as such comparison lacks practical relevance within the CJ framework. Here, n_J indicates the total number of judges, and n_K denotes the number of repeated judgments each judge performs.

In line with Thurstone's theory, SCM 4a depicts the texts' discriminal processes (T_{ia}, T_{hb}) and their discriminal difference (D_{iahbjk}) (see Section 2). Additionally, based on the arguments developed in Section 3.1.2, the SCM incorporates the judges' biases (B_{kj}). Together with the outcome, these variables constitute the preliminary set of endogenous variables, $V = \{O_{iahbjk}, D_{iahbjk}, T_{ia}, T_{hb}, B_{kj}\}$. Finally, the SCM presents the preliminary set of structural equations, $F = \{f_O, f_D\}$, which define the non-parametric dependencies among these variables.

$$\begin{aligned} O^{cp}_{iahbjk} &:= f_O(D_{iahbjk}) \\ D_{iahbjk} &:= f_D(T_{ia}, T_{hb}, B_{jk}) \end{aligned}$$

(a) SCM



(b) DAG

Figure 4: Conceptual-population model, scalar form.

Notably, every SCM has an associated DAG (Pearl et al., 2016; Cinelli et al., 2020). A DAG is a *graph* consisting of nodes connected by edges, where nodes represent random variables. The term *directed* indicates that edges or arrows extend from one node to another, indicating the direction of causal influence. The absence of an edge implies no direct relationship between the nodes. The term *acyclic* means that the causal influences do not form loops, ensuring the influences do not cycle back on themselves (McElreath, 2020). DAGs conventionally depict observed variables as solid black circles and unobserved (latent) variables as open circles (Morgan and Winship, 2014). Although DAGs conventionally omit exogenous variables for simplicity, the DAGs presented in this section includes exogenous variables to improve clarity and reveal potential issues related to conditioning and confounding (Cinelli et al., 2020).

Figure 4b displays the DAG corresponding to SCM 4a, illustrating the expected causal relationships outlined in Thurstone's theory. The graph shows that the discriminal processes of the texts (T_{ia}, T_{hb}) influence their discriminal difference (D_{iahbjk}), which in turn determines the outcome (O^{cp}_{iahbjk}). It also highlights the influence of judges' biases (B_{kj}) on the discriminal difference. Additionally, the DAG differentiates between observed endogenous variables, such as the outcome (solid black circle), and latent endogenous variables, including the texts' discriminal processes, their discriminal difference, and the judges' biases (open circles).

4.1.2. The conceptual population *data structure*

Although specifying a data structure is not mandatory when using SCMs and DAGs, defining one in this case can improve clarity and facilitate the description of the system. Thus, to re-express the scalar form of the CJ system shown in Figure 4 into an equivalent vectorized form, we first define the vectors I and J, along with the matrices IA and JK, as follows:

$$I = \begin{bmatrix} 1 \\ \vdots \\ i \\ \vdots \\ h \\ \vdots \\ n_I \end{bmatrix} ; \; J = \begin{bmatrix} 1 \\ \vdots \\ j \\ \vdots \\ n_J \end{bmatrix} ; \; IA = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & n_A \\ \vdots & \vdots \\ i & a \\ \vdots & \vdots \\ h & b \\ \vdots & \vdots \\ n_I & 1 \\ \vdots & \vdots \\ n_I & n_A \end{bmatrix} ; \; JK = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & n_K \\ \vdots & \vdots \\ j & k \\ \vdots & \vdots \\ n_J & 1 \\ \vdots & \vdots \\ n_J & n_K \end{bmatrix} \tag{1}$$

Here, each element of I represents a unique individual i or h, where n_I denotes the total number of individuals. Similarly, each element of J corresponds to a unique judge j, with n_J indicating the total number of judges. Moreover, each row of IA represents a unique pairing of individuals i, h with stimuli a, b. As a result, the matrix IA contains $n_I \cdot n_A$ rows and 2 columns, where n_A specifies the number of stimuli available per individual. Likewise, each row of JK associates a judge j with a (repeated) judgment index k. Consequently, the matrix JK has $n_J \cdot n_K$ rows and 2 columns, where n_K indicates the number of repeated judgments each judge makes.

Additionally, we construct the matrix R to map each row of the IA matrix with a corresponding row from the JK matrix. Thus, this matrix has n rows and 6 columns, where $n = \binom{n_I \cdot n_A}{2} \cdot n_J \cdot n_K$. Here, the term $\binom{n_I \cdot n_A}{2}$ represents the binomial coefficient, which quantifies the total number of unique comparisons possible between every pair of stimuli generated by each pair of individuals in the population. Thus, we define the matrix as follows:

$$R = \begin{bmatrix} 1 & 1 & 1 & 2 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 2 & 1 & n_K \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ i & a & h & b & j & k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_I & n_A - 1 & n_I & n_A & n_J & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_I & n_A - 1 & n_I & n_A & n_J & n_K \end{bmatrix} \tag{2}$$

It is easier to visualize the structure of these vectors and matrices by considering an example where $n_I = 5$, $n_A = 2$, $n_J = 3$, and $n_K = 3$. In this simple case, the vectors and matrices described in equations (1) and (2) take the following form:

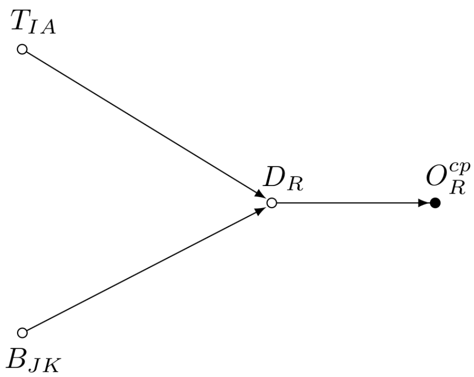
$$I = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} ; \; J = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} ; \; IA = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ 2 & 2 \\ 3 & 1 \\ 3 & 2 \\ 4 & 1 \\ 4 & 2 \\ 5 & 1 \\ 5 & 2 \end{bmatrix} ; \; JK = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 2 & 1 \\ 2 & 2 \\ 2 & 3 \\ 3 & 1 \\ 3 & 2 \\ 3 & 3 \end{bmatrix} ; \; R = \begin{bmatrix} 1 & 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 2 \\ 1 & 1 & 1 & 2 & 1 & 3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 5 & 2 & 1 & 1 \\ 1 & 1 & 5 & 2 & 1 & 2 \\ 1 & 1 & 5 & 2 & 1 & 3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 4 & 2 & 5 & 2 & 3 & 1 \\ 4 & 2 & 5 & 2 & 3 & 2 \\ 4 & 2 & 5 & 2 & 3 & 3 \\ 5 & 1 & 5 & 2 & 3 & 1 \\ 5 & 1 & 5 & 2 & 3 & 2 \\ 5 & 1 & 5 & 2 & 3 & 3 \end{bmatrix} \tag{3}$$

Now, using equations (1) and (2), we can re-express SCM 4a and DAG 4b in an equivalent vectorized form, as shown in Figure 5. In this depiction, the outcome O_R^{cp}, the texts' discriminal difference D_R, their discriminal processes T_{IA}, and the judges' biases B_{JK} are represented as vectors rather than scalar values. These vectors capture all the observations from the conceptual population. Specifically, O_R^{cp} and D_R are observed and latent vectors of length n, respectively. Moreover, T_{IA}

and B_{JK} are latent vectors of lengths $n_I \cdot n_A$ and $n_J \cdot n_K$, respectively.

$$O_R^{cp} := f_O(D_R)$$
$$D_R := f_D(T_{IA}, B_{JK})$$

(a) SCM



(b) DAG

Figure 5: Conceptual-population model, initial vectorized form.

