Research Article

Causal Inference Using Mediation Analysis or Instrumental Variables – Full Mediation in the Absence of Conditional Independence

By Thomas Otter, Max J. Pachali, Stefan Mayer and Jan R. Landwehr

Both instrumental variable (IV) estimation and mediation analysis are tools for causal inference. However, IV estimation has mostly developed in economics for causal inference from observational data. In contrast, mediation analysis has mostly developed in psychology, as a tool to empirically establish the process by which an experimental manipulation brings about its effect on the dependent variable of interest. As a consequence, many researchers well versed in IV estimation are not familiar with mediation analysis, and vice versa. In this paper, we discuss the commonalities and differences between IV estimation and mediation analysis. We highlight that IV estimation leverages an a priori assumption of full mediation for causal inference. In contrast, modern practice in mediation analysis focuses on testing the statistical significance of the indirect effect without too much regard for the specification of the estimated model. A drawback of this approach is that inferring mediation from the statistical significance of a (putative) indirect effect through the hypothesized mediator may be spurious altogether. We discuss specification issues and how they relate to inference about mediation, and specifically to the distinction between full and partial mediation. Based on this discussion we argue in favor of further developing tests that are more diagnostic about the underlying causal structure, motivated by the implication that full mediation could be more common than currently believed.

1. Introduction

In this paper, we compare and contrast the assumptions and goals of IV estimation and mediation analysis. Specifically, we investigate the relationship between full mediation at a causal theory level and what can be empirically observed in the data. This way we hope to sensitize applied researchers to the ambiguity of what is commonly referred to as "partial mediation". We argue that this is important given the current trends in methodological and empirical mediation research. In essence, the recent literature emphasizes statistical inference for indirect effects based on bootstrapping procedures (e. g., Hayes 2013; Preacher and Hayes 2004; Zhao et al. 2010) but essentially assumes that mediation is the underlying causal

Thomas Otter is Professor of Marketing at Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60623 Frankfurt am Main, Germany, Phone: +49 798 34 646, E-mail: otter@marketing.uni-frankfurt.de.
* Corresponding author

Max J. Pachali is Ph.D. Student at the Graduate School of Economics, Finance, and Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, 60623 Frankfurt am Main, Germany, Phone: +49 798 34 648, E-mail: max.pachali@hof.uni-frankfurt.de.

Stefan Mayer is Postdoctoral Researcher in the Marketing Department at Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60623 Frankfurt am Main, Germany, Phone: +49 798 34 692, E-mail: smayer@wiwi.uni-frankfurt.de.

Jan R. Landwehr is Professor of Marketing at Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60623 Frankfurt am Main, Germany, Phone: +49 798 34 631, E-mail: landwehr@wiwi.uni-frankfurt.de.

model. We do not argue with the value of improving statistical inference for coefficients and functions of coefficients believed to be different from zero in a given, undisputed model. However, we believe that some of the original appeal of mediation analysis is related to testing the underlying causal model as in Baron and Kenny (1986).

The original test for mediation proposed by Baron and Kenny (1986) builds on the connection between full mediation and conditional independence and tests conditional mean independence. Their test rests on the following set of regression equations, where t's denote intercepts (see also *Fig. 1*).

$$Y_i = t_1 + cX_i + \varepsilon_{Y,i} \quad (1)$$

$$M_i = t_2 + aX_i + \varepsilon_{M,i} \quad (2)$$

$$Y_i = t_3 + c^*X_i + bM_i + \varepsilon^*_{Y,i} \quad (3)$$

The first equation regresses Y on the randomly assigned experimental variable X. A statistically significant coefficient c establishes empirical support for the total effect from X to Y (cf. *Fig. 1*, Panel A).[1] Because of random assignment of X, the coefficient c necessarily measures a causal effect. The second equation regresses M on X. A statistically significant coefficient a establishes empirical support for the effect from X to M that is again causal by experimental design. The third equation regresses Y on randomly assigned X and on observed M. Under the assumptions of full mediation, the absence of measurement error in M, and independence between $\varepsilon_{M,i}$ and $\varepsilon^*_{Y,i}$, the hypothesis that $c^* = 0$ holds and b measures the causal effect from M to Y (cf. *Fig. 1*, Panel B). Usually, empirical support for the hypothesis of $c^* = 0$ is established based on p-values larger than some subjectively chosen level α (Baron and Kenny 1986).[2]

