MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 2, 305–327

Latently Mediating: A Bayesian Take on Causal Mediation Analysis with Structured Survey Data

Alessandro Varacca

Department of Economics and Social Sciences (DiSes), Università Cattolica del Sacro Cuore, Piacenza (PC), Italy

ABSTRACT

In this paper, we propose a Bayesian causal mediation approach to the analysis of experimental data when both the outcome and the mediator are measured through structured questionnaires based on Likert-scaled inquiries. Our estimation strategy builds upon the error-in-variables literature and, specifically, it leverages Item Response Theory to explicitly model the observed surrogate mediator and outcome measures. We employ their elicited latent counterparts in a simple g-computation algorithm, where we exploit the fundamental identifying assumptions of causal mediation analysis to impute all the relevant counterfactuals and estimate the causal parameters of interest. We finally devise a sensitivity analysis procedure to test the robustness of the proposed methods to the restrictive requirement of mediator's conditional ignorability. We demonstrate the functioning of our proposed methodology through an empirical application using survey data from an online experiment on food purchasing intentions and the effect of different labeling regimes.

KEYWORDS

Causal mediation; item response theory; latent variables; measurement error; Bayesian methods; g-computation

Introduction

Although causal inference (CI) techniques have garnered increasing interest in recent social sciences literature, researchers predominantly concentrate on identification strategies and related estimation methods aimed at solely quantifying the effect of a cause (Gelman & Imbens, 2013). Much less attention is given to a slightly different and more challenging question: *what is the mechanism through which this total effect comes into being*? In other words, is the mere magnitude of the (average) treatment effect the sole goal of CI, or does the interest lie in other related estimands? In some cases, the answer to the second question is yes. Indeed, some studies explore more than just whether an intervention succeeded in improving a target indicator. Rather, uncovering and quantifying the so-called causal mechanisms can answer to more interesting and relevant research questions (Celli, 2022). The set of statistical techniques aimed at investigating causal mechanisms go under the generic name of causal mediation analysis (CMA). Imai et al. (2010a, 2010b) provide a general scope for CMA by defining a causal mechanism as the process where a treatment influences an outcome through an intermediate variable called mediator. Therefore, CMA involves estimating three fundamental quantities: (i) the direct causal effect of the treatment on the outcome; (ii) the indirect causal effect of the mediator on the outcome; (iii) the sum of (i) and (ii) which goes under the name of total causal effect.

CMA holds relevance across various disciplines within the social sciences, offering valuable insights and applications. For example, CMA can be very useful in political science when conducting impact assessment and policy evaluation. Rather than answering the question of *whether* and by *how much* a policy is working, CMA allows to investigate *why* this is the case (Keele et al., 2015). This is achieved by identifying and dissecting the indirect effects of one or more mediators on the desired policy outcomes, all while accounting for the direct effects as well. Empirical assessments provide a versatile avenue for exploring mediators, adopting diverse approaches. Researchers can either construct *ad-hoc* models tailored to specific case studies or anchor their investigations within established theoretical frameworks. For instance, Huber et al. (2017) delved into the

efficacy of alternative tools aimed at boosting employment, while Ma et al. (2020) leveraged the Expectancy Value Theory to explore the role of "perceived value" in enhancing the success of a policy. Similar approaches find resonance in other disciplines as well, where hypotheses concerning *a priori* defined mediators can be tested from theories. In psychology, several notable mediators emerge, such as "self-efficiency" in the Social Cognitive Theory (e.g., Benight & Bandura, 2004), "trust" and "jealousy" in the Attachment Theory (e.g., Toplu-Demirtaş et al., 2022), and "intentions" in the Theory of Planned Behavior (e.g. Sultan et al., 2020). In marketing studies, researchers frequently explore mediators like "usefulness" to reveal the mechanisms of acceptance of new technologies within the Technology Acceptance Model (Blut & Wang, 2020) or 'consumer attitudes metrics' in liking marketing mix activities and sales performances (Hanssens et al., 2014). Management studies offer a fertile ground for mediator identification as well. For example, "capabilities" can mediate the relational performances of buyers/suppliers within the Resource-Based and Relational Views (e. g. Mesquita et al., 2008) and the degree of innovation or corporate social responsibility can play important roles in mediating the relationships between Total Quality Management and firms' performances (e.g. Abbas, 2020; Sadikoglu & Zehir, 2010). The applications of CMA extend beyond these disciplines into various other domains within the social sciences, including sociology, education, and communication sciences, where several other examples could be provided.

A prevalent characteristic found in the literature discussed above is the reliance on survey data, with latent variables often assuming the role of mediators. These latent variables essentially represent unobservable constructs that are assessed through multiple items, ideally derived from or aligned with well-constructed and rigorously validated scales (Boateng et al., 2018). However, resource and time constraints as well as pragmatic choices concerning the feasibility of surveys can often lead to sub-optimal choices in the measurement of latent variables. Moreover, even a latent variable measured from widely accepted and validated scales is not free from the presence of measurement errors. In fact, the concepts of latent variables and measurement error are very closely related as latency can be often framed as an information gap between surrogate indirect indicators and a corresponding unmeasurable trait. In the context of CMA, Hoyle and Kenny (1999), le Cessie et al., (2012), Vander Weele et al. (2012) and Muthén and Asparouhov (2015) have discussed how measurement errors in the moderating variable may lead to severely biased causal effects in a variety of different analytical settings. Although some authors have proposed post-hoc corrections to adjust for such inconsistencies, these can be impractical because they only apply to specific modeling strategies (le Cessie et al., 2012). Therefore, recent works have attempted to directly tackle measurement errors in mediators by extending the standard CMA methods through supplementary statistical models linking these latent components to the corresponding indirect measurements. For example, Albert et al. (2016) employ a generalized structural equation model (GSEM) assuming that the unobserved mediator follows a normal distribution with unit variance and conditional mean functionally related to a set of surrogate covariates. Similarly, Sun et al. (2021) use a linear SEM to inform several unobserved mediators through a large set of highly correlated observable surrogates and incorporate the resulting model into a Bayesian proportional hazard regression. Last, Loh et al. (2020) apply a structural after measurement (SAM—Rosseel & Loh, 2024) approach to a SEM using continuous surrogates to identify a set of latent mediators. The authors eventually resort to a modified g-computation algorithm to calculate the casual effects of interest in case of longitudinal data.

When working with survey data involving polytomous items such as Likert-scaled question, however, the characterization of individual latent traits is typically different. Given the distinctive nature of ordinal responses and because such inquiries are explicitly designed to accurately inform specific characteristics, the literature recognizes two main approaches to measurement error. These are typically referred to as Categorical Factor Analysis (CFA) and Item Factor Analysis (IFA), where the second is a generic label for a larger set of models known as Item-Response Theory (IRT) Models (Van der Linden, 2018). The goal of both these techniques is to come up with suitable statistical machineries to identify and quantify latent characteristics from sets of indirect discrete (either ordinal or multinomial) information sources. Although several authors have established equivalence relationships between CFA and a number of IRT models (Glockner-Rist & Hoijtink, 2003; Kamata & Bauer, 2008; Takane & De Leeuw, 1987), the latter have remained relatively underexplored outside the field of psychomtrics (Thomas, 2019). However, given their fully probabilistic nature, and considering the recent developments in efficient Bayesian estimation techniques (Bürkner, 2019; Furr, 2017; Luo & Jiao, 2018), it is now relatively easy to fit IRT models within complex multilevel statistical structures to control for measurement error in either dependent or independent variables. An early presentation of

this idea is given in Fox and Glas (2003), who proposed to deal with errors in predictors using an IRT normal ogive model (Lord, 1980). Fox (2005) extended their work to accommodate polytomous response data, while recent applications based on this approach include Soregaroli et al. (2022) and Stranieri et al. (2021). This solution can be readily extended to experimental settings where the mediator (or the mediators) is (are) indirectly measured through sets of Likert-valued inquiries. In these cases, one can simply define one or more measurement error models on top of the distribution functions for the outcome and the latent mediator (or mediators) so that the uncertainty in estimating the latter is automatically accounted for when imputing the potential outcomes of interest. This can be seen as a special case of multilevel Bayesian mediation analysis (Bafumi et al., 2005; Yuan & MacKinnon, 2009).

In this paper, we present how CMA can be addressed in presence of latent meditators and outcome variables when these are measured through polytomous ordinal items in a structured survey. In doing so, we also contribute to the literature showing how IRT can be used to address measurement errors in Likert-scaled inquiries devised to approximate well-defined individual latent characteristics. The core of our work focusses on integrating such corrections mechanisms within the non-parametric identification strategy for CMA proposed by Imai et al. (2010a). In particular, we show how Bayesian estimation can be used to impute latent counterfactual mediator and outcome values under sequential ignorability and randomized treatment assignment. Building on our methodological approach, we discuss a simple sensitivity analysis designed to probe the consistency of our estimates to the fundamental assumption of conditional independence between the mediator and the outcome. We finally illustrate the proposed methods through an empirical example that shows how our approach can be easily applied to many real-world experimental data where the relevant variables have been measured through sets of Likert-scaled questions. To encourage users unfamiliar with either CMA or Bayesian methods to pick up these techniques, and to facilitate the practical implementation of the proposed algorithm, we provide the full `R` codes, the corresponding `Stan` programs and a complementary `R markdown` document[1]

The reminder of this paper proceeds as follows: section "Methodology" provides a comprehensive discussion of our methodological approach, section "Simulation study" presents a simulation study addressing the estimation of the relevant parameters, section "Sensitivity analysis" outlines our sensitivity analysis test targeting the residual correlation between the outcome and the mediator, section "Empirical application" discusses an empirical application using real data from a randomized experiment, while section "Conclusions" provides some concluding remark and discussion points for future research.

[1]The replication package for this paper is available online on OSF.io at: https://osf.io/8bz4j/?view_only=9a067a723dbd41b4bbc3fd6b922ddacc

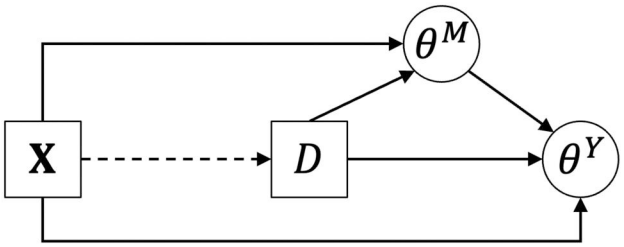

Figure 1. The causal mechanism considered throughout the paper.

Methodology

Identification of the treatment effects

The motivating causal structure that we will refer to throughout the document is the standard setup with one treatment (D), one mediator (θ^M) and one outcome of interest (θ^Y). Unlike the standard notation adopted in most CMA studies, we will refer to the mediator and the outcome using the Greek letter θ because both quantities are latent. We will also assume that the treatment is randomly assigned, as standard practice in experimental setups. Then, given pretreatment confounders $\mathbf{X}$, where $\mathbf{X}$ indicates a $P \times 1$ vector of observed variables measured with no error, our analytical scheme can be depicted as in Figure 1, where the absence of an arrow from $\mathbf{X}$ to D indicates that $\mathbf{X}$ does not affect the treatment propensity because of randomization. Although casual mechanisms have been historically approached using SEMs (Baron & Kenny, 1986), Imai et al. (2010a, 2010b) and Imai et al. (2011) argued that this approach can have several important limitations when it comes to its reliance on (generalized) linear parametric models, untestable assumptions about the error terms and misuses of the exogeneity assumption (Celli, 2022).

Despite Heckman and Pinto's (2015) discussion on how some of these shortcomings can be addressed using econometric methods, more recently an alternative non-parametric way of defining causal effects in CMA has increasingly gained popularity. This literature relies on

identification strategies that, because of their agnosticism with respect to explicit modeling assumptions, are in fact more general than the structural constraints in SEMs. These are formulated using the potential outcomes (PO) framework (Rubin, 1974) and are readily recognizable to anyone familiar with the standard CI literature (Imbens & Rubin, 2015). In brief, the key difference between the SEM and the PO approach to causal mediation analysis lies in how identification is handled. Whereas the former leverages explicit modeling of both the mediator and the outcome through structural equations, the latter defines all the relevant counterfactuals and formulates the necessary assumptions to inform the corresponding causal estimands through sample information. Only then can parametric/semi-parametric restrictions be invoked to estimate these quantities from the data.[2] Although in several circumstances these two approaches tend to overlap and yield the same (or roughly similar) estimators (see, for example, section "Bayesian estimation"), all the methods discussed hereafter are developed within the PO paradigm.

The basic idea of CMA is that there exist two pairs of POs (Imai et al., 2010a, 2010b). Let $j \in \{1, ..., N\}$ be the subscript indicating some individual in a study, then $\theta_j^M(d)$ denotes the potential value of the mediator for j when the treatment is set to $D_j = d$. Similarly, $\theta_j^Y(d, \vartheta^M)$ represents the PO for individual j when $D_j = d$ and the mediator takes value $\theta_j^M = \vartheta^M$. It follows that the measurable mediator and outcome values can be indicated as $\theta_j^M = \theta_j^M(D_j)$ and $\theta_j^Y = \theta_j^Y\left[D_j, \theta_j^M(D_j)\right]$, respectively. Clearly, of both expressions only one can be potentially informed through observed data. Under no interference and no unobserved alternative versions of the treatment (i.e.: the so-called Stable Unit Treatment Value Assumption—SUTVA), Imai et al. (2010a) use these quantities to define several causal effects, the first one being the Natural Indirect Effect (NIE). Assuming a binary treatment, the latter can be defined as:

$$\delta_j(d) = \theta_j^Y\left[d, \theta_j^M(D_j = 1)\right] - \theta_j^Y\left[d, \theta_j^M(D_j = 0)\right]$$

where $d \in \{0, 1\}$. The NIE expresses the change in θ_j^Y had one changed the mediator value from the control status, $\theta_j^M(D_j = 0)$, to the treatment status, $\theta_j^M(D_j = 1)$, while holding the treatment constant at d. The NIE can be employed to construct the average causal mediation effect (ACME), which represents the first target estimand in CMA:

$$\delta(d) = \mathbb{E}\left[\delta_j(d)\right]$$

Other important quantities are the Natural Direct Effect (NDE), which indicates the treatment effect from setting the mediator to the potential value that would occur under treatment d, and the Average Natural Direct Effect (ANDE), i.e., the second estimand of interest:

$$\zeta_j(d) = \theta_j^Y\left[D_j = 1, \theta_j^M(d)\right] - \theta_j^Y\left[D_j = 0, \theta_j^M(d)\right]$$
$$\zeta(d) = \mathbb{E}\left[\zeta_j(d)\right]$$

Consistently with Vander Weele and Vansteelandt (2009), Imai et al. (2010a, 2010b), Imai et al. (2011), Park and Kaplan (2015) and Celli (2022), both the ACME and the ANDE are non-parametrically identified under the following assumptions: (i) $0 < \Pr(D_j = 1|\mathbf{X}_j = \boldsymbol{x}) < 1$ (i.e., every individual has a strictly positive probability to receive the treatment); (ii) $\Pr\left[\theta_j^M(d)|D_j = d, \mathbf{X}_j = \boldsymbol{x}\right] > 0$ (i.e., conditional on pretreatment covariates $\mathbf{X}_j$ — where pretreatment means common causes of the treatment, mediator, and/or outcome that are measured before treatment — the mediator is not a deterministic function of the treatment); (iii) $\left\{\theta_j^Y(d', \vartheta^M), \theta_j^M(d)\right\} \perp D_j|\mathbf{X}_j = \boldsymbol{x}$ (i.e., the two PO are independent of the treatment conditional on $\mathbf{X}_j$); (iv) $\theta_j^Y(d', \vartheta^M) \perp \theta_j^M(d)|\mathbf{X}_j = \boldsymbol{x}, D_j = d$ (i.e., the mediator is ignorable—the potential outcome is independent of the potential mediator value—conditional on D_j and $\mathbf{X}_j$). Whereas randomizing the treatment ensures that assumptions (i) through (iii) are met, condition (iv) is more restrictive and more difficult to attain (Heckman & Pinto, 2015). In short, assumption (iv) makes sure that there are no treatment effects on any variable (whether observable or unobservable) that would confound the mediator-outcome relationship. In other words, the treatment effects are identified if there are no unmeasured pretreatment variables and no posttreatment confounders (i.e., common causes of the mediator and outcome that are measured after treatment). Since measuring the mediator and the outcome often occurs later than the exposure to the treatment, the identification of any mediation effect often struggles in face of this restriction. However, as discussed by Vander Weele and Vansteelandt (2009) there are cases in which assumption (iv) could prove more credible. For example, when the mediator is measured shortly after the treatment is administered, it

[2]However, in some experimental settings, these further assumptions may not be necessary.

is less likely that there might be post-treatment confounding variables. While in many situations this reads as a strong requirement, it may be less problematic for experimental settings where surrogate data for the mediator are collected through surveys or questionnaires by the end of the experiment.