4.1.3. Integrating hierarchical structural components

Building on the principles of Structural Equation Modeling (SEM) (Hoyle, 2023) and Item Response Theory (IRT) (Fox, 2010; van der Linden, 2017a), the conceptual-population model integrates two *hierarchical structural components* to examine how different teaching methods influence students' writing ability. Each structural component defines how observed or latent variables affect the primary latent variable of interest (Everitt and Skrondal, 2010). Their hierarchical nature enables researchers to test hypotheses considering the hierarchical structure of stimuli (see Section 3.1.2) and the uncertainties in trait estimation (see Section 3.2).

The top branch of DAG 6b illustrates the first component, where *relevant* [2] student-related variables (X_I), such as teaching method, and students' idiosyncratic errors e_I causally influence the latent variable representing students' writing-quality trait (T_I). The error term e_I captures variations in students' traits unexplained by X_I. Here, X_I is an observed matrix with n_I rows and q_I

[2] *Relevant variables* are those that satisfy the *backdoor criterion* (Neal, 2020, pp 37), that is, they belong to a *sufficient adjustment set* (Pearl, 2009; Pearl et al., 2016; Morgan and Winship, 2014). A *sufficient* set (potentially empty) blocks all non-causal paths between a predictor and an outcome without opening new ones (Pearl, 2009). Refer also to footnote 1.

columns (variables), and both e_I and T_I are latent vectors of length n_I. Additionally, this branch shows how T_I, along with *relevant* [3] text-related variables (X_{IA}) (e.g., text length), and texts' idiosyncratic errors e_{IA} causally influence the texts' written-quality trait (T_{IA}), the first primary latent variable of interest. The error term e_{IA} captures variations in the texts' traits that remain unexplained by T_I or X_{IA}. Here, X_{IA} is an observed matrix with dimensions $n_I \cdot n_A$ rows and q_{IA} columns (variables), while e_{IA} and T_{IA} are latent matrices with n_I rows and n_A columns.

Similarly, the bottom branch of DAG 6b depicts the second component, where *relevant* [4] judge-related variables (Z_J), such as judgment expertise, and judges' idiosyncratic errors e_J causally influence the latent variable representing judges' bias (B_J). The error e_J captures variations in judges' bias unexplained by Z_J. Here, Z_J is an observed matrix with n_J rows and q_J columns (variables), and both e_J and B_J are latent vectors of length n_J. Furthermore, the branch shows how B_J, along with *relevant* [5] judgment-related variables (Z_{JK}) (e.g., the number of judgments a judge makes), and judgments' idiosyncratic errors e_{JK} causally influence the judges' biases associated with each text B_{JK}, the second primary latent variable of interest. The error e_{JK} captures variations in judgments unexplained by B_J or Z_{JK}. Here, Z_{JK} is an observed matrix with dimension $n_J \cdot n_K$ rows and q_{JK} columns (variables), while e_{JK} and B_{JK} are latent latent matrices with n_J rows and n_K columns

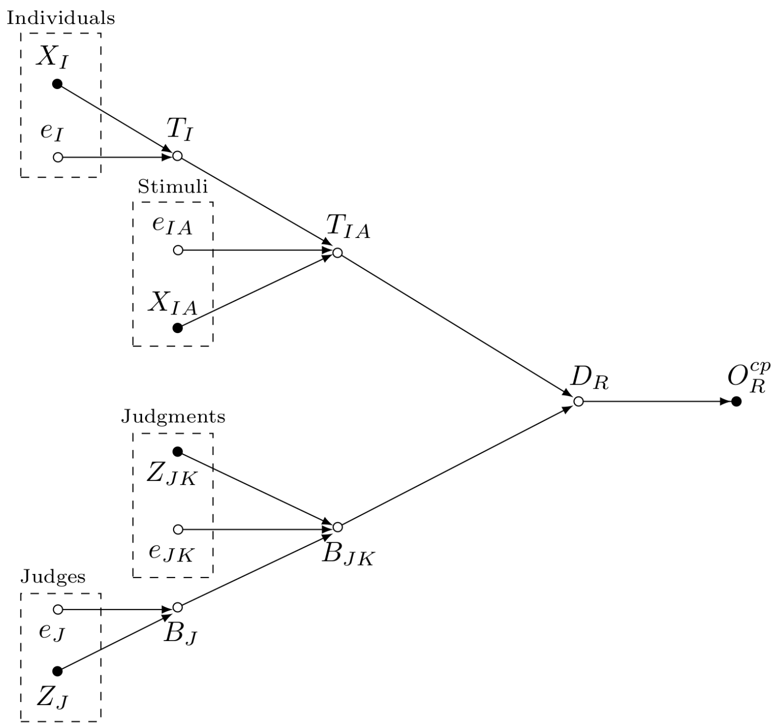
Notably, all variables and functions shown in SCM 6a and DAG 6b are part of the set of endogenous variables V, structural equations F, and exogenous variables E for the conceptual-population model. Additionally, the figures demonstrate that all exogenous variables are independent of one another, as indicated by the relationships $e_{IA} \perp \{e_I, e_{JK}, e_J\}$, $e_I \perp \{e_{JK}, e_J\}$ and $e_{JK} \perp e_J$.

Overall, the conceptual-population model extends Thurstone's general form by introducing key innovations to address the limitations discussed in Section 3.1.2 and Section 3.2. These enhancements include accounting for judges' biases and integrating hierarchical structural components. Nevertheless, despite its promise of enhancing measurement accuracy and precision, the model still depends on the unrealistic assumption that researchers have access to a *conceptual population*. Since researchers rarely meet this assumption in practice, they must consider a more realistic scenario.

[3] refer to footnote 2.
[4] refer to footnote 2.
[5] refer to footnote 2.

$$
\begin{aligned}
O_R^{cp} &:= f_O(D_R) \\
D_R &:= f_D(T_{IA}, B_{JK}) \\
T_{IA} &:= f_T(T_I, X_{IA}, e_{IA}) \\
T_I &:= f_T(X_I, e_I) \\
B_{JK} &:= f_B(B_J, Z_{JK}, e_{JK}) \\
B_J &:= f_B(Z_J, e_J) \\
e_I &\perp \{e_J, e_{IA}, e_{JK}\} \\
e_J &\perp \{e_{IA}, e_{JK}\} \\
e_{IA} &\perp e_{JK}
\end{aligned}
$$

(a) SCM



(b) DAG

Figure 6: Conceptual-population model, final vectorized form.