Baron and Kenny (1986) argued that empirical support for $c^* = 0$ is the "strongest demonstration of mediation" (p. 1176) and that if the residual path $c^* \neq 0$, "this indicates the operation of multiple mediating factors" (see also Demming et al. 2017). While we agree with the conclusion that empirical evidence supporting $c^* = 0$ is essentially proof of mediation[3], we disagree with the statement that a residual path $c^* \neq 0$ necessarily indicates additional mediating factors. In fact, the main contribution of our paper is to highlight plausible data generating mechanisms, including the empirically relevant case of measurement error in the mediator, that give rise to $c^* \neq 0$ even though full mediation holds at a causal theory level. Under these data generating mechanisms, standard mediation analysis relying on Baron and Kenny (1986), or the more modern variants of testing the statistical significance of the indirect effect (Hayes 2013; Pieters 2017; Preacher and Hayes 2004, 2008; Zhao et al. 2010) yield biased inferences about the indirect effect. Thus, conclusions about the presence and strength of mediation may be misleading.

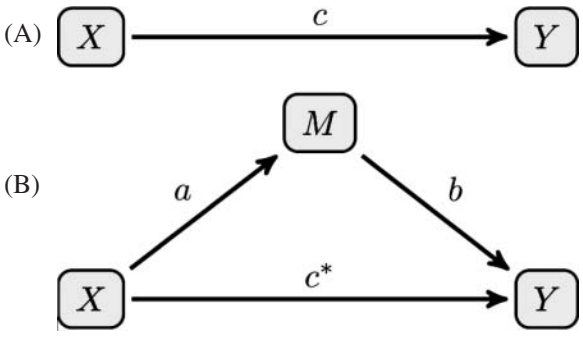


Fig. 1: Mediation according to Baron and Kenny (1986)

We view this as a first order concern and believe that the current focus on *statistically* reliable estimates of indirect effects (e. g., Hayes 2013; Rucker et al. 2011; Zhao et al. 2010) should be widened to include an assessment of the underlying causal structure to the fullest extent possible. In this context, it cannot be emphasized enough that a statistically significant indirect effect is merely consistent with mediation but may be obtained from model structures without any mediation (see also Fiedler et al. 2011), such that what is estimated as a significant indirect effect may in fact not be an indirect effect at all, but reflecting other reasons for dependence between observed variables. Hence, we argue for a careful assessment of alternative models and for the development of model comparison procedures that account for violations of conditional independence between X and Y given observed M that occur even if full mediation holds at a causal theory level.

The remainder of the paper proceeds as follows. First, we explain the relationship between full mediation – which goes hand in hand with necessarily unbiased inference for the indirect effect – and conditional independence, and describe the conceptual link between full mediation and IV estimation. Furthermore, we explain and show why a correlation between the error terms of M and Y will force any statistical procedure testing for conditional independence to indicate only partial mediation although full mediation is the data generating mechanism, while simultaneously biasing inferences about the indirect effect. Second, we introduce the data generating mechanism of true partial mediation and describe why true partial mediation is not testable by statistical procedures. Third, we exemplify how statistical procedures may spuriously point to partial mediation although the data were not generated by a mediating process at all. Fourth, we show that statistical procedures spuriously point to partial mediation although full mediation was the data generating mechanism when the mediator is measured with measurement error, while simultaneously biasing inferences about the indirect effect again. Fifth, we introduce a Bayesian test procedure for testing conditional independence (i. e., full mediation) that can positively support this hypothesis. In contrast, standard testing based on p-values can only fail to reject conditional independence. Finally, we close with a discussion of our observations, recommendations for applications, and an outlook towards future methodological developments.

2. Mediation and conditional independence

2.1. Full mediation and IV estimation

The just described classical mediation test by Baron and Kenny (1986) is essentially a test of conditional (mean) independence. That is, one tests whether *X* and *Y* become *independent* of one another, once *M* is controlled for in the third equation, i. e., if the effect of *X* on *Y* is zero *conditional* on a fixed value of *M*. Intuitively, this is consistent with all the effect from *X* onto *Y* going "through" *M*. However, simple conditioning on observed values of *M* in the data will only result in conditional independence between *X* and *Y* under the additional assumptions of error free measurements of the mediator *M* and independence between unobserved causes of *M* and *Y*.