Given assumptions (i) through (iv), we can write down an analytical expression for the conditional distribution of the POs (Imai et al., 2010b):

$$f\left(\theta_j^Y\left[d,\theta_j^M(d')\right]|\mathbf{X}_j=\boldsymbol{x}\right) = \int f\left(\theta_j^Y|\theta_j^M=\vartheta^M,D_j=d,\mathbf{X}_j=\boldsymbol{x}\right)dF_{\theta_j^M|D_j=d',\mathbf{X}_j=\boldsymbol{x}}(\vartheta^M) \tag{1}$$

where $f(\cdot)$ defines a generic probability density or mass function, while $d,d'\in\{0,1\}$. Equation (1) generalizes the results in Imai et al. (2010a), who provide similar results in terms of conditional expectation rather than probability distributions. Finally, while the ACME and the ANDE can be simply obtained by averaging (1) with the respect to the empirical distribution of $\mathbf{X}_j$ (i.e.: a simple average – Li et al., 2023) and plugging in the resulting values into the corresponding formulas, we prefer to work with Equation (1) directly as it provides for a better understanding of algorithm 1 in section "Bayesian estimation".

Measurement error model

As discussed in section "Measurement error model", the latent quantities denoted as θ^M and θ^Y are not directly observable. Rather, we can only attempt to obtain indirect information using surrogate measures, such as Likert-valued statements and questions. Since quantifying the causal effects defined in the previous section hinges on these latent features, our modeling exercise requires a coherent methodological framework connecting individual responses to the corresponding hidden attributes. As discussed in the introductory section, IRT models stand out as a compelling set of empirical tools that can inform both individual-specific and question-specific traits. Although these statistical techniques have long been solely regarded as psychrometric methods, they are steadily gaining popularity across multiple disciplines (Thomas, 2019; Yamashita, 2022), including applications to measurement error modeling (Fox, 2005; Fox & Glas, 2003; Soregaroli et al., 2022; Stranieri et al., 2021).

When participants to an experiment are asked to rate Likert-valued statements, where the lowest and highest values correspond to strong disagreement and strong agreement, respectively, the Rating Scale Model (RSM—Andrich, 2005, 2016) provides a sensible probabilistic framework to link response scores to two abstract components, i.e.: "item difficulty" and "person ability" (Wright, 1977). Unlike more complex IRT models, the standard RSM assumes that the distance between item difficulty values remains constant across all items. This assumption is reasonable when item responses are obtained using a fixed set of behavioral thresholds (e.g., Likert-type scales). Additionally, because all coefficients in a RSM represent positions on an underlying latent variable, they enable objective comparisons of individuals and items.

Mathematically, the standard RSM model can be formulated as follows (Andrich, 2005, 2016; Van der Linden, 2018: Chapter 5):

$$P_{j,r,q} = \Pr\left(r_{j,q}=r|\theta_j,\beta_q,\boldsymbol{\kappa}\right) = \frac{\exp\left\{\sum_{\ell=1}^{r}\left(\theta_j-\beta_q-\kappa_\ell\right)\right\}}{1+\sum_{k=1}^{\mathcal{R}}\exp\left\{\sum_{\ell=1}^{k}\left(\theta_j-\beta_q-\kappa_\ell\right)\right\}} \tag{2}$$

where $q\in\{1,\ldots,Q\}$ indicates the q^{th} item (i.e.: statement or question), j represents the j^{th} person (i.e.: the individual or respondent), $r\in\{1,2,\ldots,\mathcal{R}\}$ is the response given by person j to any item q, $P_{j,r,q}$ indicates the probability that person j answers r to item q, β_q stands for the q^{th} item's difficulty, $\boldsymbol{\kappa}$ is a $\mathcal{R}$-vector of thresholds $\boldsymbol{\kappa}=[\kappa_1,\ldots,\kappa_{\mathcal{R}}]$, and θ_j refers to the j^{th} person's ability. A popular extension to the model in Equation (2) is the generalized RSM (GRSM – Muraki, 1992), where the parameter set now includes a discrimination parameter, α:

$$P^{\alpha}_{j,r,q} = \Pr\left(r_{j,q}=r|\theta_j,\alpha_q,\beta_q,\boldsymbol{\kappa}\right) = \frac{\exp\left\{\sum_{\ell=1}^{r}\left(\alpha_q\theta_j-\beta_q-\kappa_\ell\right)\right\}}{1+\sum_{k=1}^{\mathcal{R}}\exp\left\{\sum_{\ell=1}^{k}\left(\alpha_q\theta_j-\beta_q-\kappa_\ell\right)\right\}} \tag{3}$$

These additional coefficients are proportional to the strength of the relationship between the latent individual characteristic and the chances of choosing option r. Therefore, positive values of α_q correspond to statements where individual with higher θ_j will choose r with higher probability, and vice versa (Bafumi et al., 2005). For simplicity, we re-formulate Equation (3) more compactly as:

$$r_{j,q}|\theta_j,\alpha_q,\beta_q,\boldsymbol{\kappa}\sim\text{GRSM}\left(\theta_j,\alpha_q,\beta_q,\boldsymbol{\kappa}\right) \tag{4}$$

Since we are interested in modeling two latent characteristics, we extend Equation (4) to:

$$r^M_{j,q}|\theta^M_j,\alpha^M_q,\beta^M_q,\boldsymbol{\kappa}^M \sim \text{GRSM}\left(\theta^M_j,\alpha^M_q,\beta^M_q,\boldsymbol{\kappa}^M\right)$$
$$r^Y_{j,q}|\theta^Y_j,\alpha^Y_q,\beta^Y_q,\boldsymbol{\kappa}^Y \sim \text{GRSM}\left(\theta^Y_j,\alpha^Y_q,\beta^Y_q,\boldsymbol{\kappa}^Y\right)$$

Before describing our estimation procedure, two important clarifications need to be made. First, the standard nomenclature of the RSM does not seem to entirely fit our analytical framework. In fact, unlike most applications in psychometric analysis, our primary objective does not involve modeling respondents' abilities or explicitly correcting for item difficulty. Instead, we exploit this probabilistic construct to map sets of Likert-valued statements onto continuous measures that share a common support. Therefore, to align the IRT terminology with our analytical framework, we will henceforth refer to the subscripts q and j as "question" and "respondent", respectively, instead of "item" and "person". Second, model (3) is clearly statistically not identified. On the one hand, adding a constant to θ_j, β_q and κ_ℓ does not change how the model predicts $P_{j,r,q}$ (this problem is typically called additive aliasing). Consequently, as suggested by Bafumi et al. (2005), Gelman and Hill (2006, chapter 14.3), Furr (2017), and Luo and Jiao (2018), we impose two simple restrictions on model (3): (i) we constrain both the last statement coefficient β_Q and the last threshold $\kappa_\mathcal{R}$ to be the negative sum of the other statement coefficients and thresholds, respectively (i.e.: $\beta_Q = -\sum_{q=1}^{Q-1}\beta_q$ and $\kappa_\mathcal{R} = -\sum_{r=1}^{\mathcal{R}-1}\kappa_r$, so that these terms average to zero); (ii) we specify a zero-mean weakly informative prior distribution for both the statement coefficients and the thresholds to allow identifying the mean-function parameters of θ_j (See sections "Bayesian estimation" and "Simulation study"). Besides additive aliasing, model (3) also suffers from two additional forms of indeterminacy known as multiplicative aliasing and reflection invariance. These can be worked out by imposing two additional model-identifying restrictions (Bafumi et al. 2005; Fox, 2005; Fujimoto & Neugebauer, 2020; Furr, 2017): (i) placing a log-normal prior on α_q, thereby restricting the sign of the discriminating parameters to positive values;[3] (ii) fixing the prior variance of θ_j to a constant value (typically $\sigma_\theta = 1$). The latter also helps in CMA applications in that it ensures that both the latent outcome and the latent mediator share the same prior scale. Not only is this a sensible assumption when both these quantities are measured through Likert-scaled questions that share the same minimum and maximum scores, but it is also a common choice in the error-in-variables literature (Albert et al., 2016).

Finally, we would like to stress that the GRSM can be either replaced by other slightly different IRT models such as the Partial Credit Model (PCM) and the Graded Response Model (GRM), or extended in several ways, including multilevel, nested (Böckenholt, 2012) as well as multivariate specifications (Fujimoto & Neugebauer, 2020). Although there exist model selection techniques to determine which formulation provides a better fit for the data (see, for example, Fox, 2005 or Luo & Jiao, 2018, for a survey of such methods), choosing between these alternatives ultimately hinges on the structure and purposes of the survey, particularly when it comes to the statements' design. However, since an exhaustive treatment of such techniques (as well as a comprehensive discussion of the many IRT modeling choices) is outside the scope of this paper, we limit our discussion to two complementary model checking strategies when discussing our empirical application in section "Empirical application".

As we illustrate in the following section, our estimation strategy hinges on the joint distribution of $r_{q,j}$, θ_j, β_q and $\boldsymbol{\kappa}$ as well as the remaining parameters for the conditional mean of θ_j. This multivariate probability function represents the full Bayesian model (Betancourt, 2020) and can be decomposed into two fundamental terms known as likelihood and prior. Using this terminology, Equation (3) represents the likelihood of the observed data, while θ_j, β_q and $\boldsymbol{\kappa}$ require their own prior distributions (Gelman et al., 2013). As recommended in Furr (2017), Luo and Jiao (2018) and Bürkner (2019), we choose weakly informative priors:[4]

$$\begin{aligned}
\kappa^z_\ell &\sim \mathcal{N}(0,3) \text{ for all } \ell \in \{1,\ldots,\mathcal{R}\}\\
\beta^z_q &\sim \mathcal{N}(0,3) \text{ for all } q \in \{1,\ldots,Q^z\}\\
\theta^z_j &\sim \mathcal{N}\left(\mu^z_j,\sigma^z\right) \text{ for all } j \in \{1,\ldots,N\}\\
\alpha^z_q &\sim \log\mathcal{N}(1,1) \text{ for all } q \in \{1,\ldots,Q^z\}
\end{aligned} \tag{5}$$

where $z \in \{M,Y\}$, Q^z is the number of questions or statements for z, $\sigma^z = 1$ following the identifying restrictions defined above, while μ^z_j are defined in the following section.

[3]This is sometimes regarded as a restrictive assumption in that discrimination is limited to the relative magnitude of α, rather than its sign and magnitude. However, this assumption remains necessary to statistically identify model (3) and it is standard practice in the (Bayesian) estimation of generalized IRT models. One alternative and less limiting approach would be to manually restrict the sign of each α based on individual characteristics, as discussed in Bafumi et al. (2005). However, since this approach hinges both on the nature of the data and the problem at hand, we do not discuss it in our work.

[4]In the empirical application discussed in section "Empirical application", we also conduct a small sensitivity analysis where we nudge the all the priors' coefficients in Equation (5) to test the stability of our estimates. Although these tests show that our results are robust to limited variations in the priors' parameters, we stress that, in general, prior influence tends to decrease with sample size (Gelman et al., 2013, p. 355).

Bayesian estimation

Following the general approach discussed in Imai et al. (2010b), Park and Kaplan (2015), Albert et al. (2016) and Loh et al. (2020), we devise a simple Bayesian g-computation algorithm[5] coupled with a tractable specification for all the terms in Equation (1). The first stage of our estimation strategy consists in defining a suitable (parametric, non-parametric or semi-parametric) model for $f(\theta_j^M|D_j, \mathbf{X}_j = \boldsymbol{x}_j)$ and $f(\theta_j^Y|\theta_j^M = \vartheta^M, D_j, \mathbf{X}_j = \boldsymbol{x}_j)$. We begin by assuming that both θ_j^M and θ_j^Y are normally distributed. Adopting a Gaussian model for the latent mediator and outcome is not only convenient in that, under linear conditional means, the model becomes immediately interpretable as a standard linear SEM (Imai et al., 2010a), but it also remains general enough since latent variables are typically given this type of distribution (Albert et al., 2016). Moreover, as discussed in section "Measurement error model", the normal distribution is also a reasonable and widely adopted prior for the respondent coefficients of the GRSM model which, likewise, represent latent characteristics (Fox, 2005; Fox & Glas, 2003). The overlap between the distributional assumptions for the outcome and the mediator, and the prior choices for the measurement error model is very important for the functioning of the algorithm discussed below. The reason lies in the two-step approach of g-computation, where step one takes care of estimating the parameters of $f(\theta_j^M|D_j, \mathbf{X}_j = \boldsymbol{x})$ and $f(\theta_j^Y|\theta_j^M = \vartheta^M, D_j, \mathbf{X}_j = \boldsymbol{x})$, while step two makes use of these estimates to simulate the POs through the formula in Equation (1) (Snowden et al., 2011).

Given a normal model for both the mediator and the outcome, characterizing the conditional distribution of θ_j^M and θ_j^Y requires specifying μ_j^M and μ_j^Y. This essentially corresponds to constructing regression equations for the two latent quantities of interest. As discussed in Imai et al. (2010b) and Preacher (2015), setting μ_j^M or μ_j^Y to linear predictors under either an identity or any other canonical link gives rise to a linear or generalized linear SEM. However, whereas tackling estimation and identification in such cases can require rather different approaches, all the procedures described here can potentially accommodate different distribution functions as well as non-parametric or semi-parametric conditional expectation functions.