4.2. The sample-comparison model

The sample-comparison model presents a more realistic scenario than the conceptual-population model. In Section 4.2.1, it explicitly assumes researchers work with a data sample consisting of a limited number of repeated judgments (n_K^s) from a sample of judges (n_J^s) and a specific number of texts (n_A^s) from a sample of students (n_I^s), all drawn from the conceptual population. Furthermore, in Section 4.2.2, the model assumes that judges do not perform *all repeated judgments* within the data sample. Instead, they conduct a sufficient number of stimuli comparisons, n_C, to ensure an accurate estimation of the proportion $P(B > A)$, as proposed by Thurstone (1927a).

4.2.1. The sample mechanism

To incorporate the sampling mechanism and facilitate the interpretation of the sample-comparison model, we first define the *data sampling process* using the binary vector variables S_I, S_J, S_{IA}, and S_{JK} as follows:

$$S_I = \begin{bmatrix} i_{(1)} \\ \vdots \\ i_{(i)} \\ \vdots \\ i_{(h)} \\ \vdots \\ i_{(nI)} \end{bmatrix} ; \; S_J = \begin{bmatrix} j_{(1)} \\ \vdots \\ j_{(j)} \\ \vdots \\ j_{(nJ)} \end{bmatrix} ; \; S_{IA} = \begin{bmatrix} ia_{(1,1)} \\ \vdots \\ ia_{(1,n_A)} \\ \vdots \\ ia_{(i,a)} \\ \vdots \\ ia_{(h,b)} \\ \vdots \\ ia_{(nI,1)} \\ \vdots \\ ia_{(nI,nA)} \end{bmatrix} ; \; S_{JK} = \begin{bmatrix} jk_{(1,1)} \\ \vdots \\ jk_{(1,n_K)} \\ \vdots \\ jk_{(j,k)} \\ \vdots \\ jk_{(nJ,1)} \\ \vdots \\ jk_{(nJ,nK)} \end{bmatrix} \tag{4}$$

Where each element of S_I is a binary value indicating the presence or absence of corresponding elements in the vector I, as in equation (5). We apply the same logic to S_J using vector J (not shown). Thus, the vectors S_I and S_J contains n_I and n_J elements, respectively.

$$i_{(i)} = \begin{cases} 1 & \text{if data element } i \text{ from } I \text{ is sampled} \\ 0 & \text{if data element } i \text{ from } I \text{ is missing} \end{cases} \tag{5}$$

Similarly, each element of S_{IA} is a binary value indicating the presence or absence of data rows in the matrices IA, as defined in equation (6). We apply the same logic to S_{JK} using the matrix JK (not shown). Thus, the vectors S_{IA} and S_{JK} contains $n_I \cdot n_A$ and $n_J \cdot n_K$ elements, respectively.

$$ia_{(i,a)} = \begin{cases} 1 & \text{if data elements } i, a \text{ from } IA \text{ are sampled} \\ 0 & \text{if data elements } i, a \text{ from } IA \text{ are missing} \end{cases} \tag{6}$$

We can visualize these vectors more clearly using the example in equation (3). Suppose researchers exclude the second student and the third judge, and they also omit the second text from each individual and the third repeated judgment from each judge. Given $n_I = 5$, $n_A = 2$, $n_J = 3$, and $n_K = 3$, these vectors take the following structure:

$$S_I = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} ; \; S_J = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} ; \; S_{IA} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} ; \; S_{JK} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

Equation (7) shows that missing observations in the vectors S_I and S_J–which represent unsampled students and judges–directly determine which observations are missing in S_{IA} and S_{JK}. In other words, researchers can only observe texts and judgments from students and judges initially included in the sample. The equation also shows that the sum of observed elements in S_I equals the number of sampled students (n_I^s) and that a similar sum in vector S_J equals the sampled judges (n_J^s). Conversely, the sum of observed elements in S_{IA} represents the total sampled texts across all sampled students ($n_I^s \cdot n_A^s$), while a similar sum in vector S_{JK} represents the total sampled repeated judgments across all sampled judges ($n_J^s \cdot n_K^s$). Notice that because the design systematically excludes every third repeated judgment, researchers can also express S_{JK} using $n_K = n_K^s = 2$.

Finally, we define the *sample mechanism* S in equation (8), which maps each element of S_{IA} to every element of S_{JK}. Each element $s_{(i,a,h,b,j,k)}$ is a binary value indicating the presence or absence of data rows in the matrix R resulting from the sample mechanism, as in equation (9). Thus, the vector contains n elements, matching the number of rows in R, and the sum of its elements represents the total data sample: $n^s = \binom{n_I^s \cdot n_A^s}{2} \cdot n_J^s \cdot n_K^s$. Here, the term $\binom{n_I^s \cdot n_A^s}{2}$ represents the binomial coefficient, which quantifies the total number of unique comparisons possible between every pair of sampled stimuli generated by each pair of sampled individuals.

$$S = \begin{bmatrix} s_{(1,1,1,2,1,1)} \\ \vdots \\ s_{(1,1,1,2,1,n_K)} \\ \vdots \\ s_{(i,a,h,b,j,k)} \\ \vdots \\ s_{(n_I,n_A-1,n_I,n_A,n_J,1)} \\ \vdots \\ s_{(n_I,n_A-1,n_I,n_A,n_J,1)} \end{bmatrix} \tag{8}$$

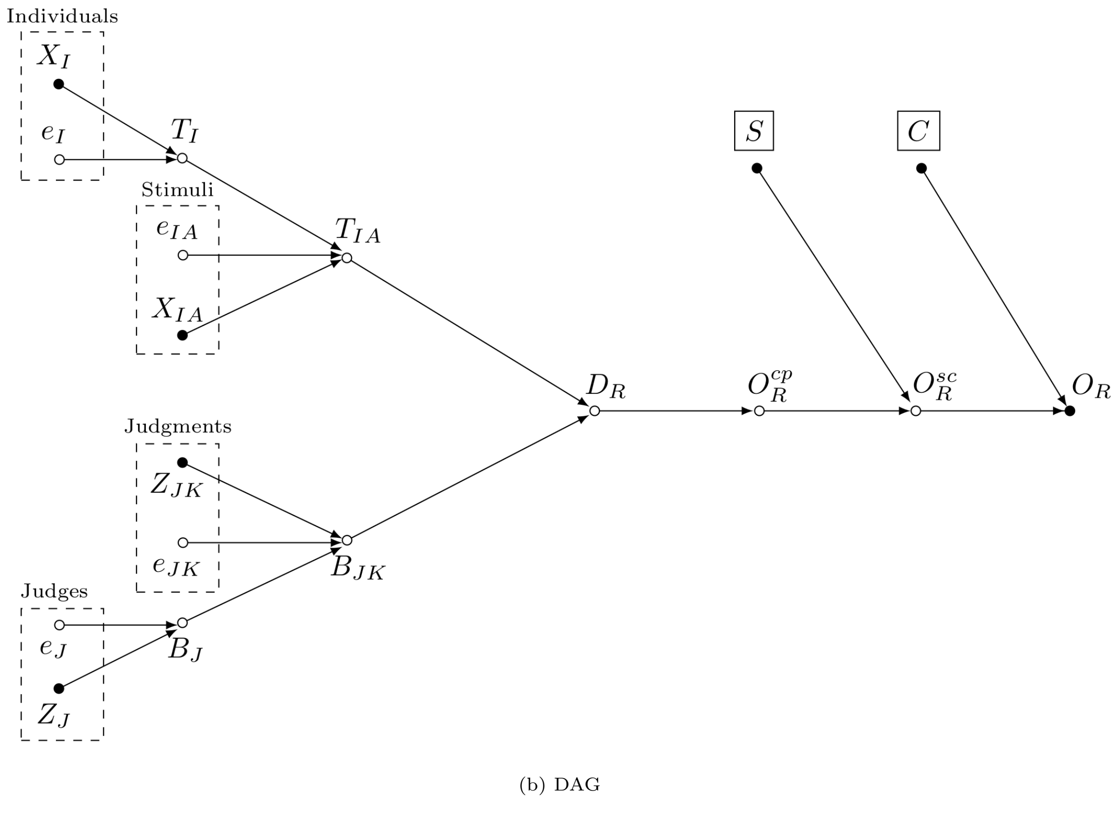
$$s_{(i,a,h,b,j,k)} = \begin{cases} 1 & \text{if data elements } i,a,h,b,j,k \text{ from } R \text{ are sampled} \\ 0 & \text{if data elements } h,i,a,b,j,k \text{ from } R \text{ are missing} \end{cases} \tag{9}$$