The directed acyclic graph (DAG) that corresponds to full mediation under these additional assumptions is depicted in *Fig. 2*. In a DAG directed arrows from one variable to another indicate direct causal effects (see also Rohrer 2018). The absence of a direct connection indicates that two variables are only indirectly connected, if at all. A causal effect in general means that manipulating the cause (the variable at the origin of an arrow) will consistently affect the consequence (the variable the arrow points to).

The arrows connecting *X* to *M* and *M* to *Y* indicate directed causal effects. The arrows pointing to *X*, *M*, and *Y* coming from *U*s indicate the influence of independent, unobserved disturbances. The graph shows the data generating mechanism (i. e., the theoretical model) of full mediation. Accordingly, we use β_1 and β_2 to denote the theoretical causal pathways (cf. *a* and *b*, which we used earlier to denote the empirical estimates of β_1 and β_2 under the assumptions embedded in *Fig. 2*).

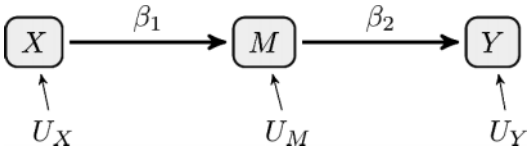


Fig. 2: Full mediation (directed acyclic graph)

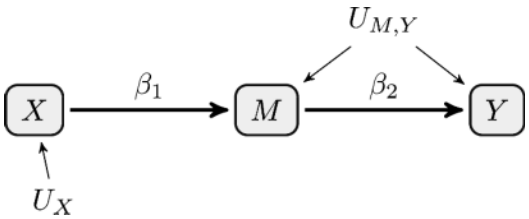


Fig. 3: Instrumental variable (directed acyclic graph)

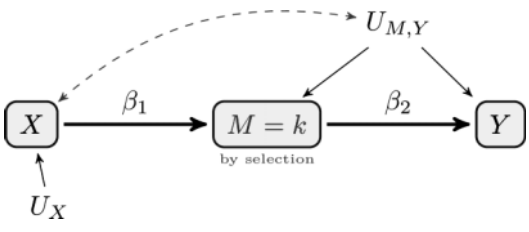


Fig. 4: Conditioning on the mediator M (directed acyclic graph)

The DAG shows that an intervention that forces *M* to take a particular fixed value blocks the link from *X* to *Y*, such that manipulations of *X* no longer translate into *Y*-changes. *In addition*, the DAG implies conditional independence between *X* and *Y* given *M*, i. e., $p(Y|X,M) = p(Y|M)$. Conditional independence in the data generating mechanism translates into a probability α to reject the true Null hypothesis that the direct effect of *X* on *Y* conditional on *M* equals zero. Moreover, it leads to a Bayes factor supporting the model that excludes the direct effect from *X* to *Y* over the model that includes this effect.

In empirical applications, independence between U_X and $\{U_M,U_Y\}$ is guaranteed if *X* is subject to experimental manipulation with random assignment to treatment conditions. Independence between U_M and U_Y is an assumption (Imai et al. 2010a), just like the absence of a direct effect from *X* on *Y* conditional on *M*, when assuming full mediation. In this context, it is useful to compare the DAG in *Fig. 2* to the DAG that motivates IV estimation depicted in *Fig. 3*.[4] The only difference to *Fig. 2* is that the unobserved background variables influencing *M* and *Y* are no longer independently distributed. In other words, there exist unobservables $U_{M,Y}$ that jointly affect *M* and *Y*.

The DAG in *Fig. 3* again shows that an intervention that forces *M* to take a particular fixed value blocks the link from *X* to *Y*, such that manipulations of *X* no longer translate into *Y*-changes, i. e., full mediation. However, the DAG in *Fig. 3* no longer implies conditional independence between *X* and *Y* given *M*, even if we can be sure of the independence between U_X and $U_{M,Y}$ as in an experimental study that manipulates *X* with random assignment.[5] *Fig. 4* depicts the consequences of conditioning on *M* in this scenario graphically and points to the reason for not observing conditional independence although full mediation is the causal mechanism.