For the sake of illustration, we hereafter assume that both μ_j^M and μ_j^Y are linear in D_j, $\mathbf{X}_j$ and θ_j^M (for an extension using a semi-parametric specification see Kim et al., 2018). We also include an interaction (i.e.: treatment heterogeneity) between the mediator and the treatment in the outcome predictor:

$$\begin{aligned}\mu_j^M &= \Lambda_0^M + \Lambda_D^M D_j + \mathbf{X}_j'\boldsymbol{\Lambda}_{\mathbf{X}}^M \\ \mu_j^Y &= \Lambda_0^Y + \Lambda_D^Y D_j + \mathbf{X}_j'\boldsymbol{\Lambda}_{\mathbf{X}}^Y + \Lambda_M^Y\theta_j^M + \Lambda_{M,D}^Y D_j\theta_j^M\end{aligned} \tag{6}$$

where $\boldsymbol{\Lambda}_{\mathbf{X}}^M$ and $\boldsymbol{\Lambda}_{\mathbf{X}}^Y$ indicate $P \times 1$ vectors of regression coefficients for the $P \times 1$ covariate set $\mathbf{X}_j$ in the latent moderator and outcome equation, respectively. Combining Equations (1) and (6) under normally distributed θ_j^M and θ_j^Y yields:

$$\begin{aligned}&f\left(\theta_j^M|D_j, \mathbf{X}_j = \boldsymbol{x}\right) = \mathcal{N}\left(\Lambda_0^M + \Lambda_D D_j + \boldsymbol{x}_j'\boldsymbol{\Lambda}_{\mathbf{X}}^M, \sigma^M\right) \\ &f\left(\theta_j^Y|\theta_j^M = \vartheta^M, D_j, \mathbf{X}_j = \boldsymbol{x}\right) \\ &= \mathcal{N}\left(\Lambda_0^Y + \Lambda_D^Y D_j + \boldsymbol{x}_j'\boldsymbol{\Lambda}_{\mathbf{X}}^Y + \Lambda_M^Y\vartheta_j^M + \Lambda_{M,D}^Y D_j\vartheta_j^M, \sigma^Y\right)\end{aligned} \tag{7}$$

Equation (7) also completes the prior specification in Equation (5), to which we add the following weakly informative priors for all the "slope" parameters:

$$\begin{aligned}&\Lambda_0^M, \Lambda_0^Y, \Lambda_D^M, \Lambda_D^Y, \Lambda_M^Y, \Lambda_{M,D}^Y \sim \mathcal{N}(0,1) \\ &\Lambda_{p,\mathbf{X}}^M, \Lambda_{p,\mathbf{X}}^Y \sim \mathcal{N}(0,1) \text{ for all } p \in \{1, ..., P\}\end{aligned} \tag{8}$$

where $\Lambda_{p,\mathbf{X}}^M \in \boldsymbol{\Lambda}_{\mathbf{X}}^M$ and $\Lambda_{p,\mathbf{X}}^Y \in \boldsymbol{\Lambda}_{\mathbf{X}}^Y$. These distributional choices follow the general principle of avoiding flat uninformative priors that, in case of poorly informative likelihoods, can cause severe mixing problems in Markov chain Monte Carlo (MCMC) sampling algorithms (Gelman et al., 2017; Lemoine, 2019; Park & Kaplan, 2015; Smid et al., 2020). Specifically, since $\sigma^M = \sigma^Y = 1$ and provided that all the binary and continuous variables in $\mathbf{X}_j$ have been centered or standardized, respectively, the $\mathcal{N}(0,1)$ represents a good default for linear regression models (Gelman et al., 2008; Ghosh et al., 2018). In the simulation exercise discussed in section "Sensitivity analysis", however, we also test our model against wider priors for the Λ parameters.[6] Equations (3), (5), (7) and (8),

[5]Notice that, unlike the approaches described in Imbens and Rubin (2015, p. 150), Ding and Li (2018, p. 223) and Li et al. (2023, p. 6), our statistical model does not attempt to explicitly define a joint distribution for the potential outcomes and the model parameters. Rather, the full Bayesian model only serves to combine the modelling steps in Equations (3) and (7) with the identification strategy described in Section "Measurement error model". In other words, whereas the estimation stage is fully Bayesian, identification follows a different conceptual path.

[6]All the results discussed in the empirical application presented in Section "Empirical application" also show stability to different choices of prior for Λ such as $\mathcal{N}(0, 2.5)$ and $\mathcal{N}(0, 5)$.

make up the full Bayesian model from which we can sample *via* MCMC methods (see Appendices[7] A1 and A2 for details, supplementary material).

Suppose now that S uncorrelated samples were successfully collected from the joint posterior distribution of the parameters in Equation (7). Then, calculating the ACME and the ANDE involves: (i) plugging such estimates in the corresponding formulas to approximate the conditional distributions of θ_j^M and θ_j^Y; (ii) imputing the POs *via* Monte Carlo integration of Equation (1). Park and Kaplan (2015), who provide a Bayesian alternative to the algorithm proposed by Imai et al. (2010b), illustrate one way of computing these quantities using the posterior distribution of the coefficients in Equation (6). Their idea essentially resides in sampling POs from the conditional means $\mathbb{E}[\theta_j^M|D_j,\mathbf{X}_j=\boldsymbol{x}]=\mu_j^M$ and $\mathbb{E}[\theta_j^Y|\theta_j^M=\vartheta^M,D_j,\mathbf{X}_j=\boldsymbol{x}]=\mu_j^Y$, and use them to directly quantify all the causal estimands (see Imai et al., 2010a: Theorem 1). Although the authors show that this procedure yields unbiased estimates of $\delta(d)$ and $\zeta(d)$, it does not make full use of the distributions in Equation (7) since the potential values of both the mediator and the outcome are only generated using their expected values. In this respect, imputation through the conditional mean disregards the variance of θ_j^M and θ_j^Y, i.e.: σ^M and σ^Y, thereby resulting in overconfident (i.e., narrower) credible intervals for the resulting predictions. This can be especially limiting when θ_j^M and θ_j^Y are latent, as incorporating the uncertainty in these unobservable respondent characteristics is an essential feature of the measurement error model. To this extent, imputing potential values using $f(\theta_j^M|D_j,\mathbf{X}_j=\boldsymbol{x})$ and $f(\theta_j^Y|\theta_j^M=\vartheta^M,D_j,\mathbf{X}_j=\boldsymbol{x})$ is consistent with the problem in Equation (1), where the integration is defined with respect to the whole conditional distributions of the counterfactuals (see Imai et al., 2010b: Theorem 1). A similar argument is sustained by Keil et al. (2018), who devise a g-computation formula based on the posterior predictive distribution (PPD) of the POs in a standard binary treatment setup.

To provide more conservative estimates of $\delta(d)$ and $\zeta(d)$, we propose the procedure is stylized in algorithm 1, where the set $\{\tilde{\vartheta}_j^{Y,(s)}(d,d')|s\in 1,\ldots,S\}$, for all $j\in\{1,\ldots,N\}$, represents[8] S draws from $f(\theta_j^Y[d,\theta_j^M(d')]|\mathbf{X}_j=\boldsymbol{x})$, while $\tilde{\Lambda}_0^{M,(s)}$, $\tilde{\Lambda}_D^{M,(s)}$, $\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{M,(s)}$, $\tilde{\Lambda}_0^{Y,(s)}$, $\tilde{\Lambda}_D^{Y,(s)}$, $\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{Y,(s)}$, $\tilde{\Lambda}_M^{Y,(s)}$, $\tilde{\Lambda}_{M,D}^{Y,(s)}$, $\tilde{\sigma}^{M,(s)}$ and $\tilde{\sigma}^{Y,(s)}$ indicate samples from the posterior distributions of the parameters in Equation (7),[9] and steps (1) through (3) make use of the assumption that θ_j^M and θ_j^Y are normally distributed.

Algorithm 1: Bayesian g-computation

For all $s\in\{1,\ldots,S\}$ do:

For all $j\in\{1,\ldots,N\}$ do:

(1) Sample from $\mathcal{N}\left(\tilde{\Lambda}_0^{M,(s)}+\tilde{\Lambda}_D^{M,(s)}d'+\boldsymbol{x}_j'\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{M,(s)},\ \tilde{\sigma}^{M,(s)}\right)$ and obtain $\tilde{\vartheta}_j^{M,(s)}(d')$

(2) Sample from $\mathcal{N}\left(\tilde{\Lambda}_0^{M,(s)}+\tilde{\Lambda}_D^{M,(s)}d+\boldsymbol{x}_j'\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{M,(s)},\ \tilde{\sigma}^{M,(s)}\right)$ and obtain $\tilde{\vartheta}_j^{M,(s)}(d)$

(3.1) Sample from $\mathcal{N}\left(\tilde{\Lambda}_0^{Y,(s)}+\tilde{\Lambda}_D^{Y,(s)}d'+\boldsymbol{x}_j'\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{Y,(s)}+\tilde{\Lambda}_M^{Y,(s)}\tilde{\vartheta}_j^{M,(s)}(d)+\tilde{\Lambda}_{M,D}^{Y,(s)}d'\tilde{\vartheta}_j^{M,(s)}(d),\ \tilde{\sigma}^{Y,(s)}\right)$ and obtain $\tilde{\vartheta}_j^{Y,(s)}(d',d)$

(3.2) Sample from $\mathcal{N}\left(\tilde{\Lambda}_0^{Y,(s)}+\tilde{\Lambda}_D^{Y,(s)}d+\boldsymbol{x}_j'\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{Y,(s)}+\tilde{\Lambda}_M^{Y,(s)}\tilde{\vartheta}_j^{M,(s)}(d')+\tilde{\Lambda}_{M,D}^{Y,(s)}d\tilde{\vartheta}_j^{M,(s)}(d'),\ \tilde{\sigma}^{Y,(s)}\right)$ and obtain $\tilde{\vartheta}_j^{Y,(s)}(d,d')$

(3.3) Sample from $\mathcal{N}\left(\tilde{\Lambda}_0^{Y,(s)}+\tilde{\Lambda}_D^{Y,(s)}d+\boldsymbol{x}_j'\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{Y,(s)}+\tilde{\Lambda}_M^{Y,(s)}\tilde{\vartheta}_j^{M,(s)}(d')+\tilde{\Lambda}_{M,D}^{Y,(s)}d\tilde{\vartheta}_j^{M,(s)}(d'),\ \tilde{\sigma}^{Y,(s)}\right)$ and obtain $\tilde{\vartheta}_j^{Y,(s)}(d,d)$

(3.4) Sample from $\mathcal{N}\left(\tilde{\Lambda}_0^{Y,(s)}+\tilde{\Lambda}_D^{Y,(s)}d'+\boldsymbol{x}_j'\tilde{\boldsymbol{\Lambda}}_{\mathbf{X}}^{Y,(s)}+\tilde{\Lambda}_M^{Y,(s)}\tilde{\vartheta}_j^{M,(s)}(d')+\tilde{\Lambda}_{M,D}^{Y,(s)}d'\tilde{\vartheta}_j^{M,(s)}(d'),\ \tilde{\sigma}^{Y,(s)}\right)$ and obtain $\tilde{\vartheta}_j^{Y,(s)}(d',d')$

(4.1) Compute $\tilde{\delta}_G^{(s)}(d)=N^{-1}\sum_{j=1}^N\left[\tilde{\vartheta}_j^{Y,(s)}(d,1)\right]-N^{-1}\sum_{j=1}^N\left[\tilde{\vartheta}_j^{Y,(s)}(d,0)\right]$

(4.2) Compute $\tilde{\zeta}_G^{(s)}(d)=N^{-1}\sum_{j=1}^N\left[\tilde{\vartheta}_j^{Y,(s)}(1,d)\right]-N^{-1}\sum_{j=1}^N\left[\tilde{\vartheta}_j^{Y,(s)}(0,d)\right]$

Even though the posterior quantities produced by algorithm 1 will be approximately centered around the values obtained with the methods discussed in Park and Kaplan (2015), the former will entail more uncertainty, resulting in larger credible intervals. How wider these intervals will be is going to depend on the index of dispersion of θ_j^M and θ_j^Y

Finally, it is important to stress that, under linear μ_j^M and μ_j^Y and normal priors on θ_j^M and θ_j^Y, one

[7]Appendix A1 and A2 can be found in the supplementary material available online on OSF.io at: https://osf.io/8bz4j/?view_only=9a067a723dbd41b4bbc3fd6b922ddacc.

[8]Conditioning on $\mathbf{X}_j=\boldsymbol{x}$ is omitted to simplify notation.

[9]Notice that, using the GRSM as a measurement error model, the variance of θ_j^M and θ_j^Y is fixed to a constant, so $\tilde{\sigma}^{z,(s)}=\sigma^z$ for all $s\in\{1,\ldots,S\}$.

could directly use the posterior distribution of Λ_0^M, Λ_D^M, Λ_M^Y, $\Lambda_{M,D}^Y$, Λ_D^Y and $\mathbf{\Lambda}_{\mathbf{X}}^M$ to quantify the two estimands of interest *via* coefficients' multiplication (Imai et al., 2010b):

$$\tilde{\delta}^{(s)}(d) = \tilde{\Lambda}_D^{M,(s)}\left(\tilde{\Lambda}_M^{Y,(s)} + \tilde{\Lambda}_{M,D}^{Y,(s)} d\right)$$
$$\tilde{\zeta}^{(s)}(d) = \tilde{\Lambda}_D^{Y,(s)} + \tilde{\Lambda}_{M,D}^{Y,(s)}\left(\tilde{\Lambda}_0^{M,(s)} + \tilde{\Lambda}_D^{M,(s)} d + \bar{\boldsymbol{x}}'\tilde{\mathbf{\Lambda}}_{\mathbf{X}}^{M,(s)}\right)$$

where $\bar{\boldsymbol{x}} = N^{-1}\sum_{j=1}^N \boldsymbol{x}_j$. Notice that, under standardized covariates, both $\Lambda_0^M = 0$ and $\bar{\boldsymbol{x}} = 0$, so $\tilde{\Lambda}_0^{M,(s)}$ and the $\bar{\boldsymbol{x}}'\tilde{\mathbf{\Lambda}}_{\mathbf{X}}^{M,(s)}$ term drop out of the equation for $\tilde{\zeta}^{(s)}(d)$. The multiplication approach is commonly used in Bayesian SEM (Lawson et al., 2023; McCandless & Somers, 2019; Yuan & MacKinnon, 2009). However, unlike $\tilde{\delta}_G^{(s)}(d)$ and $\tilde{\zeta}_G^{(s)}(d)$, the results obtained in this way do not marginalize over σ^M and σ^Y and, because of that, they will perfectly align with the estimates obtained through imputation the conditional means-based imputation. The empirical application presented in section "Empirical application" provides a comparison between $\delta(d)$ and $\zeta(d)$ when these are estimated using either the marginal or the conditional predictive distribution of the latent outcome and the latent mediator.