With the definition of S, we incorporate the sample mechanism into the conceptual-population model. Following the convention of McElreath (2020) and Deffner et al. (2022), DAG 7b represents the conceptual-population outcome O_R^{cp} as unobserved, emphasizing that researchers cannot directly access this outcome due to the sampling mechanism. The DAG also depicts the *sample design* vector S as a causal factor influencing the sample-comparison outcome O_R^{sc}. A square encloses this vector, indicating that it is a conditioned variable. Here, *conditioning* means that researchers restrict their focus to the elements of O_R^{cp} that satisfy $s_{(i,a,h,b,j,k)} = 1$ (Neal, 2020; McElreath, 2020). Essentially, S serves as the vector selecting *all repeated judgments a subset of judges makes for the subset of stimuli produced by the sampled individuals*. This process generates the sample-comparison outcome O_R^{sc}.

Notably, the DAG shows that S is independent of all other variables in the model. This implies that DAG 7b applies exclusively to Simple Random Sampling (SRSg) designs. In these designs, each repeated judgment, judge, stimulus, and individual has the same probability of being included in the sample as any other observation within their respective groups (Lawson, 2015).

$$
\begin{aligned}
O_R &:= f_C(O_R^{sc}, C) \\
O_R^{sc} &:= f_S(O_R^{cp}, S) \\
O_R^{cp} &:= f_O(D_R) \\
D_R &:= f_D(T_{IA}, B_{JK}) \\
T_{IA} &:= f_T(T_I, X_{IA}, e_{IA}) \\
T_I &:= f_T(X_I, e_I) \\
B_{JK} &:= f_B(B_J, Z_{JK}, e_{JK}) \\
B_J &:= f_B(Z_J, e_J) \\
e_I &\perp \{e_J, e_{IA}, e_{JK}\} \\
e_J &\perp \{e_{IA}, e_{JK}\} \\
e_{IA} &\perp e_{JK}
\end{aligned}
$$

(a) SCM



(b) DAG

Figure 7: Sample-comparison model, final vectorized form

However, due to concerns about the practical feasibility of the comparison task (Boonen et al., 2020), CJ experiments rarely implement an exhaustive pairings of sampled judges, stimuli, and individuals. Thus, a realistic scenario must account for the fact that judges typically compare only a subset of stimuli authored by certain individuals.

4.2.2. The comparison mechanism

As in the previous section, we begin defining the *comparison mechanism* using the binary vector variable C to facilitate the interpretation of the sample-comparison model (see equation (10)). The vector contains n elements corresponding to the number of rows in the R matrix, with each element $c_{(i,a,h,b,j,k)}$ being a binary value indicating the presence or absence of data rows in R, a definition similar to that of $s_{(i,a,h,b,j,k)}$ in equation (9).

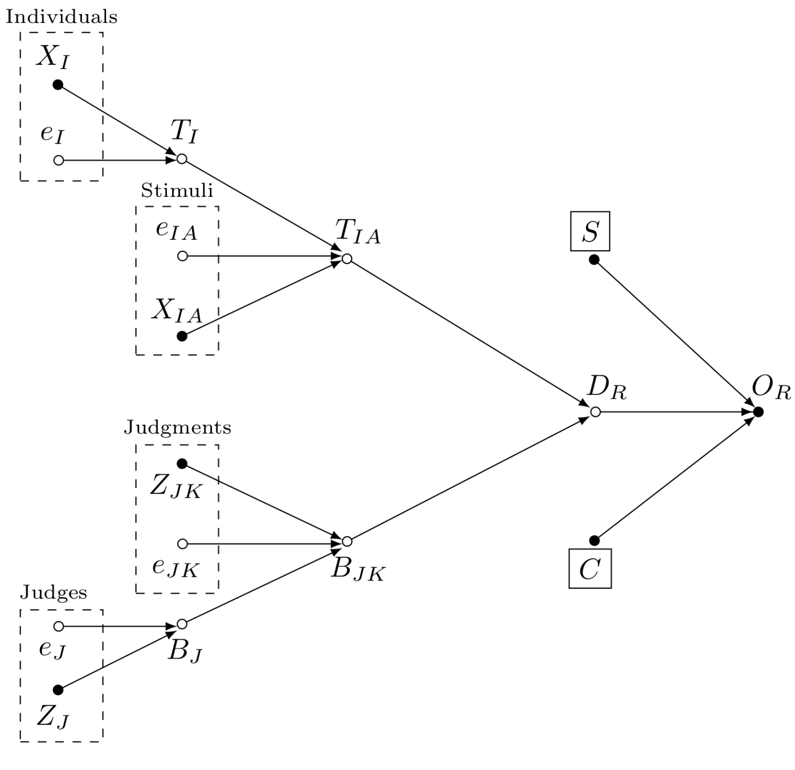
$$C = \begin{bmatrix} c_{(1,1,1,2,1,1)} \\ \vdots \\ c_{(1,1,1,2,1,n_K)} \\ \vdots \\ c_{(i,a,h,b,j,k)} \\ \vdots \\ c_{(n_I,n_A-1,n_I,n_A,n_J,1)} \\ \vdots \\ c_{(n_I,n_A-1,n_I,n_A,n_J,1)} \end{bmatrix} \tag{10}$$

The DAG 7b also integrates the *comparison mechanism* C into the conceptual-population model, ultimately deriving the outcome O_R. It shows the sample-comparison outcome O_R^{sc} as unobserved, highlighting that researchers cannot directly access to O_R^{sc} due to the comparison mechanism. Moreover, the DAG depicts C as a causal factor influencing O_R, enclosing it in a square to indicate that it is a conditioned variable. Specifically, C determines *which repeated judgments the subset of judges makes for the subset of stimuli produced by the sampled individuals.* In essence, C reflects the assumption that judges *do not* perform all possible repeated judgments for the selected stimuli. Instead, they carry out a "sufficient number of comparisons" n_C to allow for accurate estimation of the proportion $P(B > A)$ for each stimulus pair (Thurstone, 1927a, pp. 267).