Conditioning on $M = k$ is broadly speaking the same as selecting all those combinations of *X* and $U_{M,Y}$ that yield $M = k$ in an infinite data set, leaving *X* and $U_{M,Y}$ dependent once conditioned on $M = k$ (the dashed, double-headed arrow in *Fig. 4* indicates this dependency). That is, once *M* is fixed to a particular value *k* and a value of *X* is selected, the value of $U_{M,Y}$ is not free to vary but directly dependent on *X*, and vice versa to be consistent with $M = k$. In turn, this conditional dependence between *X* and $U_{M,Y}$ gives rise to conditional dependence between *X* and *Y* because $U_{M,Y}$ causes *Y*. As a consequence, any statistical procedure designed to test for conditional independence between *X* and *Y*, including the regression approach proposed by Baron and Kenny (1986), will consistently reject the (true) Null hypothesis that the direct effect of *X* on *Y* equals zero even though the underlying causal model in *Fig. 3* implies full mediation via *M*.[6]

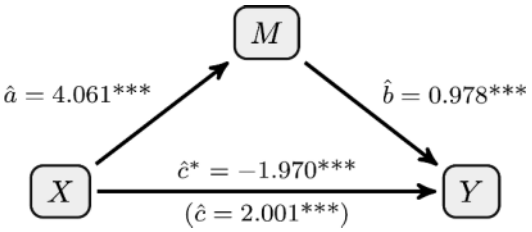


Fig. 5: Results of a mediation analysis according to Baron and Kenny (1986) for the simulated data in Section 2.1 violating conditional independence

Next, we provide a numerical illustration using a, compared to usual experimental sample sizes, large simulated sample ($N = 2000$) to showcase the consistent violation of conditional independence that results from correlated M and Y disturbances. We simulate X from a uniform distribution, generate M as $M = \text{int}_1 + X\beta_1 + U_M$ with the intercept term $\text{int}_1 = 1$ and $\beta_1 = 4$. Y is defined as $Y = \text{int}_2 + M\beta_2 + U_Y$ with $\text{int}_2 = 1$ and $\beta_2 = 0.5$. Finally, the errors $U_{M,Y} = \begin{pmatrix} U_M & U_Y \end{pmatrix}'$ are dependently distributed according to $U_{M,Y} \sim N(\mu,\Sigma)$ with $\mu' = \begin{pmatrix} 0 & 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma_M & \rho \\ \rho & \sigma_Y \end{pmatrix}$ with $\rho = 0.5$ and $\sigma_M = \sigma_Y = 1$.

Fig. 5 shows the results of analyzing the simulated data according to Baron and Kenny (1986). Full mediation according to Baron and Kenny (1986) requires statistically significant $c = \beta_1 \cdot \beta_2$, which we find ($\hat{c} = 2.001$, $t(1998) = 19.857$, $p < .001$), statistically significant $a = \beta_1$, which we also find ($\hat{a} = 4.061$, $t(1998) = 53.125$, $p < .001$), and $c^* = 0$ in the regression $Y = t_3 + Xc^* + Mb + \varepsilon_Y^*$ which we strongly reject ($\hat{c}^* = -1.970$, $t(1997) = -18.766$, $p < .001$). In combination with $\hat{b} = 0.978$ ($t(1997) = 49.433$, $p < .001$), the conclusion could be "partial mediation". The significance of the (strongly biased) estimate of the indirect effect seems to also support this conclusion (the 95 % CI from 1,000 bootstrapped samples ranges from 3.742 to 4.201 whereas the data generating value is $\beta_1 \cdot \beta_2 = 2$). However, at a causal theory level, M fully mediates X 's influence on Y and therefore an experimental procedure that fixes M would in fact fully block the causal connection between X and Y. Also note that under this DAG a is a consistent estimator of β_1, c is a consistent estimator of $\beta_1 \cdot \beta_2$, but b fails to consistently estimate β_2, which motivates IV estimation. Thus, as already seen in our example, standard estimators yield inconsistent estimates of the indirect causal effect from X to Y, even though the total causal effect is *only* through the indirect effect and can be estimated consistently. The direction of the bias in b relative to the data generating β_2 is a function of the error correlation ρ. For positive (negative) ρ, b will be larger (smaller) than β_2, biasing the inference about the indirect effect. At the same time, c^* will be biased upwards for $\rho < 0$ and biased downwards for $\rho > 0$ away from the data generating causal direct effect from X to Y denoted as β_3. If the data generating β_3 is equal to zero as in a model with full mediation, the results will misleadingly suggest "partial mediation".