Simulation study

The non-parametric identification of the causal estimands discussed in section "Identification of the treatment effects" clearly hinges on the parameters of Equation (7) being themselves statistically identified.[10] Although section "Identification of the treatment effects" briefly mentions that the restrictions applied to the discrimination, question and threshold parameters in the GRSM model allow identifying the slope coefficients in the mean function of θ_j^M and θ_j^Y, the fact that μ_j^Y also includes latent regressors in the form of θ_j^M may further complicate the estimation of these parameters. In other words, had the coefficients Λ_0^M, Λ_D^M and $\mathbf{\Lambda}_{\mathbf{X}}^M$ not been identified though the constraints discussed above, the posterior distribution of θ_j^Y would likely be biased and so would the estimates of $\Lambda_0^Y, \Lambda_D^Y, \Lambda_M^Y, \Lambda_{M,D}^Y$ and $\mathbf{\Lambda}_{\mathbf{X}}^Y$. Since the g-computation algorithm presented in section "Bayesian estimation" (as well as the coefficient multiplication method) simulates counterfactuals through the posterior distribution of the parameters above (and also σ^M, σ^Y in the basic RSM specification), these need to be correctly estimated when sampling from the full Bayesian model implied by Equations (4), (5) and (8). In this section, we discuss an extensive simulation study where we assess the approximated posterior distribution of the coefficients in μ_j^M and μ_j^Y as well as the corresponding causal effects. To do so, we exploit the class of simulation-based calibration methods developed by Cook et al. (2006), Talts et al. (2018) and Schad et al. (2021). The idea is to provide a fully Bayesian implementation of the standard routines designed to validate frequentist estimators. The main difference between the two approaches is that, in the Bayesian case, the target parameters are not fixed quantities, but they have their own distribution. Therefore, our evaluation exploits the properties of the full Bayesian model to address the coherence between a range of '*true*' parameter values generated through the prior distributions and the resulting posterior (Betancourt, 2020; Talts et al., 2018).

Following the notation introduced and discussed in Appendix A1 (supplementary material), let $\mathbf{\Lambda}^M = [\Lambda_0^M, \Lambda_D^M, \mathbf{\Lambda}_{\mathbf{X}}^M]$ be a $(P+2)\times 1$ coefficient vector, call $\mathbf{\Lambda}^Y$ the $(P+4)\times 1$ parameter vector $[\Lambda_0^Y, \Lambda_D^Y, \mathbf{\Lambda}_{\mathbf{X}}^Y, \Lambda_M^Y, \Lambda_{M,D}^Y]$, and let $L = (P+2)+(P+4)$. Define also the L-dimensional vector $\mathbf{\Lambda} = [\mathbf{\Lambda}^M, \mathbf{\Lambda}^Y]$ and let $f(\mathbf{\Lambda})$ be the corresponding (joint) prior distribution. With a slight abuse of notation, define the sets of remaining (complement) GRSM parameters as $\mathbf{\Lambda}^{M,c}$ and $\mathbf{\Lambda}^{Y,c}$, respectively, which we also concatenate into $\mathbf{\Lambda}^c = [\mathbf{\Lambda}^{M,c}, \mathbf{\Lambda}^{Y,c}]$ with (joint) prior $f(\mathbf{\Lambda}^c)$. Furthermore, indicate with $\boldsymbol{r}_j^z$ the $Q^z\times 1$ vector $\boldsymbol{r}_j^z = [r_{j,1}^z, ..., r_{j,Q^z}^z]$. We can aggregate the latter over j into a $NQ^z\times 1$ array $\boldsymbol{r}^z = [\boldsymbol{r}_1^z, ..., \boldsymbol{r}_N^z]$ and form the full set of observed responses $\boldsymbol{r} = [\boldsymbol{r}^M, \boldsymbol{r}^Y]$ with joint likelihood function $f(\boldsymbol{r}|\mathbf{\Lambda}, \mathbf{\Lambda}^c) = f(\boldsymbol{r}^M|\mathbf{\Lambda}^M, \mathbf{\Lambda}^{M,c})f(\boldsymbol{r}^Y|\mathbf{\Lambda}^Y, \mathbf{\Lambda}^{Y,c})$. Finally let $\mathbf{\Lambda}^*$, $\mathbf{\Lambda}^{c,*}$ represent '*true*' parameter values drawn from the joint prior $f(\mathbf{\Lambda}, \mathbf{\Lambda}^c) = f(\mathbf{\Lambda})f(\mathbf{\Lambda}^c)$ using the configurations in Equations (5) and (8), and consider observations $\boldsymbol{r}^*$ obtained from $f(\boldsymbol{r}|\mathbf{\Lambda}^*, \mathbf{\Lambda}^{c,*})$. Then, the tuple $[\boldsymbol{r}^*, \mathbf{\Lambda}^*]$ represents a draw from the joint distribution $f(\boldsymbol{r}, \mathbf{\Lambda}) \propto f(\mathbf{\Lambda}|\boldsymbol{r})$, implying that $\mathbf{\Lambda}^*$ is itself a draw from the posterior distribution $f(\mathbf{\Lambda}|\boldsymbol{r})$. Therefore, given a $S\times L$ matrix of posterior samples $\tilde{\mathbf{\Lambda}} = [\tilde{\mathbf{\Lambda}}^{(1)}, ..., \tilde{\mathbf{\Lambda}}^{(S)}]'$ obtained by fitting $f(\mathbf{\Lambda}|\boldsymbol{r}^*)$, the marginal distribution of $\mathbf{\Lambda}^*$ should be the same as that of any $\tilde{\mathbf{\Lambda}}^{(s)} \in \tilde{\mathbf{\Lambda}}$. If not,

[10] Notice that causal identification differs from statistical identification. Whereas the former makes sure that the causal estimands defined in section "Methodology" capture the intended treatment effect under assumption (i) through (iv), the latter refers to the extent to which the data can inform the prior within the ensemble defined by the full Bayesian model (Gelman et al., 2013; Schad et al., 2021). Throughout this section, we discuss statical identification.

the sampler is likely ill-designed and the resulting parameter estimates might be unreliable.

Among the strategies designed to validate an algorithmic approximation to the posterior of interest, Talts et al. (2018) discuss how to construct rank statistics based on $\boldsymbol{\Lambda}^*$ and $\tilde{\boldsymbol{\Lambda}}^{(s)}$. Since the sampler should not produce parameter values that are larger or smaller than the true posterior, nor the variance of the posterior samples should exceed the true posterior dispersion, one can use knowledge of $\boldsymbol{\Lambda}^*$ to assess $\tilde{\boldsymbol{\Lambda}}^{(s)}$ as sketched in algorithm 2.

Algorithm 2: Simulation-based calibration

For all $k \in \{1, ..., K\}$ do:

(1) Sample from $f(\boldsymbol{\Lambda})$ and $f(\boldsymbol{\Lambda}^c)$ to obtain $\boldsymbol{\Lambda}_k^*$ and $\boldsymbol{\Lambda}_k^{c,*}$

(2) Sample from $f\left(\boldsymbol{r}|\boldsymbol{\Lambda}_k^*, \boldsymbol{\Lambda}_k^{c,*}\right)$ and obtain $\boldsymbol{r}_k^*$

(3) Fit $f\left(\boldsymbol{\Lambda}|\boldsymbol{r}_k^*\right)$ and obtain the $S \times L$ matrix $\tilde{\boldsymbol{\Lambda}}_k$

For all $l \in \{1, ..., L\}$ do:

(3.1) For all $s \in \{1, ..., S\}$ calculate

$$i_{l,k}^{(s)} = \mathbb{I}\left\{\Lambda_{l,k}^* < \tilde{\Lambda}_{l,k}^{(s)}\right\}$$

(3.2) Calculate ranks as $R_{l,k} = \sum_{s=1}^{S} i_{l,k}^{(s)}$

Intuitively, had the computation been successful, the posterior distribution should overlap with the prior, yielding $\boldsymbol{\Lambda}_k^* \approx \tilde{\boldsymbol{\Lambda}}_k^{(s)}$ for all $s \in \{1, ..., S\}$. In this case, the sequence of rank statistics $\{R_{l,k}|k \in 1, ..., K\}$ for some parameter $\Lambda_l \in \boldsymbol{\Lambda}$ should be approximately uniformly distributed. Otherwise, the rank distribution will take on different shapes depending on the local differences between the prior and the posterior[11] (Cook et al., 2006; Talts et al., 2018).

Discrepancies between $\boldsymbol{\Lambda}^*$ and $\tilde{\boldsymbol{\Lambda}}^{(s)}$ can be also investigated using two useful summary statistics discussed in Schad et al. (2021). One the one hand, one can calculate z-scores as:

$$z_{l,k} = \frac{\hat{\mathbb{E}}\left[\tilde{\boldsymbol{\Lambda}}_{l,k}\right] - \Lambda_{l,k}^*}{\hat{\sigma}\left(\tilde{\boldsymbol{\Lambda}}_{l,k}\right)}$$

where $\tilde{\boldsymbol{\Lambda}}_{l,k}$ indicates the $S \times 1$ vector of posterior draws for parameter l in iteration k (i.e., the l^{th} column of the $\tilde{\boldsymbol{\Lambda}}_k$ matrix), while $\hat{\mathbb{E}}[\cdot]$ and $\hat{\sigma}(\cdot)$ represent the empirical posterior mean and standard deviation, respectively (see algorithm 3). These indicators capture how much the posterior mean overlaps with the true parameter, weighted by the posterior uncertainty. Therefore, z-scores quantify how accurately the computed posterior recovers the true model configuration through a combination of bias and precision. Specifically, smaller values of $z_{l,k}$ indicate that the posterior more strongly concentrates around the true value, while larger values suggest that the posterior concentrates away from ground truth. The last metric that we use in our simulation exercise is the posterior contraction:

$$c_{l,k} = 1 - \left[\frac{\hat{\sigma}\left(\tilde{\boldsymbol{\Lambda}}_{l,k}\right)}{\sigma\left(\Lambda_l^*\right)}\right]^2$$

where $\sigma\left(\Lambda_l^*\right)$ is the prior standard deviation of parameter l (see Equation 8). Since the additional information provided by the likelihood should reduce uncertainty, the posterior variance is expected to be smaller than the prior variance. Therefore, in case of highly informative data, the numerator decreases, thereby triggering posterior contraction and bringing $c_{l,k}$ close to one. In this case, we say that the parameter is statistically identified (Gelman et al., 2013; Schad et al., 2021). Vice versa, when the data fails to inform Λ_l, the ratio between the two variances will be close to one, pushing $c_{l,k}$ toward zero. Both z-scores and posterior contraction measures are calculated following algorithm 3.

Algorithm 3: z-scores and posterior contraction

For all $k \in \{1, ..., K\}$ do:

(1) Sample from $f(\boldsymbol{\Lambda})$ and $f(\boldsymbol{\Lambda}^c)$ to obtain $\boldsymbol{\Lambda}_k^*$ and $\boldsymbol{\Lambda}_k^{c,*}$

(2) Sample from $f\left(\boldsymbol{r}|\boldsymbol{\Lambda}_k^*, \boldsymbol{\Lambda}_k^{c,*}\right)$ and obtain $\boldsymbol{r}_k^*$

(3) Fit $f\left(\boldsymbol{\Lambda}|\boldsymbol{r}_k^*\right)$ and obtain the $S \times L$ matrix $\tilde{\boldsymbol{\Lambda}}_k$

For all $l \in \{1, ..., L\}$ do:

(3.1) Calculate $\hat{\mathbb{E}}\left[\tilde{\boldsymbol{\Lambda}}_{l,k}\right] = S^{-1} \sum_{s=1}^{S} \tilde{\Lambda}_{l,k}^{(s)}$

(3.2) Calculate $\hat{\sigma}\left(\tilde{\boldsymbol{\Lambda}}_{l,k}\right) = \sqrt{S^{-1} \sum_{s=1}^{S} \left(\tilde{\Lambda}_{l,k}^{(s)} - \hat{\mathbb{E}}\left[\tilde{\boldsymbol{\Lambda}}_{l,k}\right]\right)^2}$

(3.3) Calculate z-scores as $z_{l,k} = \frac{\hat{\mathbb{E}}\left[\tilde{\boldsymbol{\Lambda}}_{l,k}\right] - \Lambda_{l,k}^*}{\hat{\sigma}\left(\tilde{\boldsymbol{\Lambda}}_{l,k}\right)}$

(3.4) Calculate posterior concentration as

$$c_{l,k} = 1 - \left[\frac{\hat{\sigma}\left(\tilde{\boldsymbol{\Lambda}}_{l,k}\right)}{\sigma\left(\Lambda_l^*\right)}\right]^2$$

Ideally, the distribution of $z_{l,k}$ and $c_{l,k}$ across the K generated dataset should be inspected jointly, meaning that optimal performance corresponds to both z-scores

[11] Notice that calibrated sampling algorithms make sure that the credible intervals obtained from the resulting posterior distribution provide (approximate) nominal coverage (Schad et al., 2021; Talts et al., 2018). Therefore, across all simulations, any X% posterior credible intervals will include the '*true*' parameters in approximately X out of 100 replications. Since simulating implies that there can be many different X% credible intervals, the average coverage will be X% for all of them.

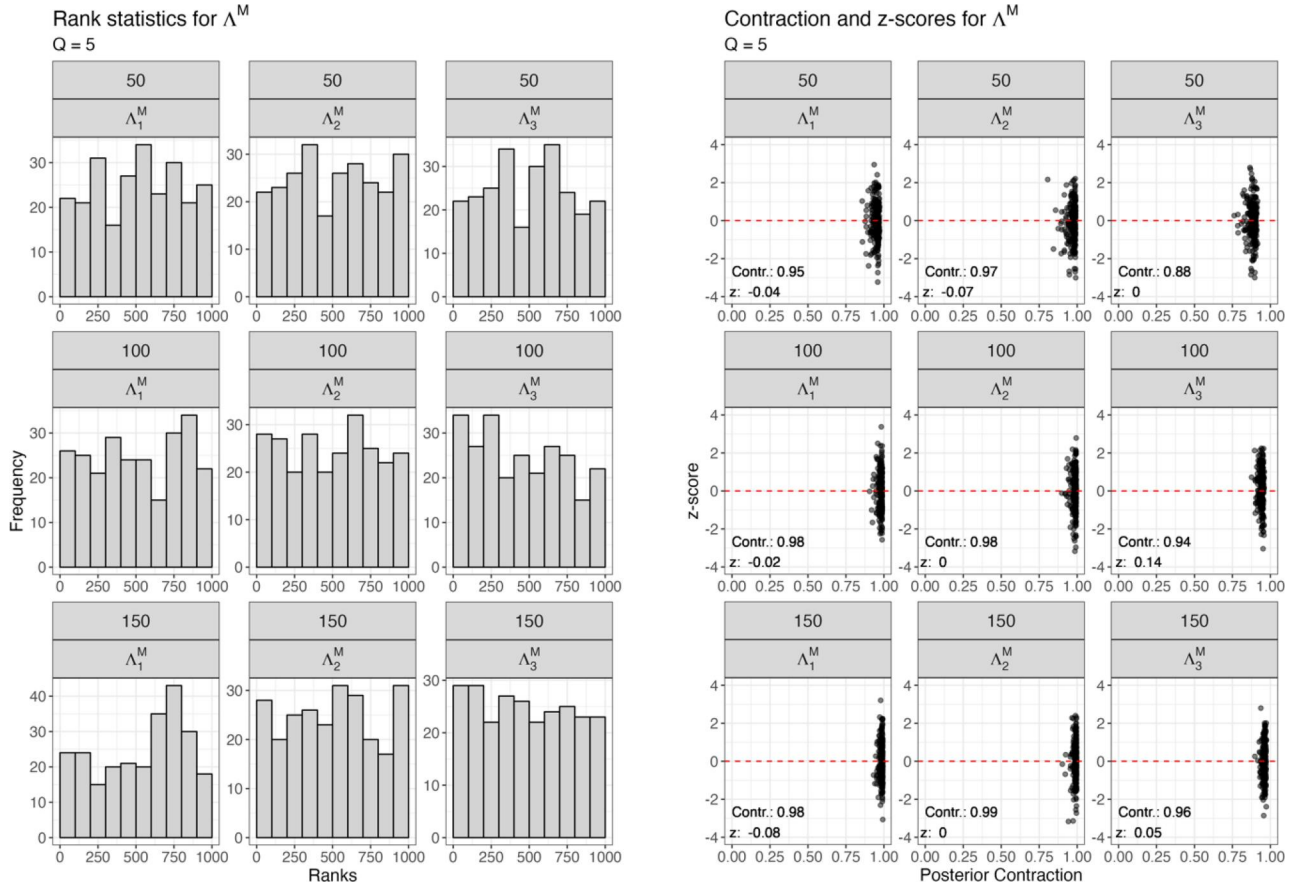

Figure 2. Simulation results for $\mathbf{\Lambda}^M$, $Q = 5$ and $\mathcal{I} \in \{50, 100, 150\}$. Each row corresponds to a different sample size, $\mathcal{I}$. The over-impressed values indicate the average posterior contraction (top) and the average z-score (bottom).

centered around zero and posterior contractions skewed toward one.