Notably, DAG 7b also shows that C is independent of all other variables in the model. This independence implies that the conceptual model represented by the DAG applies exclusively to Random Allocation Comparative Designs (Bramley, 2015) or Incomplete Block Designs (Lawson, 2015). In these designs, every repeated judgment has an equal probability of being included in the sample.

$$
\begin{aligned}
O_R &:= f_O(D_R, S, C) \\
D_R &:= f_D(T_{IA}, B_{JK}) \\
T_{IA} &:= f_T(T_I, X_{IA}, e_{IA}) \\
T_I &:= f_T(X_I, e_I) \\
B_{JK} &:= f_B(B_J, Z_{JK}, e_{JK}) \\
B_J &:= f_B(Z_J, e_J) \\
e_I &\perp \{e_J, e_{IA}, e_{JK}\} \\
e_J &\perp \{e_{IA}, e_{JK}\} \\
e_{IA} &\perp e_{JK}
\end{aligned}
$$

(a) SCM



(b) DAG

Figure 8: Comparative judgment model

Finally, since it is standard to assume that the distribution of the conceptual-population outcome O_R^{cp} also holds for O_R^{sc} and O_R, we can reformulate the sample-comparison model in Figure 7 into the equivalent form shown in Figure 8. This reformulation produces a model that applies directly to a sample of comparative data. In this version, the unobserved outcomes O_R^{cp} and O_R^{sc} are omitted, and O_R inherits the structural equation f_O that originally defined O_R^{cp}. Moreover, the definition of O_R now reflects its direct dependence on the discriminal difference D_R and the sample and comparison mechanisms, S and C.

In summary, the SCM 8a and DAG 8b extend Thurstone's general form to address several limitations of the BTL model. These extensions account for judge biases (see Section 3.1.2), reflect the hierarchical structure of stimuli (see Section 3.1.2), and incorporate measurement error in trait estimation and hypothesis testing (see Section 3.2). However, they do not resolve concerns about the assumption of equal dispersions among stimuli discussed in Section 3.1.1. Because this issue involves statistical assumptions about the distribution of the discriminal process, it requires a statistical model, which we develop in the next section.

5. From SCMs to a statistical model

With the structure in SCM 8a, we can derive a statistical model that addresses violations of the equal dispersion assumption in CJ experiments (see Section 3.1.1). This derivation is possible because a fully specified SCM encodes functional and probabilistic information, which we can replace with appropriate functions and probabilistic assumptions (Pearl et al., 2016). Specifically, SCM 8a allows us to express the joint distribution of our CJ system as a product of simpler conditional probability distributions (CPDs)[6], as shown in equation (11). For clarity, we treat expressions such as $Y := f_Y(X)$, $P(Y \mid X)$, and $Y \sim f(Y \mid X)$ as equivalent. Here, $P(Y \mid X)$ and $f(Y \mid X)$ represent the CPD of Y given X.

$$
\begin{aligned}
P(O_R, &S, C, D_R, T_{IA}, X_{IA}, e_{IA}, T_I, X_I, e_I, B_{JK}, Z_{JK}, e_{JK}, B_J, Z_J, e_J) \\
&= P(O_R \mid D_R, S, C) \cdot P(S) \cdot P(C) \cdot P(D_R \mid T_{IA}, B_{JK}) \\
&\quad \cdot P(T_{IA} \mid T_I, X_{IA}, e_{IA}) \cdot P(T_I \mid X_I, e_I) \\
&\quad \cdot P(B_{JK} \mid B_J, Z_{JK}, e_{JK}) \cdot P(B_J \mid Z_J, e_J) \\
&\quad \cdot P(X_{IA}) \cdot P(X_I) \cdot P(Z_{JK}) \cdot P(Z_J) \\
&\quad \cdot P(e_{IA}) \cdot P(e_I) \cdot P(e_{JK}) \cdot P(e_J)
\end{aligned}
\tag{11}
$$

[6]This re-expression is possible because the *chain rule* of probability and the *Bayesian Network Factorization (BNF)* property. For further details, see Pearl et al. (2016) and Neal (2020).

Each CPD in equation (11) rests on specific assumptions, which we outline in the SCM 9a, the probabilistic model 9a, and the statistical model 9a. First, we assume that O_R follows a Bernoulli distribution[7], reflecting the binary nature of CJ outcomes. Furthermore, following the conventions of Generalized Linear Latent and Mixed Models (GLLAMMs) (Rabe-Hesketh et al., 2004; Skrondal and Rabe-Hesketh, 2004; Rabe-Hesketh et al., 2007), the distribution links O_R to the latent discriminal difference vector D_R using an inverse-logit function: $\text{inv\_logit}(x) = 1/(1 + \exp(-x))$. Note that while the probabilistic model 9b includes the probability distributions for the sampling and comparison mechanisms, $P(S)$ and $P(C)$, we omit these from the statistical model 9c. We do so because, although these distributions contribute to the joint distribution of the data, they do not affect the relationship between O_R and D_R. This lack of influence is due to the random selection designs that govern S and C[8].

Second, we define D_R as the difference between the discriminal processes of two texts, $T_{IA}[i,a]$ and $T_{IA}[h,b]$, plus the corresponding repeated judge biases, $B_{JK}[j,k]$. The discriminal processes represent the underlying written-quality trait of the texts. Note that if we assume that $B_{JK}[j,k]$ reflects the difference in stimulus-specific biases, i.e., $B_{JK}[j,k] = B_{JK}[i,a,j,k] - B_{JK}[h,b,j,k]$, we can re-write the discriminal difference as:

$$\begin{aligned} D_R &= (T_{IA}[i,a] - T_{IA}[h,b]) + B_{JK}[j,k] \\ &= (T_{IA}[i,a] + B_{JK}[i,a,j,k]) - (T_{IA}[h,b] + B_{JK}[h,b,j,k]) \\ &= T^*_{IA}[i,a] - T^*_{IA}[h,b] \end{aligned} \tag{12}$$

This formulation reveals that the discriminal difference captures a *pure interaction effect*, in which neither the texts' discriminal processes nor the judges' biases alone determine the outcome, but their interaction does (Attia et al., 2022). In simple terms, the discriminal processes of the stimuli

[7]The binomial distribution–including its special case, the Bernoulli distribution– represent a maximum entropy distribution for binary events (McElreath, 2020, pp. 34). This means that when only two un-ordered outcomes are possible, and the expected numbers of each type of event are assumed to be constant, then the binomial distribution is the one that is most consistent with these constraints (McElreath, 2020, pp. 310). For a detailed discussion of the binomial as a maximum entropy distribution, see McElreath (2020, chap. 10.1.2).