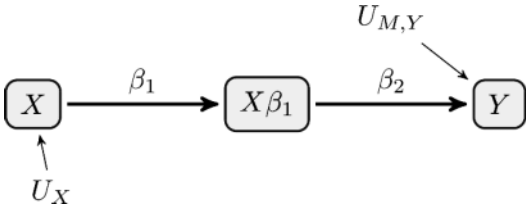


Fig. 6: IV estimation (directed acyclic graph)

IV estimation "solves" the problem posed by correlated unobservables $U_{M,Y}$ by assuming full mediation *a priori*. IV estimation proceeds by replacing M by its consistent estimate $X \cdot a = X\beta_1$ (see *Fig. 6*). Regressing Y on a consistent estimate of $X\beta_1$ then consistently estimates β_2 because $X\beta_1$ is orthogonal to $U_{M,Y}$ by construction. Note that an alternative estimation approach could exploit the relation between $c = \beta_1 \cdot \beta_2$ and $a = \beta_1$ to obtain β_2 from the ratio c/a. As we will show later, the assumption of full mediation leveraged by IV estimation defies testing which contrasts with the intent of mediation analysis.

A classical example is the estimation of price elasticities from observational data in situations where unobserved variables $U_{M,Y}$ (e. g., advertising) are strongly suspected to connect price (M) to sales (Y) variation in the data. Thereby, when regressing Y on M the pure, i. e., the direct causal effect of an exogenous price change is confounded with the effects of the unobserved variables $U_{M,Y}$. As a consequence, this regression fails to consistently estimate the effect of a price change forced upon the system from outside of the system (e. g., a price-manager independently changing the price of a product with the goal of stimulating demand). Now, if there exists an observed variable X like, for instance, the cost of product ingredients that influences prices but does not have any direct effect on sales (i. e., price fully mediates the effect of cost of product ingredients on sales), it can serve to isolate exogenous variation in price, i. e., $X \cdot a = X\beta_1$ that in turn identifies causal price effects (subject to functional form assumptions). That is, the portion of the variance in price that is exclusively explained by cost of product ingredients is used to infer the causal effect of price on sales. In applications like this, in contrast to experiments, independence between X (cost of product ingredients in the example) and the unobserved variables $U_{M,Y}$ connecting price and sales that complicate the inferential problem in the first place is an important assumption (e. g., Ebbes et al. 2016; Wooldridge 2010).[7]

Thus, IV estimation is predicated on assuming full mediation a priori. This is because the regression of Y on $X\beta_1$ fails to estimate β_2 consistently, unless M fully mediates X 's causal influence on Y. In contrast, mediation analysis aims at establishing evidence for mediation. The goal of IV estimation is a consistent estimate of β_2. This estimate enters "what-if" calculations that pertain to hypothetical experimental manipulations of M, e. g., someone "stepping into the system" and forcing, e. g., the prices in the above example to take different values by deliberate, independent management action. Somewhat in contrast,

mediation analysis focuses on measuring the indirect effect, i. e., the product of β_1 and β_2. However, while a consistent estimate of β_1 is guaranteed in the context of an experiment that randomly assigns *X*, consistent estimates of β_2, and thus a consistent estimate of the indirect effect, and of a potential direct effect are not guaranteed. Obviously, biased inferences about the indirect effect and a potential direct effect may mislead conclusions about the (psychological) process by which *X* brings about its effect on *Y* and bias the assessment of the precise theoretical contribution and appropriateness of a given experimental manipulation.

2.2. True partial mediation

It is instructive to compare *Fig. 3* to a model with true partial mediation (see *Fig. 7*). Obviously, stepping into the system in *Fig. 7* and fixing *M* to a particular value by intervention no longer fully blocks the transmission from *X*-manipulations to changes in *Y*. As another consequence, *X* and *Y* will be dependent, even conditional on *M* in data generated from this model. Importantly, the model in *Fig. 7* can rationalize any valid covariance matrix of manipulated variable *X* and observed variables *M* and *Y*. In other words, the theoretical model does not make predictions that could be falsified based on data. The model simply fits any pattern of empirical relationships between the measured variables. What may be somewhat surprising, however, is that the model in *Fig. 3* is equally flexible and therefore empirically not distinguishable from the model in *Fig. 7* based on observations of *X*, *M*, and *Y*, even if *X* is randomly assigned.[8]