Our simulation exercise proceeds as follows: we begin by constructing a simple $\mathcal{I} \times 2$ covariate matrix using $X_{i,1} \sim \mathcal{N}(2, 1)$ and $X_{i,2} \sim \text{Gamma}(1, 2)$, for all $i \in \{1, ..., \mathcal{I}\}$, where $\mathcal{I}$ is the number of respondents in the simulated data. We next randomly assign each observation to either a treatment or a control group *via* $D_i \sim \text{Bernoulli}(0.5)$. To test our algorithm against differently sized samples, we set $\mathcal{I}$ to 50, 100 and 150. Finally, we generate responses for M and Y on a 1 to 5 Likert scale *via* $r^z_{i,q}|\theta^z_i, \alpha^z_q, \beta^z_q, \boldsymbol{\kappa}^z \sim \text{GRSM}(\theta^z_i, \alpha^z_q, \beta^z_q, \boldsymbol{\kappa}^z)$, with Q set to either 3, 5 or 8 (step 2 of Algorithms 2 and 3), while we randomly extract '*true*' parameters for $\theta^z_i, \alpha^z_q, \beta^z_q$ and $\boldsymbol{\kappa}^z$ from the priors defined in Equation (5), (6) and (7) (step 1 of Algorithms 2 and 3). We also simulate the regression coefficients in both the mediator and the outcome conditional mean function using the priors in Equation (8). Since step 3 in algorithms 2 and 3 requires fitting the model multiple times, we set $K = 250$ to keep computation time reasonably low.

Figure 2 displays the calculated ranks, contractions and z-scores[12] for $\mathbf{\Lambda}^M = [\Lambda^M_1, \mathbf{\Lambda}^M_{\mathbf{X}}]$, where Λ^M_1 corresponds to Λ^M_D in Equations (6) and (7), $\mathbf{\Lambda}^M_{\mathbf{X}} = [\Lambda^M_2, \Lambda^M_3]$ indicates the 2×1 coefficient vector[13] for the 2×1 simulated covariate vector $\mathbf{X}_i = [X_{i,1}, X_{i2}]$, $\mathcal{I} \in \{50, 100, 150\}$ and $Q = 5$. Inspecting the rank plots revels that the sampler is well calibrated in that the posterior of all $\Lambda^M_l \in \mathbf{\Lambda}^M$ recovers the prior distribution rather precisely (for $K = 250$). Similarly, all the z-scores are relatively close to zero, with most contained between the values -2 and of 2, indicating very good recovery of the '*true*' parameters $\mathbf{\Lambda}^{M,*}$ (vertical axis). Contraction ranges also exhibit encouraging results, with the most parameters' posterior draws successfully incorporating data information and yielding very strong contraction in general (i.e., close to 1).

[12]Note that here we are only addressing the posterior distribution of the slope parameters as they are essential to correctly estimate the casual estimands of interest. This kind of analyses, however, can be easily extended to all other model parameters to investigate whether our computation provides accurate posterior estimates.

[13]Since we standardize both X_1 and X_2, there is no need to simulate intercepts Λ^M_0 and Λ^Y_0.

Unsurprisingly, posterior contractions tend to improve as $\mathcal{I}$ grows from 50 to 150, although these differences are barely noticeable.

Similar results also hold for $\mathbf{\Lambda}^Y = [\Lambda_1^Y, \mathbf{\Lambda}_{\mathbf{X}}^Y, \Lambda_4^Y, \Lambda_5^Y]$, where Λ_1^Y, Λ_4^Y and Λ_5^Y correspond to, respectively, Λ_D^Y, Λ_M^Y and $\Lambda_{M,D}^Y$ in Equations (6) and (7), while $\mathbf{\Lambda}_{\mathbf{X}}^Y = \left[\Lambda_2^Y, \Lambda_3^Y\right]$ as defined in the paragraph above (Figure 3). However, since the equation for the latent outcome includes variables measured with error in the mean function, our simulation produces higher dispersion in the information gain statistics across the K replications (horizontal axis). In fact, the distribution of the posterior contractions appears more spread out for $\mathcal{I} = 50$ and collapses to one as $\mathcal{I}$ grows to 150. However, even in cases where $\mathcal{I}$ is low, the average contraction statistic floors at roughly 0.8, which indicates a very good performance. The z-scores for $\mathbf{\Lambda}^Y$ are also symmetrically distributed around zero, suggesting that the corresponding simulated parameters were, on average, successfully recovered.

Appendix A4[14] (figures A5 to A8, supplementary material) presents simulation results for both $Q = 3$ and $Q = 8$. Providing more questions does not seem to contribute much to the already good z-scores and posterior contractions of $\mathbf{\Lambda}^M$. On the other hand, setting $Q = 3$ yields higher dispersion in the contraction values of $\mathbf{\Lambda}^Y$ while, conversely, $Q = 8$ improves on $Q = 5$ through more concentrated contraction statistics. In either case, however, the average c_l and z_l floor at 0.75 and -0.08 across all $\mathbf{\Lambda}^Y$, indicating high information gain and precision across replications.

Unlike the priors in Equation (5), the distributions in Equation (8) are not explicitly referenced in the literature. Rather, they reflect a reasonable range of expected coefficients values, given standardized covariates and $\sigma^M = \sigma^Y = 1$. We thus repeat the analysis using more diffuse priors for both $\mathbf{\Lambda}^M$ and $\mathbf{\Lambda}^Y$. This configuration is more challenging to address because there are higher chances of observing large $\mathbf{\Lambda}^{M,*}$ which may contribute to generate noisier θ_i^M. Consequently, θ_i^Y will also exhibit higher dispersion because of the interaction between $\Lambda_4^{Y,*}, \Lambda_5^{Y,*}$ and θ_i^M. Figures A9 and A10 in Appendix 4 (supplementary material) suggest that placing less informative distributions on the slope parameters of the mediator's/outcome's conditional mean changes their rank, z-score, and contraction statistics only marginally. As a result, algorithm 1 should not be too sensitive to weaker prior knowledge about its core components.

Finally, although $\mathbf{\Lambda}^M$ and $\mathbf{\Lambda}^Y$ are key to construct the causal estimands of interest, they are not of direct relevance in mediation analysis. Rather, our interest lies in the ACME and the ANDE, which are both derived by combining the individual components of the two above parameter sets. Therefore, the final step of our simulation addresses the ranks, z-scores, and contraction values of the four causal quantities introduced in section "Identification of the treatment effects". Unfortunately, not all the evaluation metrics discussed in this section can be directly applied to the g-computation output discussed in section "Bayesian estimation". Since algorithm 1 generates counterfactuals by sampling from a normal distribution whose mean depends on the posterior estimates of the slope coefficients, calculating contraction values becomes challenging because the prior variance of the corresponding quantities is not available. Furthermore, the very definition of rank statistic conflicts with the goal of algorithm 1 to incorporate more uncertainty into the posterior distribution of the causal estimands. Consequently, only z-scores can be consistently calculated and examined. To do so, we modify algorithm 3 by both adding step (1.1), where we construct $\delta_G^*(\mathcal{d})$ and $\zeta_G^*(\mathcal{d})$ as in algorithm 1, and re-defining:

$$z_k^{\delta} = \frac{\hat{\mathbb{E}}\left[\tilde{\boldsymbol{\delta}}_{G,k}(\mathcal{d})\right] - \delta_{G,k}^*(\mathcal{d})}{\hat{\sigma}\left[\tilde{\boldsymbol{\delta}}_{G,k}(\mathcal{d})\right]}; z_k^{\zeta} = \frac{\hat{\mathbb{E}}\left[\tilde{\boldsymbol{\zeta}}_{G,k}(\mathcal{d})\right] - \zeta_{G,k}^*(\mathcal{d})}{\hat{\sigma}\left[\tilde{\boldsymbol{\zeta}}_{G,k}(\mathcal{d})\right]}$$

where both $\tilde{\boldsymbol{\delta}}_{G,k}(\mathcal{d})$ and $\tilde{\boldsymbol{\zeta}}_{G,k}(\mathcal{d})$ are $S \times 1$ vector of posterior draws for the two causal estimands, respectively. Figure 4 reports the z-scores of the four causal estimands, indicating that our posterior estimates correctly recover the '*true*' ACME and ANDE values. Results are also consistent for $Q = 3$ and $Q = 8$ (see Appendix A4, figures A11 and A12, supplementary material).

To assess contraction values and ranks, we test our estimation approach against $\delta^*(\mathcal{d})$ and $\zeta^*(\mathcal{d})$ values generated *via* coefficient multiplication which, as discussed at the end of section "Bayesian estimation", corresponds to a g-computation algorithm using μ^Y and μ^M only. For these quantities, we can obtain prior standard deviations $\sigma(\delta^*)$ and $\sigma(\zeta^*)$ using the properties of the normal distribution:

$$\sigma(\delta^*) = \sigma\left(\Lambda_D^{M,*}\right)\left[\sigma\left(\Lambda_M^{Y,*}\right) + \sigma\left(\Lambda_{M,D}^{Y,*}\right)\mathcal{d}\right]$$

$$\sigma(\zeta^*) = \sigma\left(\Lambda_D^{Y,*}\right) + \sigma\left(\Lambda_{M,D}^{Y,*}\right)\left[\sigma\left(\Lambda_D^{M,*}\right)\mathcal{d} + \sum_{p=1}^{P} \overline{x}_p \sigma\left(\Lambda_{p,\mathbf{X}}^{M,*}\right)\right]$$

[14]Appendix A4 can be found in the supplementary material available online on OSF.io at: https://osf.io/8bz4j/?view_only=9a067a723dbd41b4bbc3fd6b922ddacc

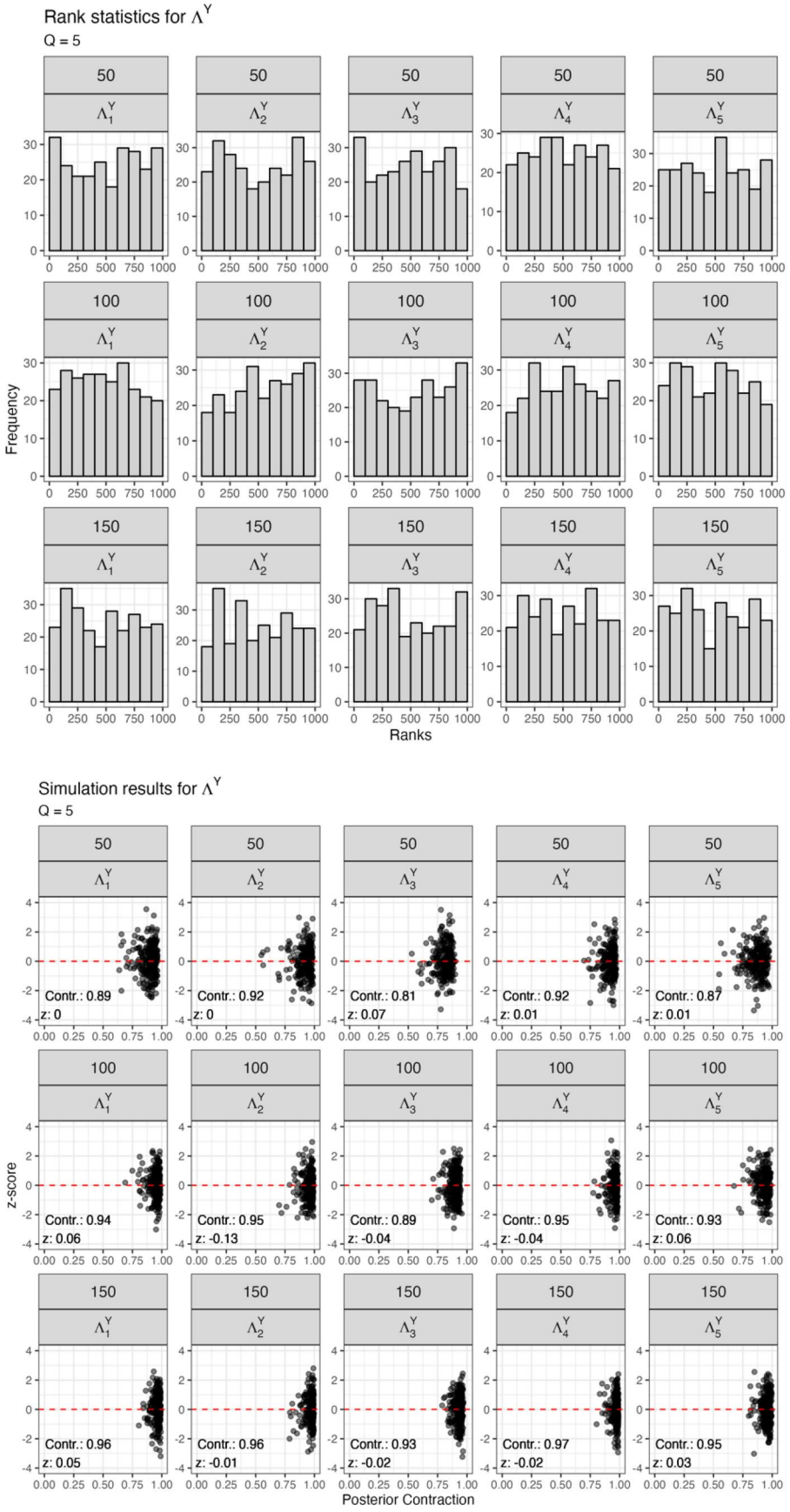

Figure 3. Simulation results for $\mathbf{\Lambda}^Y$, $Q = 5$ and $\mathcal{I} \in \{50, 100, 150\}$. Each row corresponds to a different sample size, $\mathcal{I}$. The over-impressed values indicate the average posterior contraction (top) and the average z-score (bottom).