[8]Randomization ensures that data—and, by extension, an estimator—satisfies several key identification properties, such as common support, no interference, and consistency. The most critical property, however, is the elimination of confounding. *Confounding* occurs when an external variable, such as X_I, simultaneously influences both the outcome (e.g., O_R) and a variable of interest (e.g., S), resulting in spurious associations between the two (Everitt and Skrondal, 2010). Randomization addresses this issue by effectively decoupling the association between the variable of interest and any other variable, except for the outcome itself. For a more detailed discussion on the benefits of randomization, see Pearl (2009), Morgan and Winship (2014), Neal (2020), and Hernán and Robins (2020).

become an observable outcome only through the judges' perceptions (biases). It is important to clarify that the square brackets in D_R merely indicate the relevant indices for each trait vector, without implying any subsetting of the data.

$$
\begin{aligned}
O_R &:= f_O(D_R, S, C) \\
D_R &:= f_D(T_{IA}, B_{JK}) \\
T_{IA} &:= f_T(T_I, X_{IA}, e_{IA}) \\
T_I &:= f_T(X_I, e_I) \\
B_{JK} &:= f_B(B_J, Z_{JK}, e_{JK}) \\
B_J &:= f_B(Z_J, e_J) \\
e_I &\perp \{e_J, e_{IA}, e_{JK}\} \\
e_J &\perp \{e_{IA}, e_{JK}\} \\
e_{IA} &\perp e_{JK}
\end{aligned}
$$

(a) SCM

$$
\begin{aligned}
&P(O_R \mid D_R, S, C) \cdot P(S) \cdot P(C) \\
&P(D_R \mid T_{IA}, B_{JK}) \\
&P(T_{IA} \mid T_I, X_{IA}, e_{IA}) \\
&P(T_I \mid X_I, e_I) \\
&P(B_{JK} \mid B_J, Z_{JK}, e_{JK}) \\
&P(B_J \mid Z_J, e_J) \\
&P(e_I) \\
&P(e_J) \\
&P(e_{IA}) \\
&P(e_{JK})
\end{aligned}
$$

(b) Probabilistic model

$$
\begin{aligned}
O_R &\overset{iid}{\sim} \text{Bernoulli}\left[\text{inv\_logit}(D_R)\right] \\
D_R &= (T_{IA}[i,a] - T_{IA}[h,b]) + B_{JK}[j,k] \\
T_{IA} &= T_I + \beta_{XA} X_{IA} + e_{IA} \\
T_I &= \beta_{XI} X_I + e_I \\
B_{JK} &= B_J + \beta_{ZK} Z_{JK} + e_{JK} \\
B_J &= \beta_{ZJ} Z_J + e_J \\
e_I &\overset{iid}{\sim} \text{Normal}(0, s_{XI}) \\
e_J &\overset{iid}{\sim} \text{Normal}(0, s_{ZJ}) \\
e_{IA} &\overset{iid}{\sim} \text{Normal}(0, p_{IA}) \\
e_{JK} &\overset{iid}{\sim} \text{Normal}(0, p_{JK})
\end{aligned}
$$

(c) Statistical model

Figure 9: Comparative judgment model, SCM, probabilistic and statistical model assuming different discriminal dispersions for the student's traits

Third, we define T_{IA} as a linear combination of the students' underlying writing-quality traits T_I, the effects of relevant text-related variables $\beta_{XA}X_{IA}$ (e.g., text length on quality), and the texts' idiosyncratic errors e_{IA}. We assume that each element of e_{IA} is independent and identically distributed (*iid*) according to a Normal distribution, with a mean of zero and a standard deviation p_{IA}. The parameter p_{IA} is defined as a proportion of 1 to establish the scale of the latent trait, which is required in any latent variable model (Depaoli, 2021). Note that because e_{IA} follows a Normal distribution, T_{IA} is also normally distributed. This result implies that our model reverts to assuming Normal discriminal processes for the stimuli, as originally proposed by Thurstone (1927a) (see Table 1).

Fourth, we define T_I as a linear combination of the effects of relevant student-related variables, $\beta_{XI}X_I$, and students' idiosyncratic errors, e_I. As before, we assume that each element of e_I is *iid* Normal with a mean of zero. However, we consider the standard deviation, s_{XI}, varies depending on the teaching method group to which each student belongs. Given the example in Section 4, where the teaching method $X_I = 1, 2$, the model sets the following constraint to anchor the scale of the latent trait (Depaoli, 2021): $\sum_{g=1}^{2} s_{XI}[g]/2 = 1$. Note that this parametrization of e_I relaxes the equal dispersion assumption for the stimuli, effectively addressing the concerns raised in Section 3.1.1.

Fifth, we define B_{JK} as a linear combination of the judges' individual bias B_J, the effects of relevant judgment-related variables $\beta_{ZK} Z_{JK}$ (e.g., how the number of judgments a judge makes affect the assessment of quality), and judgment-specific idiosyncratic errors e_{JK}. We assume that each element of e_{JK} is *iid* Normal with a mean of zero and a standard deviation p_{JK}, defined as a proportion of 1, to anchor the scale of the latent trait (Depaoli, 2021).

Sixth, we define B_J as a linear combination of the effects of relevant judge-level variables $\beta_{ZJ} Z_J$ (e.g., judgment expertise) and judge-specific idiosyncratic errors e_J. Each element of e_J is likewise assumed to be *iid* Normal with mean zero and a standard deviation s_{ZJ} that depends on the groups to which each judge belongs. For instance, if Z_J is a variable that can take on three values representing three groups of judges with varying expertise, the model imposes the constraint $\sum_{g=1}^{3} s_{ZJ}[g]/3 = 1$ to fix the scale of the latent trait (Depaoli, 2021).

Finally, to transform model 9c into a practical statistical tool for analyzing paired comparison data, we use *Bayesian inference methods*. The **Declarations** section of this document includes a link to the model, along with an alternative specification that assumes equal discriminal dispersions. We implemented both versions of the model using `Stan` (version 2.26.1, Stan Development Team., 2021). Although the implementation was successful, we do not report the results here, as they fall outside the scope of this document.

6. Discussion

6.1. Limitations and further research

7. Conclusion

Declarations

Funding: The Research Fund (BOF) of the University of Antwerp funded this project.

Financial interests: The authors declare no relevant financial interests.

Non-financial interests: The authors declare no relevant non-financial interests.

Ethics approval: The University of Antwerp Research Ethics Committee confirmed that this study does not require ethical approval.

Consent to participate: Not applicable

Consent for publication: All authors have read and approved the final version of the manuscript for publication.

Data availability: This study did not use any data.

Materials and code availability: The `CODE LINK` section at the top of the digital document located at: https://jriveraespejo.github.io/paper2_manuscript/ provides access to all materials and code.

AI-assisted technologies in the writing process: The authors used various AI-based language tools to refine phrasing, optimize wording, and enhance clarity and coherence throughout the manuscript. They take full responsibility for the final content of the publication.

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

8.1. Statistical and Causal inference

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).

8.1.1. 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 13, where $E[\cdot]$ denotes the expected value.

$$E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 2] \tag{13}$$

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). 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 10, visually represents the transition from estimands to estimates.