The intuition for the flexibility of the model in *Fig. 3* is that we can always adjust the estimated effect from *M* to *Y* to perfectly account for the covariance between *X* and *Y*, given the effect from *X* to *M*, while perfectly matching the observed covariance between *M* and *Y* based on the error covariance in the usual linear setting (see Appendix A.2 for a formal proof). The effect from *X* to *M* only needs to rationalize the covariance between these two variables. As a consequence, these two models that differ fundamentally regarding the causal role of *M* cannot be distinguished based on the covariance-matrix of the manipulated variable *X* and observed variables *M* and *Y*.

This equivalence further burdens IV estimation with untestable assumptions. It implies that the error covariance induced by $U_{M,Y}$ and a possible direct effect from *X* to *Y* cannot be jointly identified from the data. As a consequence, IV estimation has to rely on the strong assumption that the direct effect equals zero (i. e., full mediation). Although this assumption is not directly testable, it is possible to conduct sensitivity analyses using different informative prior assumptions about a potential direct effect (Conley et al. 2012; see Imai et al. 2010a, 2010b for the analogous idea in the context of mediation analysis). Muthén et. al (2016, p. 159) show how to conduct sensitivity analysis of β_2 and β_3 for the influence of $U_{M,Y}$ using structural equation modeling and specifically the software Mplus.

For researchers applying mediation analysis, the previous arguments imply that empirical results consistent with partial mediation are difficult to interpret because the DAG shown in *Fig. 7* does not constrain the observed covariance matrix between *X*, *M*, and *Y* in a way that would unequivocally imply true partial mediation as the data generating mechanism. Moreover, as we will show in the following two sections, observed partial mediation can be produced by very different mechanisms ranging from completely spurious mediation due to a non-causal relation between *M* and *Y*, to full mediation that is misclassified as partial mediation due to measurement error in the mediator.

2.3. Seemingly partial mediation due to non-causal relations

In this section, we show that the empirical results consistent with partial mediation can occur even in the absence of any mediation through *M* (i. e., seemingly partial mediation due to non-causal relations). In particular, deleting the path from *M* to *Y* and adding a direct effect from *X* to *Y* in *Fig. 3* (see *Fig. 8*) yields yet another model that is observationally indistinguishable from those in *Fig. 3* and *Fig. 7*. However, this is a model without mediation. If one were able to step into this system and fix *M* to a particular value by intervention while experimentally manipulating *X*, *Y* would change in the same way as without the intervention that fixes *M*. Thus, in a scenario with experimental control over *M* (i. e., moderation-of-process design; Spencer et al. 2005) one would be able to

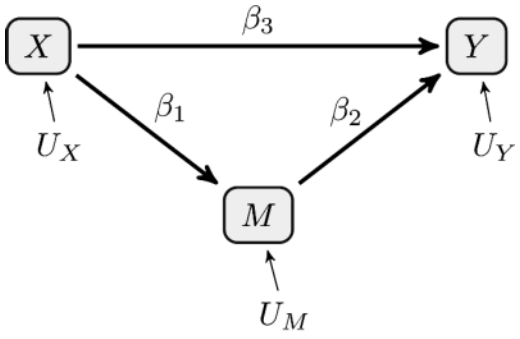


Fig. 7: Partial mediation (directed acyclic graph)

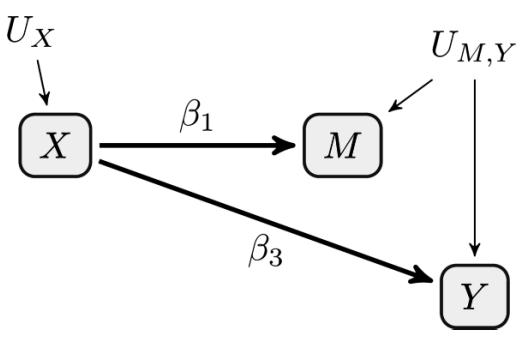
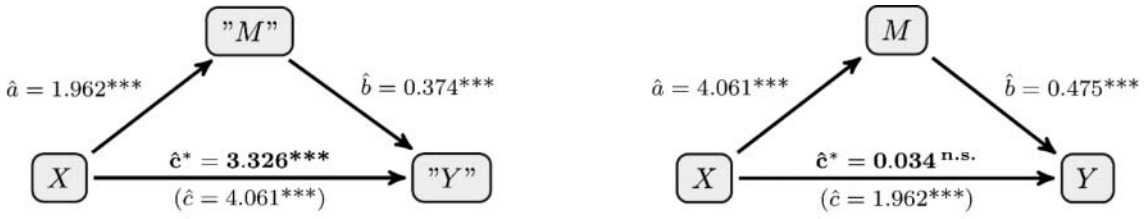