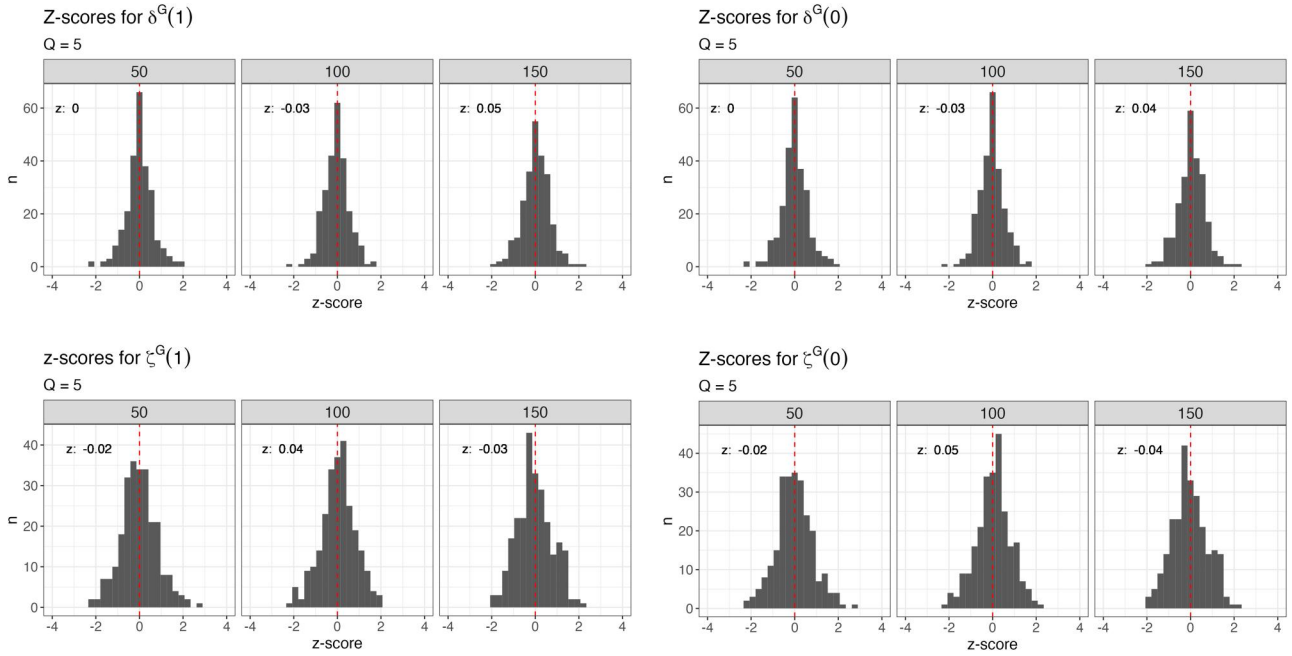

Figure 4. Simulation results for $\delta_G(d)$ and $\zeta_G(d)$, $Q = 5$ and $\mathcal{I} \in \{50, 100, 150\}$. Each box corresponds to a different sample size, $\mathcal{I}$. The overimpressed values indicate the average z-score.

where each individual component can be found in Equation (8). All the metrics displayed in Figure 5 confirm the evidence collected in the previous analyses: the simulated ACME and ANDE are correctly recovered, as indicated by the zero-centered z-scores. Average contractions are also satisfactory, although $\delta(0)$ exhibits a slightly larger dispersion of information gain across the replicated datasets. However, increasing the sample size quickly skews these values toward one. Ranks also appear uniformly distributed irrespective of the estimand or the corresponding sample size. These results also hold for $Q = 3$ and $Q = 8$ (see Appendix A4, figures A13 and A14, supplementary material).

Sensitivity analysis

One common criticism to the approach described in sections "Identification of the treatment effects" through "Bayesian estimation" is that the conditional independence assumption between the latent mediator and the latent outcome is often too strong of an identifying condition (Celli, 2022; Heckman & Pinto, 2015). Likewise, Equation (7) implicitly postulates zero residual correlation between θ_j^M and θ_j^Y, implying no association between the two quantities after controlling for the mediator and the treatment. One way that this assumption does not hold is the presence of unobserved confounding variables that affect both the mediator and the outcome (Imai et al., 2010a, 2010b). Since the credibility of the estimated ACME and ANDE hinges on mediator's conditional ignorability, we present a simple sensitivity test targeting the potential correlation between θ_j^M and θ_j^Y. Building on the model introduced in section "Bayesian estimation", we can re-write Equation (7) as:

$$\begin{pmatrix} \theta_j^M \\ \theta_j^Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} \mu_j^M \\ \mu_j^Y \end{bmatrix}, \begin{bmatrix} \sigma^M & \rho \\ \rho & \sigma^Y \end{bmatrix} \right) \tag{9}$$

where both μ_j^M and μ_j^Y are defined in Equation (6) and $\rho \in [-1, 1]$. Clearly, for $\rho = 0$, Equation (9) reduces to Equation (7), which holds under the five assumptions listed in section "Identification of the treatment effects". Therefore, to assess the extent to which failures of sequential ignorability could impact the causal estimands of interest, we replace the univariate prior distributions on both θ_j^M and θ_j^Y with a multivariate normal where $\sigma^M = \sigma^Y = 1$ and ρ is fixed to some sensitivity value. To probe the consistency of our results against different degrees of dependence between the two latent constructs, we re-estimate the ACME and the ANDE by setting ρ to -0.8, -0.5, 0.5, and 0.8 and compare the resulting posterior distributions to those obtained under independence (i.e., $\rho = 0$). If the estimates obtained with values of $\rho \neq 0$ deviate too much from the baseline, the assumption of no unobserved confounding becomes more central to the credibility of the analysis.

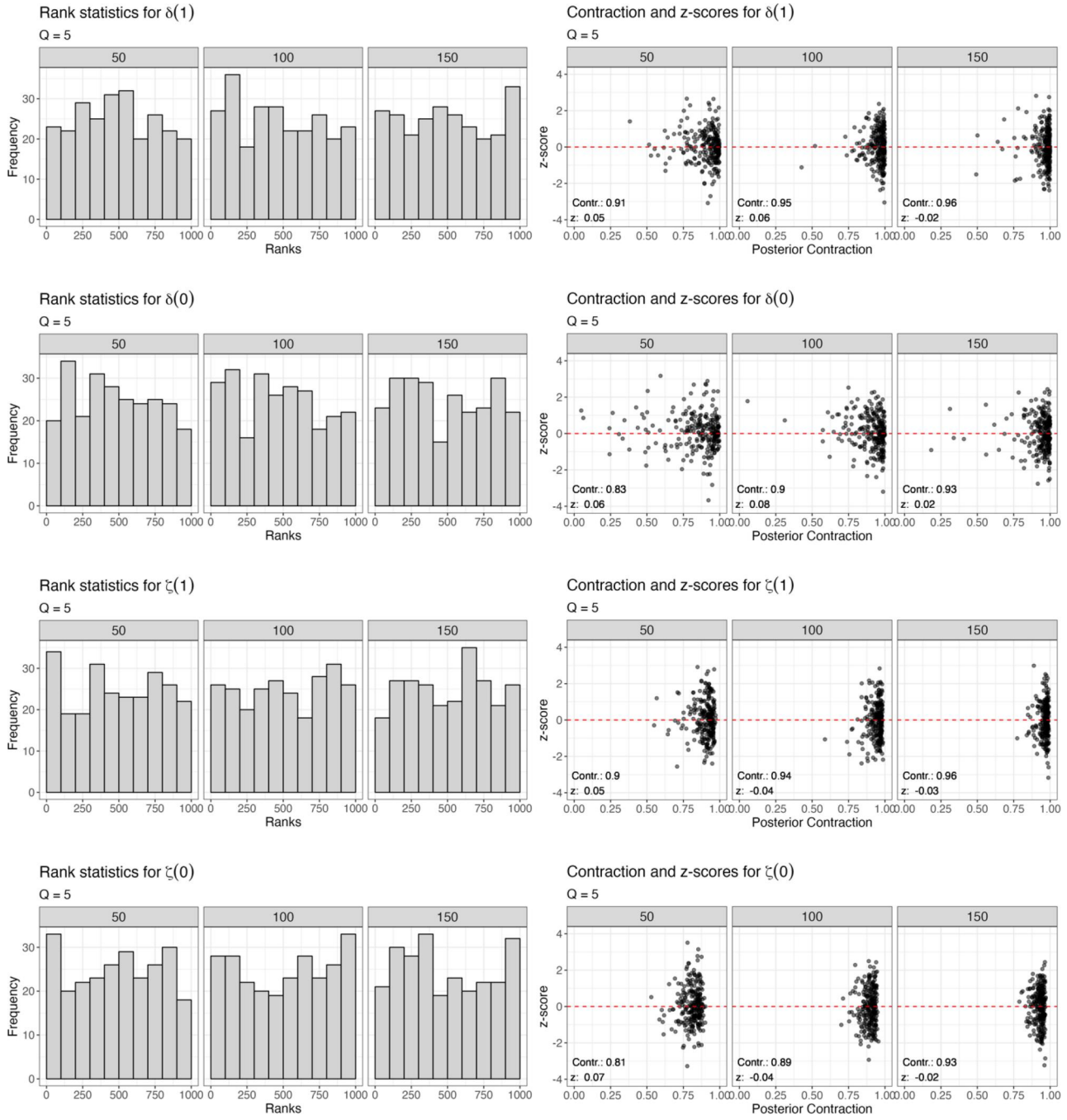

Figure 5. Simulation results for $\delta(d)$ and $\zeta(d)$, $Q = 5$ and $\mathcal{I} \in \{50, 100, 150\}$. Each row corresponds to a different sample size, $\mathcal{I}$. The overimpressed values indicate the average posterior contraction (top) and the average z-score (bottom).

Empirical application

The identification and estimation approaches presented in section "Methodology" can be easily applied to several experimental settings. As discussed in section "Methodology", Likert-scaled questions and CMA are in fact the bread and butter of many survey-based studies where the two main quantities of interest (i.e.: the outcome and the mediator) can only be indirectly elicited *via* structured questionnaires or other similar tools. In the following, we briefly present an empirical application from a randomized experiment designed to assess the willingness to buy for a food product under two different labeling schemes. Participants were either shown a label describing the product's nutritional characteristics (the so-called Nutriscore) or a label indicating that the food belongs to a so-called geographical indication (GI). The hypothesis is that the effect of the labels on the purchase intention is mediated by a latent construct that corresponds to the "healthiness" of the product as perceived by the

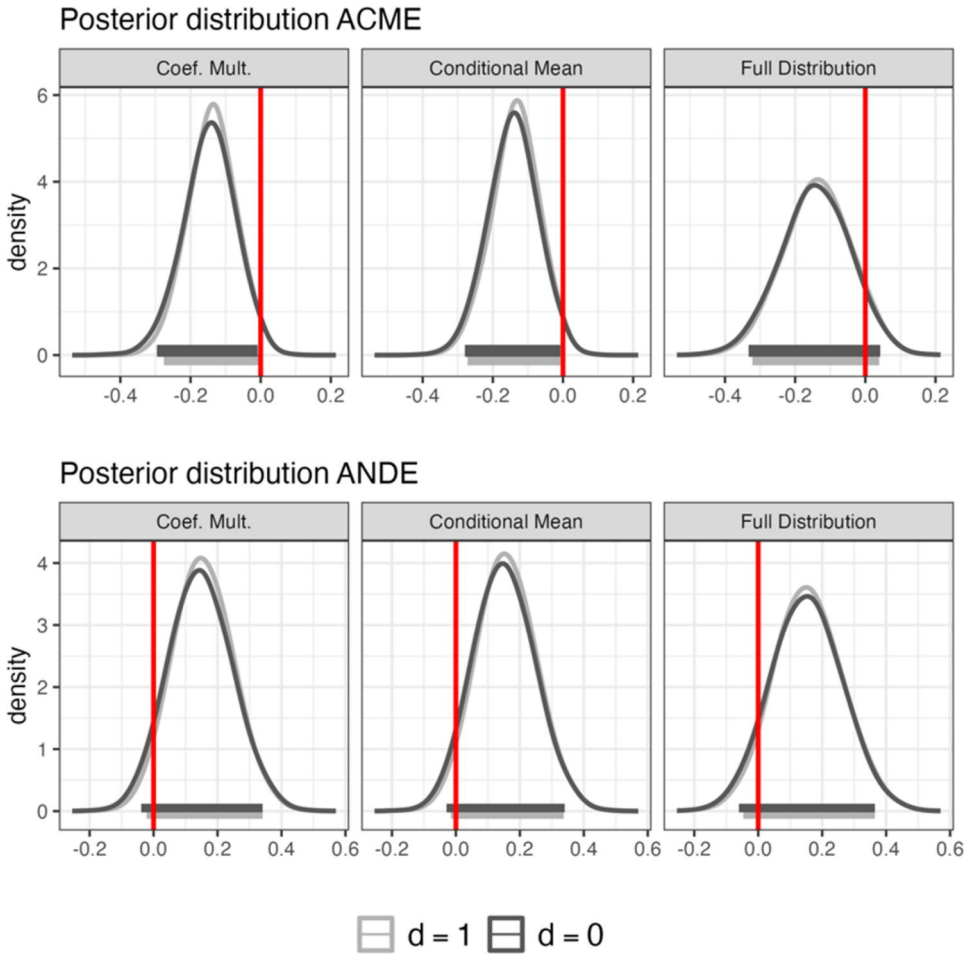

Figure 6. Posterior distributions for the ACME (top panels) and the ANDE (bottom panels) using the model implied by Equations (3) (5), (7) and (8) in Algorithm 1 (Full Distribution – rightmost panel), by solely resorting to conditional mean imputations (Conditional Mean – central panel) and through coefficients' multiplication (Coef. Mult. – leftmost panel). The solid lines at the bottom of the plots represent 95% credible intervals, while the solid vertical lines indicate zero (no) effect.

respondents (Ikonen et al., 2020). Both the purchase intention and the perceived healthiness were captured by two sets of Likert-valued questions (3 and 7 questions, respectively) which we use to elicit θ_j^Y and θ_j^M, respectively. Alongside these scores, control variables including respondents' nationality, degree, age, income, and family size were also collected. The experimental design consisted of a 2×2 factorial design which identifies four labeling conditions: no label (condition A), just one of the two labels (GI only—condition B; Nutriscore only—condition C) or both (condition D). Data were collected utilizing a between subject sampling scheme, where respondents were randomly assigned to one of the four experimental conditions. The survey was conducted online and administered throughout the Qualtrics platform.[15] To be eligible for the questionnaire, the respondents had to be above 18 years old, and declare to be at least partially involved in the shopping for their household. The final database included validated responses from 1,524 individuals living in Italy and The Netherlands.

To illustrate our methodological contribution, we focus on the Nutriscore label (condition C) and assess the ACME and ANDE when the this is compared against condition A (no label). In doing so, we contrast the results attained through algorithm 1 to the estimates obtained using the coefficients' product approach or the corresponding conditional g-computation method. Appendix A3 (supplementary material)[16] also discusses several convergence checks for the sampling algorithm that we used to approximate the posterior densities of the parameters in Equations (5) and (8). Figure 6 shows the distribution of the two

[15]All participants involved in the experimental survey have duly completed and signed the informed consent form. Additionally, it is important to note that the study adheres to all other requirements mandated by national legislations.

[16]Appendix A3 can be found in the supplementary material available online on OSF.io at: https://osf.io/8bz4j/?view_only=9a067a723dbd41b4bbc3fd6b922ddacc.