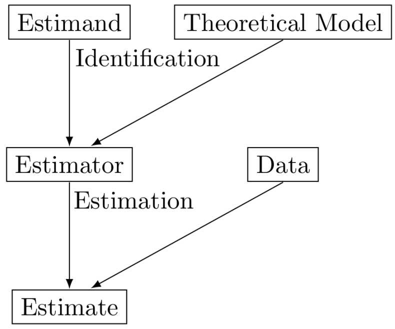


Figure 10: Identification-Estimation flowchart. Extracted and slightly modified from Neal (2020, pp. 32)

Identification is a necessary condition to ensure *consistent* estimators. An estimator achieves *consistency* when it converges to the "true" value of an estimand as the data size approaches infinity (Everitt and Skrondal, 2010). Without identification, researchers cannot achieve consistency, even with "infinite" and error-free data. As a result, deriving meaningful insights about an estimand from finite data becomes impossible (Schuessler and Selb, 2023, pp. 5). Therefore, to ensure accurate and reliable estimates, researchers prioritize estimators with desirable identification properties. For instance, the Z-test is a widely used estimator for comparing group proportions, yielding accurate estimates when its underlying assumptions are satisfied (Kanji, 2006). Furthermore, researchers can interpret estimates from the Z-test as causal, provided the data is 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 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).

8.1.2. 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 8.1.1, the approach begins by defining the *individual causal effect* (ICE) as the difference between each student's potential outcomes, as in Equation 14.

$$\tau_i = Y_i \mid do(T_i = 1) - Y_i \mid do(T_i = 2) \tag{14}$$

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 15), representing the average difference between the students' observed potential outcomes and their counterfactual counterparts.

$$\begin{aligned} \tau &= E[\tau_i] \\ &= E[Y_i \mid do(T_i = 1)] - E[Y_i \mid do(T_i = 2)] \end{aligned} \tag{15}$$

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 16).

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 8.1.1 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 16)

$$CACE_t = E[Y_i \mid T_i = t, X] \tag{16}$$

From the CACE, researchers can identify the ACE from associational quantities as in Equation 17. 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).

$$
\begin{aligned}
\tau &= E[Y_i \mid do(T_i = 1)] - E[Y_i \mid do(T_i = 2)] \\
&= E_X[CACE_1 - CACE_2] \\
&= E_X\left[E[Y_i \mid T_i = 1, X] - E[Y_i \mid T_i = 2, X]\right]
\end{aligned}
\tag{17}
$$

Notably, the approach extends the ACE identification for a continuous variable T as in Equation 18, ensuring broad applicability across different causal scenarios (Neal, 2020, pp. 45)

$$
\begin{aligned}
\tau &= E[Y_i \mid do(T_i = t)] \\
&= dE_X\left[E[Y_i \mid T_i = t, X]\right]/dt
\end{aligned}
\tag{18}
$$

8.1.3. 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). Thus, using these tools, researchers can determine which statistical models can identify (ACE, CACE, or other), assuming the depicted causal structure is correct (McElreath, 2020), thus enabling valid causal inference. Figure 10 shows the role of theoretical models in the inference process.

SCMs and DAGs support identification analysis 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).

Thus, Section 8.1.3.1 and Section 8.1.3.2 elaborate on the first advantage, while Section 8.1.3.2 and Section 8.1.3.3 do so for the second. Finally, Section 8.1.3.4 explains how researchers use SCMs and DAGs alongside Bayesian inference methods in the estimation process.

8.1.3.1. The five fundamental block for SCMs and DAGs.

Figures 11, 12, 13, 14, and 15 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 $V = \{X_1, X_2, X_3\}$, 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). 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.

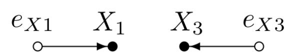
Notably, every SCM has an associated DAG (Pearl et al., 2016; Cinelli et al., 2020). The right panels of the figures display these DAGs. A DAG is a graph consisting of nodes connected by edges, where the nodes represent random variables. The term *directed* means that the edges extend from one node to 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

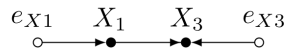
Figure 11: Two unconnected nodes

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

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

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

(a) SCM



(b) DAG

Figure 12: Two connected nodes or descendant

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

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

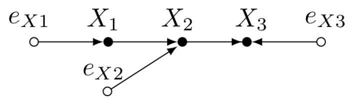
$$X_3 := f_{X3}(X_2, e_{X3})$$

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

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

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

(a) SCM



(b) DAG

Figure 13: Chain or mediator

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

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

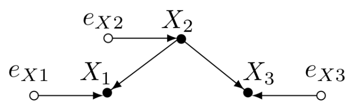
$$X_3 := f_{X3}(X_2, e_{X3})$$

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

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

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

(a) SCM



(b) DAG

Figure 14: Fork or confounder

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

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

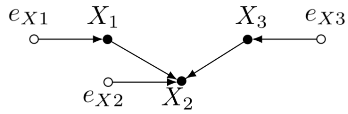
$$X_3 := f_{X3}(e_{X3})$$

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

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

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

(a) SCM



(b) DAG

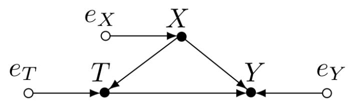
Figure 15: Collider or inmorality

A careful examination of these building blocks highlights the theoretical assumptions underlying their observed variables. SCM 11a and DAG 11b depict two unconnected nodes, representing a scenario where variables X_1 and X_3 are independent or not causally related. SCM 12a and DAG 12b 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 13a and DAG 13b 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 14a and DAG 14b 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 15a and DAG 15b 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 8.1.2. SCM 16a and DAG 16b 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. The figures display multiple descendant relationships such as $X \rightarrow T$, $X \rightarrow Y$, and $T \rightarrow Y$. 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 $\{X, T, e_Y\} \to Y$.

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

(a) SCM



(b) DAG

Figure 16: Plausible causal structure the scenario outlined in Section 8.1.2.

8.1.3.2. 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 as a direct cause of its children (Neal, 2020, pp. 22). Figure 17 illustrates how these assumptions influence the statistical and causal interpretations of graphs.

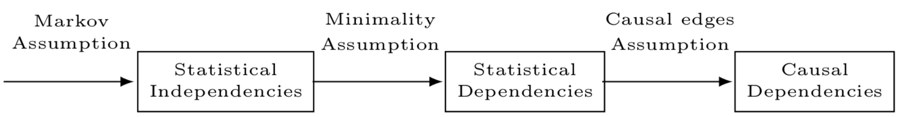


Figure 17: The flow of association and causation in graphs. Extracted and slightly modified from Neal (2020, pp. 31)

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 19) (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.

$$\begin{aligned} P(X_1, X_2, \dots, X_P) &= X_1 \cdot \prod_{p=2}^{P} P(X_i \mid X_{i-1}, \dots, X_1) \quad &\text{(by chain rule)} \\ &= X_1 \cdot \prod_{p=2}^{P} P(X_i \mid pa(X_i)) \quad &\text{(by BNF)} \end{aligned} \tag{19}$$

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.

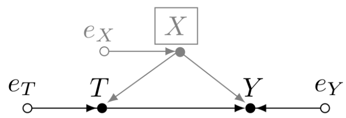
The five fundamental building blocks described in Section 8.1.3.1 clearly illustrate which (in)dependencies can SMCs and DAGs predict. For instance, applying the BNF to the causal structure shown in the SCM 11a and DAG 11b enables researchers to express the joint probability distribution of the observed 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 12a and DAG 12b, 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 8.1.3.1 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 13a and DAG 13b 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 ($X_1 \perp X_3 \mid X_2$). Moreover, in the fork structure shown in the SCM 14a and DAG 14b, 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 15a and DAG 15b 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 8.1.2. In this context, applying the BNF to the SCM 18a 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 ($Y \not\perp T$). 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 18b), 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 18b).

$$X := x$$
$$T := f_T(x, e_T)$$
$$Y := f_Y(T, x, e_Y)$$
$$e_T \perp e_X$$
$$e_T \perp e_Y$$
$$e_X \perp e_Y$$

(a) SCM



(b) Conditioned DAG

Figure 18: Plausible causal structure the scenario outlined in Section 8.1.2.

8.1.3.3. From probability to causality.

The structural approach to causal inference translates probabilistic insights into actionable strategies seeking to identify the ACE from associational quantities. The approach 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 20, where S represents the subset of variables X_p directly influenced by the intervention, while an example illustrating these concepts follows below.

$$P(X_1, X_2, \dots, X_P \mid do(S)) = \begin{cases} \prod P(X_p \mid pa(X_p)) & \text{if } p \notin S \\ 1 & \text{otherwise} \end{cases} \tag{20}$$

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 21. 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).

$$P(X_a \mid do(X_b = x)) = \sum_{Xp} P(X_a \mid X_b = x, X_p) P(X_p) \tag{21}$$

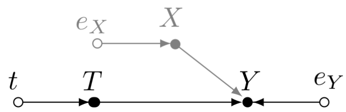
Ultimately, the backdoor adjustment enables researchers to express the ACE as:

$$\begin{aligned} \tau &= E[X_a \mid do(X_b = 1)] - E[X_a \mid do(X_b = 2)] \\ &= E_{Xp}\left[E[X_a \mid do(X_b = 1), X_p] - E[X_a \mid do(X_b = 2), X_p]\right] \\ &= \sum_{Xp} X_a \cdot P(X_a \mid X_b = 1, X_p) \cdot P(X_p) - \sum_{Xp} X_a \cdot P(X_a \mid X_b = 2, X_p) \cdot P(X_p) \end{aligned} \tag{22}$$

With these new insights, researchers revisiting the scenario in Section 8.1.3.2 can infer that the socio-economic status of the school, X, satisfies the backdoor criterion, assuming the causal structure depicted by the SCM 18a and DAG 18b 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 $T \rightarrow Y$. Notably, for the purpose of identification, the conditioned DAG 18b is equivalent to the manipulated DAG 19b, because X satisfies the backdoor criterion.

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

(a) SCM



(b) Manipulated DAG

Figure 19: Plausible causal structure the scenario outlined in Section 8.1.3.2.

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 23 (see Section 8.1.2).

$$\begin{aligned}
\tau &= E[Y_i \mid do(T_i = 1)] - E[Y_i \mid do(T_i = 2)] \\
&= E_X \left[E[Y_i \mid T_i = 1, X] - E[Y_i \mid T_i = 2, X] \right] \\
&= \sum_X Y_i \cdot P(Y_i \mid T_i = 1, X) \cdot P(X) - \sum_X Y_i \cdot P(Y_i \mid T_i = 2, X) \cdot P(X)
\end{aligned} \tag{23}$$

8.1.3.4. 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, $P(X_1, X_2, \dots, X_P \mid \theta)$, 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 24:

$$P(\theta \mid X_1, X_2, \dots, X_P) = \frac{P(X_1, X_2, \dots, X_P \mid \theta) \cdot P(\theta)}{P(X_1, X_2, \dots, X_P)} \tag{24}$$

Given that the denominator on the right-hand side of Equation 24 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 25. Ultimately, they normalize these results to obtain a valid probability distribution.

$$P(\theta \mid X_1, X_2, \ldots, X_P) \propto P(X_1, X_2, \ldots, X_P \mid \theta) \cdot P(\theta) \tag{25}$$

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 26, where the likelihood distribution is expressed by multiple chained CPDs.

$$P(X_1, X_2, \ldots, X_P \mid \theta) = P(X_1 \mid \theta) \prod_{p=2}^{P} P(X_i \mid X_{i-1}, \ldots, X_1, \theta) \tag{26}$$

Nevertheless, researchers can manage the complexity of the likelihood by assuming specific local (in)dependencies among variables. SCMs and DAGs provide a formal framework to represent these assumptions, as detailed in Section 8.1.3.2. These assumptions improve model tractability and simplify the estimation process by enabling the derivation of the BNF of the likelihood (Equation 27). With this simplified structure, any probabilistic programming language can model the system and compute the parameter's posterior distribution using Equation 24.

$$P(X_1, X_2, \ldots, X_P \mid \theta) = P(X_1 \mid \theta) \prod_{p=2}^{P} P(X_i \mid pa(X_i), \theta) \tag{27}$$

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.

Attia, J., Holliday, E., Oldmeadow, C., 2022. A proposal for capturing interaction and effect modification using dags. International Journal of Epidemiology 51, 1047–1053. doi:10.1093/ije/dyac126.

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., 2015. Investigating the reliability of adaptive comparative judgment. URL: http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf. cambridge Assessment Research Report.

Bramley, T., Vitello, S., 2019. The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. Assessment in Education: Principles, Policy and Practice 71, 1–25. doi:10.1080/0969594X.2017.1418734.

Casalicchio, G., Tutz, G., Schauberger, G., 2015. Subject-specific bradley–terry–luce models with implicit variable selection. Statistical Modelling 15, 526–547. doi:10.1177/1471082X15571817.

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.

Deffner, D., Rohrer, J., McElreath, R., 2022. A causal framework for cross-cultural generalizability. Advances in Methods and Practices in Psychological Science 5. doi:10.1177/25152459221106366.

Depaoli, S., 2021. Bayesian Structural Equation Modeling. Methodology in the social sciences, The Guilford Press.

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

Fox, J., 2010. Bayesian Item Response Modeling, Theory and Applications. Statistics for Social and Behavioral Sciences, Springer.

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.

Grubbs, F., 1969. Procedures for detecting outlying observations in samples. Technometrics 11, 1–21. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657, doi:10.1080/00401706.1969.10490657.

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.

Lawson, J., 2015. Design and Analysis of Experiments with R. Chapman and Hall/CRC.

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.

McElreath, R., 2021. Science before statistics: Causal inference. https://www.youtube.com/watch?v=KNPYUVmY3NM. Last accessed 30 April 2024.

McElreath, R., 2024. Statistical rethinking, 2024 course. URL: https://github.com/rmcelreath/stat_rethinking_2024. last accessed 15 March 2025.

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.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modeling. Psychometrika 69, 167–190. doi:10.1007/BF02295939.

Rabe-Hesketh, S., Skrondal, A., Zheng, X., 2007. Multilevel structural equation modeling, in: Lee, Y. (Ed.), Handbook of Latent Variable and Related Models. Elsevier. chapter 10, pp. 209–227.

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.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Interdisciplinary Statistics, Chapman & Hall/CRC Press.

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

Stan Development Team., 2021. Stan Modeling Language Users Guide and Reference Manual, version 2.26. Vienna, Austria. URL: https://mc-stan.org.

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.