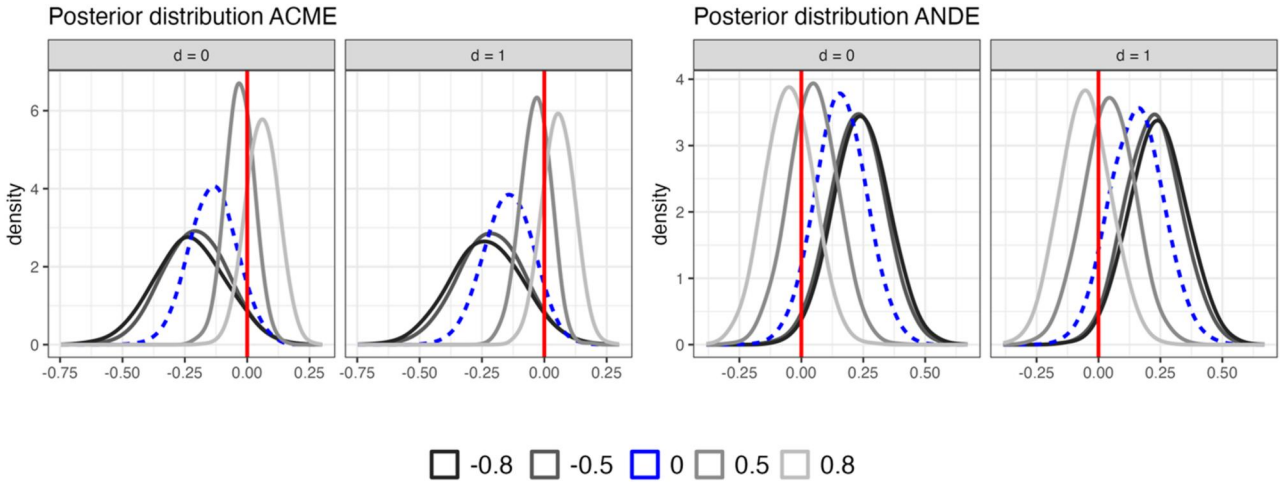

Figure 7. Sensitivity for the ACME (left panels) ANDE (right panels) obtained through Algorithm 1 for values of $\rho \in [-0.8, -0.5, 0.5, 0.8]$. The solid vertical lines indicate zero (no) effect.

causal quantities when these are estimated through algorithm 1 (Full Distribution), by the g-formula in Park and Kaplan (2015) (Conditional Mean) or *via* coefficient multiplication (as in linear SEMs).

Our results suggest strong evidence of a positive direct effect of the Nutriscore label on the respondents' purchase intentions (i.e.: the latent outcome - top panels), while the effect mediated by the perceived healthiness of the product (i.e.: the latent mediator—bottom panels) exhibits the opposite sign. Focusing on the latter, whereas the 95% credible intervals (solid horizontal bars) calculated using predictions from the conditional mean function (central panel) or coefficients' multiplication (leftmost panel) do not include zero (solid vertical lines), the one constructed using Algorithm 1 (rightmost panel) does. The reason for this difference is the larger dispersion in the posterior distributions of $\delta(\boldsymbol{d})$, when this is approximated by simulating POs from Equation (7) rather than Equation (6). This difference is also noticeable when switching to $\zeta(\boldsymbol{d})$, albeit less striking and neutral on the results.

At this point, it is worth mentioning that the latent constructs estimated through IRT models are typically best understood within the context of the analysis. In other words, our modeling approach assumes that the meaning researchers will attribute to the latent variables, as well as the structure of corresponding surveys, have been validated either before or while running the experiment (using, for example, a pilot sample). Conveniently, IRT models can be also used for measurement scales validation. Put differently, one can exploit the same modeling techniques presented in section "Measurement error model" to understand the extent to which a set of questions helps identifying the unobservable individual traits, conditional on the chosen model configuration. To this end, one could study the discrimination and difficulty parameters, the item characteristic curves and the amount of information provided by each item (question) to fine-tune the corresponding measurement system. However, since latent features are always defined on a dimensionless scale with approximately known range, one can only make sense of the relative size of $\delta(d)$ and $\zeta(d)$ by comparing them against the empirical distribution of $\hat{\theta}_j^Y = \hat{\mathbb{E}}\left[\tilde{\boldsymbol{\theta}}_j^Y\right]$, where $\tilde{\boldsymbol{\theta}}_j^Y = \left[\tilde{\theta}_j^{Y,(1)}, ..., \tilde{\theta}_j^{Y,(S)}\right]$, for all $j \in \{1, ..., N\}$. If one wished to give the latent mediator and/or the latent outcome (and, consequently, to the causal estimands) a practical interpretation, a sensible strategy could involve mapping these constructs to a measurable quantity that can be more easily appreciated within the scope of the experiment. To provide a concrete example, one could easily design a follow-up or contextual study to address if and to what extent a (unit) change along the θ^Y scale (i.e., purchase intention) translates to an observed (or stated) purchasing behavior (i.e., willingness to pay) using, for example, a choice experiment. Yet, given the contextual specificity of these exercises, we believe that a deeper discussion of these complementary aspects is best suited for a future applied work.

Figure 7 shows the results of the sensitivity analysis presented in section "Sensitivity analysis". These plots report the posterior distributions of both $\delta_G(\boldsymbol{d}, \rho)$ (left panels) and $\zeta_G(\boldsymbol{d}, \rho)$ (right panels) for $\rho \in \{-0.8, -0.5, 0, 0.5, 0.8\}$. What emerge is that, although our estimated ACME decreases for negative values of ρ, the corresponding uncertainty tends to get larger as we move away from null effects (solid

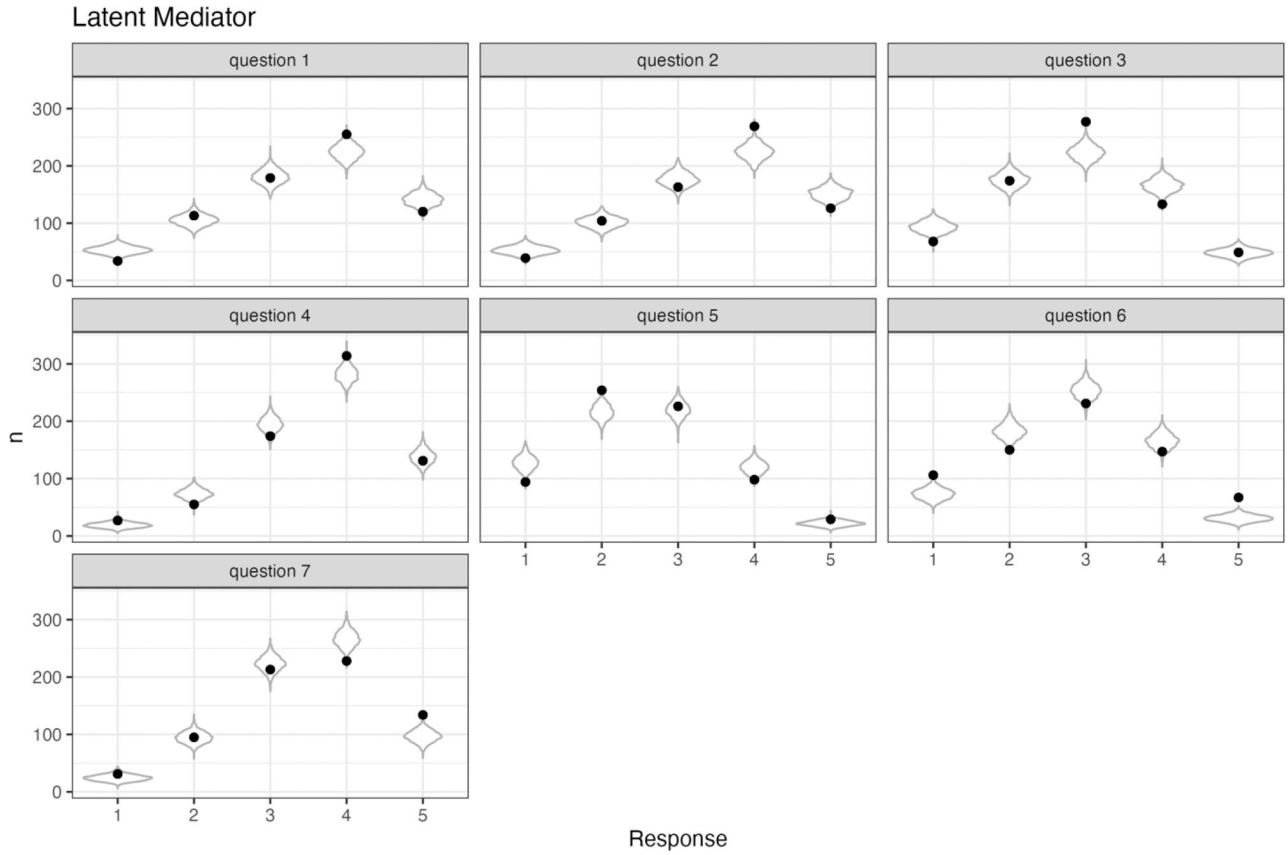

Figure 8. Posterior predictive checks for the latent mediator's model in Equation (3). The black dots represent the observed sum of responses equal to 1, 2, 3, 4 and 5 for each question. The grey density plots indicate the corresponding predicted sum obtained by sampling from model (12).

vertical line). On the other hand, when ρ takes on positive values, the ACME shifts toward zero and its posterior distribution contracts around the point estimate. Conversely, the ANDE remains positive and grows for negative correlation values, while it shifts back to zero when ρ moves into positive territory. It also worth noticing that both casual estimands move very little for large changes in negative correlation (i.e., the posterior distributions of the ACME and ANDE change very little between $\rho = -0.5$ and $\rho = -0.8$), while they tend to be more responsive to similar changes in positive ρ (i.e., when the correlation parameter moves from $\rho = 0.5$ to $\rho = 0.8$). Overall, these results indicate that our estimated effects can be quite sensitive to violations of the sequential ignorability assumption. Therefore, caution is recommended when interpreting these coefficients causally, unless it is reasonable to posit the absence of unobserved confounders.

As put forward in section "Measurement error model", we finally discuss two simple model checking techniques aimed at investigating how well the GRSM fits our survey data. All the plots depicted in Figures 8 through 10 hinge on the PPD of $r^z_{j,q}$, which corresponds to the marginalization (Gelman et al., 2013; Kruschke, 2014):

$$f\left(\tilde{r}^z_{j,q}|\boldsymbol{r}^z\right) = \int f\left(r^z_{j,q}|\boldsymbol{\Lambda} = \boldsymbol{\lambda}, \boldsymbol{\Lambda}^c = \boldsymbol{\lambda}^c\right) dF_{\boldsymbol{\Lambda},\boldsymbol{\Lambda}^c|\boldsymbol{r}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^c) \tag{10}$$

where $dF_{\boldsymbol{\Lambda},\boldsymbol{\Lambda}^c|\boldsymbol{r}}$ indicates the posterior distribution of $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}^c$. This quantity is approximated by sampling parameter values $\tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Lambda}}^c \sim dF_{\boldsymbol{\Lambda},\boldsymbol{\Lambda}^c|\boldsymbol{r}}$ and using the resulting draws to sample from $f\left(r^z_{j,q}|\boldsymbol{\Lambda} = \tilde{\boldsymbol{\lambda}}, \boldsymbol{\Lambda}^c = \tilde{\boldsymbol{\lambda}}^c\right)$. Figures 8 and 9 depict the sum of responses equal to 1, 2, 3, 4 and 5 for each question used to identify the corresponding unobservable quantity (i.e.: the latent mediator or the latent outcome). These representations are inspired by the score plots in Béguin and Glas (2001), where the authors evaluate a 3-paramters logit model by comparing the expected and predicted number of correct answers. The black dots in figures 8 and 9 indicate the observed sums, while the grey density plots represent the posterior predictions of the same quantities obtained by extracting S samples from the empirical approximation of Equation (10). Since the black dots

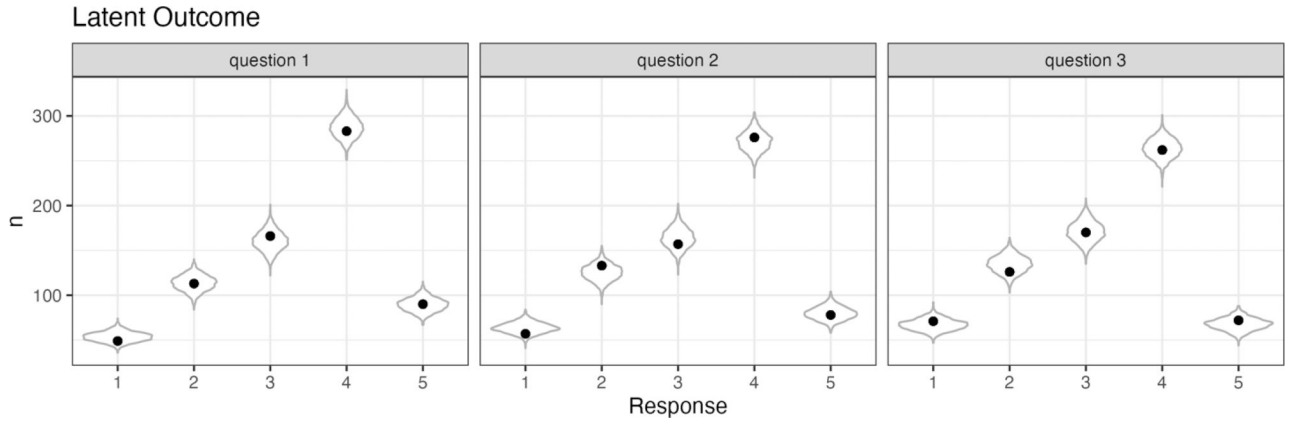

Figure 9. Posterior predictive checks for the latent outcome's model in Equation (3). The black dots represent the observed sum of responses equal to 1, 2, 3, 4 and 5 for each question. The grey density plots indicate the corresponding predicted sum obtained by sampling from model (12).

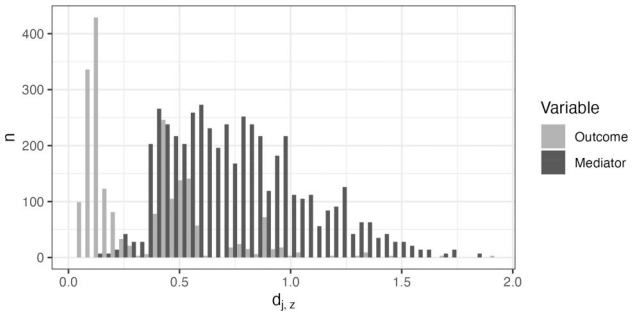

Figure 10. Posterior predictive checks for the latent mediator's model in Equation (3). The dark grey bars represent the respondent-specific distances for d_j^z and $z = M$, while the light grey bars represent the respondent-specific distances for d_j^z and $z = Y$.

fall either within or very close to the predictive distribution in most cases, this indicates that the GRSM does a reasonable job in fitting the data structure of our sample. However, results for the latent mediator suggest that there is room for improvement, especially for questions 5 and 6 where predictions appear farther apart from the observed scores.

Our second model assessment test is based on the frequency distribution of the respondent-specific average distance between observed and predicted responses. We calculate such distances as:

$$d_j^z = \sqrt{\sum_{q=1}^{Q^z} \frac{\left(r_{j,q}^z - \hat{\mathbb{E}}\left[\tilde{\boldsymbol{r}}_{j,q}^z\right]\right)^2}{N_q}}$$

where $\tilde{\boldsymbol{r}}_{j,q}^z$ is a S-vector of simulations $\tilde{\boldsymbol{r}}_{j,q}^z = \left[\tilde{r}_{j,q}^{z,(1)}, \ldots, \tilde{r}_{j,q}^{z,(S)}\right]$. The corresponding plots are reported in Figure 10, which essentially corroborates the insights provided by Figure 8 and Figure 9. Indeed, the average distances for the outcome model seem to suggest a satisfying fit, while the predicted responses pertaining to the mediator model indicate that the latter could be improved. However, most of the distances are in both cases less than one, propounding an overall reasonable performance of the GRSM.

Conclusions

Causal mediation analysis in an important approach to causal inference in that it not only allows to tackle the problem of quantifying the total effect of a treatment on a given outcome, but it also enables to break down this effect into a direct and indirect component. The latter postulates the existence of a mediating variable (i.e.: the so-called mediator) and often represents the causal estimand of interest known as average casual mediation effect. In the context of casual mediation analysis, several authors have discussed how measurement errors in the mediator may lead to biased estimates, recommending to control for error-in-variables when approaching this type of analysis. This issue is particularly relevant when working with survey data, where the mediator, the outcome, or both are typically measured through surrogate polytomous items such as Likert-scaled question. When this type of data represents the only way through which latent measures can be quantified, item response theory models provide a theoretically sound approach to map indirect discrete proxies on a continuous scale that characterizes the unmeasurable quantity of interest. In this paper, we have exploited the probabilistic nature of these statistical techniques to construct a simple Bayesian algorithm aimed at estimating both the direct and the indirect effect of a binary treatment on a latent outcome variable, through a latent mediator. Our identification strategy closely followed that of

previous seminal works in that it leveraged the potential outcome framework and the conditional ignorability of the mediator to work out tractable expressions for the causal estimands. In this respect, we discussed how our methodology compares to other similar proposals in casual mediation analysis and highlighted the main differences between our approach and the existing literature. In particular, we emphasized the importance of marginalizing over all the parameters of the counterfactual distributions to obtain conservative estimates for both the average causal mediation effect and the average natural direct effect.

This paper also contributes to the topic of sensitivity analysis with respect to the critical identifying assumptions of the underlying identification strategy. Specifically, we proposed a straightforward robustness check targeting the residual correlation between the latent outcome and the latent mediator. This sensitivity parameter represents the extent to which the presence of unobserved post-treatment cofounders can invalidate the assumption of independence between the outcome and the mediator, conditional on the treatment and observed exogenous covariates. We next showed how the proposed algorithm can be used in practice through an empirical application. Using data from a randomized experiment, we illustrated how the respondents' health consciousness can produce a negative mediation effect on the purchase intention for a specific food product, when the participants are exposed to a treatment label. Aimed at fostering the implementation of these techniques by applied researchers, we concluded by discussing the issue of computation time and show that estimation can be reliably sped up *via* variational inference methods. We also provide all the complete `R` and `Stan` scripts as well as a complementary `rmarkdown` document to guide potential users through all the methods and application discussed in the manuscript.

The present work can be extended in several ways. First, the Generalized Rating Scale Model can be too restrictive for some applications, particularly when the independence assumption between the latent mediator and outcome is deemed unrealistic because of peculiar experimental setting or the presence of unintended confounding mechanisms. In these situations, improving the current measurement error model through its multivariate counterpart might improve model fit and provide a more realistic representation of the underlying latent quantities. Second, the current casual model is limited to the simple case of one mediator, one outcome. In many applications, however, this framework can be too limiting, as causal mechanisms often involve several mediators (either latent or observable) as well as many treatments, not necessarily binary. Although such setups are still an active area of research, we believe that our proposed approach can be extended quite naturally to more complex structural models. Third, given the recent advances in Bayesian non-parametrics, a natural direction that the proposed approach could steer toward is modeling the conditional mean of the unobservable characteristics by flexible regression models such as Gaussian Processes, Bayesian Splines, Bayesian Additive Regression Trees (BART—Chipman et al., 2010) or similar techniques. This would allow going beyond simple average treatment effects and provide a solid framework for the estimation of heterogenous treatment effects.

Article information

Conflict of interest disclosures: The author signed a form for disclosure of potential conflicts of interest. The author did not report any financial or other conflicts of interest in relation to the work described.

Ethical principles: The author affirms having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work has was not supported.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Abbas, J. (2020). Impact of total quality management on corporate green performance through the mediating role of corporate social responsibility. *Journal of Cleaner Production*, *242*, 118458. https://doi.org/10.1016/j.jclepro.2019.118458

Albert, J. M., Geng, C., & Nelson, S. (2016). Causal mediation analysis with a latent mediator. *Biometrical Journal. Biometrische Zeitschrift*, *58*(3), 535–548. https://doi.org/10.1002/bimj.201400124

Andrich, D (2005). The Rasch model explained. In R. Maclean (Ed.), *Applied Rasch Measurement: A book of exemplars*. Education in the Asia-Pacific Region: Issues, Concerns and Prospects, vol 4 Springer. https://doi.org/10.1007/1-4020-3076-2_3

Andrich, D. (2016). Rasch rating-scale model. In W. J. van der Linder (Ed.) *Handbook of item response theory, Volume 1: Models* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315374512

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*(2), 171–187. https://doi.org/10.1093/pan/mpi010

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561. https://doi.org/10.1007/BF02296195

Benight, C. C., & Bandura, A. (2004). Social cognitive theory of posttraumatic recovery: The role of perceived self-efficacy. *Behaviour Research and Therapy*, *42*(10), 1129–1148. https://doi.org/10.1016/j.brat.2003.08.008

Betancourt, M. (2020). *Towards a principled Bayesian workflow* (RStan). https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Blut, M., & Wang, C. (2020). Technology readiness: A meta-analysis of conceptualizations of the construct and its impact on technology usage. *Journal of the Academy of Marketing Science*, *48*(4), 649–669. https://doi.org/10.1007/s11747-019-00680-8

Bürkner, P. C. (2019). Bayesian item response modeling in R with brms and Stan. ArXiv. https://doi.org/10.48550/arXiv.1905.09501

Celli, V. (2022). Causal mediation analysis in economics: Objectives, assumptions, models. *Journal of Economic Surveys*, *36*(1), 214–234. https://doi.org/10.1111/joes.12452

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-AOAS285

Cook, S. R., Gelman, A., & Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, *15*(3), 675–692. https://doi.org/10.1198/106186006X136976

Ding, P., & Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, *33*(2), 214–237. https://doi.org/10.1214/18-STS645

Fox, J. P., & Glas, C. A. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*(2), 169–191. https://doi.org/10.1007/BF02294796

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous response data. *The British Journal of Mathematical and Statistical Psychology*, *58*(Pt 1), 145–172. https://doi.org/10.1348/000711005X38951

Fujimoto, K. A., & Neugebauer, S. R. (2020). A general Bayesian multidimensional item response theory model for small and large samples. *Educational and Psychological Measurement*, *80*(4), 665–694. https://doi.org/10.1177/0013164419891205

Furr, D. C. (2017). Edstan: Stan Models for Item Response Theory. R package version 1.0.6. https://CRAN.R-project.org/package=edstan

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gelman, A., & Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions. National Bureau of Economic Research, Working Paper, 19614. https://doi.org/10.3386/w19614

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383. https://doi.org/10.1214/17-BA1051

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*(4), 544–565. https://doi.org/10.1207/S15328007SEM1004_4

Hanssens, D. M., Pauwels, K. H., Srinivasan, S., Vanhuele, M., & Yildirim, G. (2014). Consumer attitude metrics for guiding marketing mix decisions. *Marketing Science*, *33*(4), 534–550. https://doi.org/10.1287/mksc.2013.0841

Heckman, J. J., & Pinto, R. (2015). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric Reviews*, *34*(1-2), 6–31. https://doi.org/10.1080/07474938.2014.944466

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). SAGE.

Huber, M., Lechner, M., & Mellace, G. (2017). Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. *Review of Economics and Statistics*, *99*(1), 180–183. https://doi.org/10.1162/REST_a_00632

Ikonen, I., Sotgiu, F., Aydinli, A., & Verlegh, P. W. (2020). Consumer effects of front-of-package nutrition labeling:

An interdisciplinary meta-analysis. *Journal of the Academy of Marketing Science, 48*, 360–383. https://doi.org/10.1007/s11747-019-00663-9

Imai, K., Keele, L., & Yamamoto, T. (2010a). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51. https://doi.org/10.1214/10-STS321

Imai, K., Keele, L., & Tingley, D. (2010b). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765–789. https://doi.org/10.1017/S0003055411000414

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136–153. https://doi.org/10.1080/10705510701758406

Keele, L., Tingley, D., & Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management*, *34*(4), 937–963. https://doi.org/10.1002/pam.21853

Keil, A. P., Daza, E. J., Engel, S. M., Buckley, J. P., & Edwards, J. K. (2018). A Bayesian approach to the g-formula. *Statistical Methods in Medical Research*, *27*(10), 3183–3204. https://doi.org/10.1177/0962280217694665

Kim, C., Daniels, M., Li, Y., Milbury, K., & Cohen, L. (2018). A Bayesian semiparametric latent variable approach to causal mediation. *Statistics in Medicine*, *37*(7), 1149–1161. https://doi.org/10.1002/sim.7572

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Le Cessie, S., Debeij, J., Rosendaal, F. R., Cannegieter, S. C., & Vandenbroucke, J. P. (2012). Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, *23*(4), 551–560. https://doi.org/10.1097/EDE.0b013e318254f5de

Lawson, A. B., Kim, J., Johnson, C., Hastert, T., Bandera, E. V., Alberg, A. J., Terry, P., Akonde, M., Mandle, H., Cote, M. L., Bondy, M., Marks, J., Peres, L., Ratnapradipa, K. L., Xin, Y., Schildkraut, J., & Peters, E. S. (2023). Deprivation and segregation in ovarian cancer survival among African American women: A mediation analysis. *Annals of Epidemiology*, *86*, 57–64. https://doi.org/10.1016/j.annepidem.2023.07.001

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912–928. https://doi.org/10.1111/oik.05985

Li, F., Ding, P., & Mealli, F. (2023). Bayesian causal inference: A critical review. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *381*(2247), 20220153. https://doi.org/10.1098/rsta.2022.0153

Loh, W. W., Moerkerke, B., Loeys, T., Poppe, L., Crombez, G., & Vansteelandt, S. (2020). Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomized studies. *Multivariate Behavioral Research*, *55*(5), 763–785. https://doi.org/10.1080/00273171.2019.1681251

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203056615

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, *78*(3), 384–408. https://doi.org/10.1177/0013164417693666

Ma, Y., Wang, H., & Kong, R. (2020). The effect of policy instruments on rural households' solid waste separation behavior and the mediation of perceived value using SEM. *Environmental Science and Pollution Research International*, *27*(16), 19398–19409. https://doi.org/10.1007/s11356-020-08410-2

McCandless, L. C., & Somers, J. M. (2019). Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis. *Statistical Methods in Medical Research*, *28*(2), 515–531. https://doi.org/10.1177/0962280217729844

Mesquita, L. F., Anand, J., & Brush, T. H. (2008). Comparing the resource-based and relational views: Knowledge transfer and spillover in vertical alliances. *Strategic Management Journal*, *29*(9), 913–941. https://doi.org/10.1002/smj.699

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Muthén, B., & Asparouhov, T. (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling*, *22*(1), 12–23. https://doi.org/10.1080/10705511.2014.935843

Park, S., & Kaplan, D. (2015). Bayesian causal mediation analysis for group randomized designs with homogeneous and heterogeneous effects: Simulation and case study. *Multivariate Behavioral Research*, *50*(3), 316–333. https://doi.org/10.1080/00273171.2014.1003770

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, *66*(1), 825–852. https://doi.org/10.1146/annurev-psych-010814-015258

Rosseel, Y., & Loh, W. W. (2024). A structural after measurement approach to structural equation modeling. *Psychological Methods*, *29*(3), 561–588. https://doi.org/10.1037/met0000503

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. https://doi.org/10.1037/h0037350

Sadikoglu, E., & Zehir, C. (2010). Investigating the effects of innovation and employee performance on the relationship between total quality management practices and firm performance: An empirical study of Turkish firms. *International Journal of Production Economics*, *127*(1), 13–26. https://doi.org/10.1016/j.ijpe.2010.02.013

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Soregaroli, C., Varacca, A., Ricci, E. C., Platoni, S., Tillie, P., & Stranieri, S. (2022). Voluntary standards as meso-institutions: A Bayesian investigation of their relationships

with transaction governance and risks. *Applied Economic Perspectives and Policy*, *44*(4), 1660–1681. https://doi.org/10.1002/aepp.13252

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, *173*(7), 731–738. https://doi.org/10.1093/aje/kwq472

Stranieri, S., Varacca, A., Casati, M., Capri, E., & Soregaroli, C. (2021). Adopting environmentally-friendly certifications: Transaction cost and capabilities perspectives within the Italian wine supply chain. *Supply Chain Management*, *27*(7), 33–48. https://doi.org/10.1108/SCM-12-2020-0598

Sultan, P., Tarafder, T., Pearson, D., & Henryks, J. (2020). Intention-behaviour gap and perceived behavioural control-behaviour gap in theory of planned behaviour: Moderating roles of communication, satisfaction and trust in organic food consumption. *Food Quality and Preference*, *81*, 103838. https://doi.org/10.1016/j.foodqual.2019.103838

Sun, R., Zhou, X., & Song, X. (2021). Bayesian causal mediation analysis with latent mediators and survival outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 778–790. https://doi.org/10.1080/10705511.2020.1863154

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. arXiv. https://doi.org/10.48550/arXiv.1804.06788

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597

Toplu-Demirtaş, E., Akcabozan-Kayabol, N. B., Araci-Iyiaydin, A., & Fincham, F. D. (2022). Unraveling the roles of distrust, suspicion of infidelity, and jealousy in cyber dating abuse perpetration: An attachment theory perspective. *Journal of Interpersonal Violence*, *37*(3-4), NP1432–NP1462. https://doi.org/10.1177/0886260520927505

Van der Linden, W. J. (Ed.). (2018). *Handbook of item response theory: Three volume set. Volume 1: Models*. CRC Press. https://doi.org/10.1201/9781315119144

Vander Weele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, *2*(4), 457–468. https://doi.org/10.4310/SII.2009.v2.n4.a7

Vander Weele, T. J., Valeri, L., & Ogburn, E. L. (2012). The role of measurement error and misclassification in mediation analysis. *Epidemiology*, *23*(4), 561–564. https://doi.org/10.1097/EDE.0b013e318258f5e4

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Yamashita, T. (2022). Analyzing Likert scale surveys with Rasch models. *Research Methods in Applied Linguistics*, *1*(3), 100022. https://doi.org/10.1016/j.rmal.2022.100022

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301–322. https://doi.org/10.1037/a0016972