Fig. 8: No mediation (directed acyclic graph)



(a) Mistaking M for Y and vice versa *(b) Correctly labeling M and Y*

Fig. 9: Example for mistaking mediator M and outcome Y

provide direct evidence against mediation. However, in the common measurement-of-mediation design this complete lack of mediation cannot be diagnosed from the data. Applying, for instance, Baron and Kenny's procedure to data generated from this model will result in "partial mediation". And because, given the data, we can replace $U_{M,Y}$ with a direct effect from *M* to *Y* in our theoretical model without changing the fit to the empirical data, the estimated "indirect effect" (i. e., product of the path from *X* to *M* and from *M* to *Y*) will be significant given a sufficiently strong (conditional) covariance between *M* and *Y* induced by $U_{M,Y}$. In fact, there are more causal structures that produce statistical results consistent with "partial mediation" where the hypothesized *M* does not function as a mediator at all, including the situation in which the analyst mistakes the mediator *M* for the dependent variable *Y* (see also Fiedler et al. 2011; Pieters 2017).

We numerically illustrate this problem by generating data from the model in *Fig. 2* and mistaking *M* for *Y* and vice versa in the analysis. We consider the same numerical setting as in the previous illustration before in Section 2.1 but set $\rho = 0$ such that the errors U_M and U_Y are independently distributed. We then relabel the original *M* to "*Y*" and the original *Y* to "*M*". Applying the Baron and Kenny steps, we find evidence for partial mediation, indicated by a significant total effect of $\hat{c} = 4.061$ ($t(1998) = 53.125$, $p < .001$) from *X* on "*Y*", a significant effect of $\hat{a} = 1.962$ ($t(1998) = 22.795$, $p < .001$) from *X* on "*M*", and finally a significant direct effect $\hat{c}^* = 3.326$ ($t(1997) = 42.739$, $p < .001$) from *X* on "*Y*" conditional on "*M*". In combination with significant $\hat{b} = 0.374$ ($t(1997) = 20.779$, $p < .001$) the conclusion could be "partial mediation", again supported by a significant indirect effect with a bootstrapped 95 % CI ranging from 0.644 to 0.829. However, stepping into the system and experimentally fixing "*M*" (actually *Y*) to a particular value of course does not affect the causal transmission from *X* to "*Y*" (actually *M*) at all. Note that conditional independence between *Y* and *X* given *M* of course holds, when we retain the correct labeling of the variables (see *Fig. 9b*). Regressing *Y* on *X* and *M* using the correct labeling of the variables, we obtain $\hat{b} = 0.475$ ($t(1997) = 20.779$, $p < .001$) while $\hat{c}^* = 0.034$ ($t(1997) = 0.281$, $p = 0.779$) is not statistically different from zero. The bootstrapped 95 % CI of the indirect effect ranges from 1.735 to 2.112 and contains the data generating value $\beta_1 \cdot \beta_2 = 2$.

2.4. Seemingly partial mediation due to full mediation with measurement error in *M*

We next discuss another, and we believe practically important situation that results in observed conditional dependence even though full mediation holds at a conceptual, theory level – seemingly partial mediation produced by full mediation with measurement error in *M*. When the mediator *M* is not directly observed but only an indicator variable *m* subject to measurement error ε_m (see *Fig. 10*), conditioning on *m* leaves *X* and *Y* dependent, despite full mediation by *M*.[9] As a consequence, standard mediation analysis relying on Baron and Kenny (1986), including the modern variants of testing the significance of direct and indirect effects (e. g., Hayes 2013; MacKinnon et al. 2002; Preacher and Hayes 2004; Zhao et al. 2010), will consistently reject the hypothesis of full mediation, and produce results consistent with partial mediation.[10] We numerically illustrate this in Section 3.

The intuition for conditional dependence is relatively simpler in this case. Think again about the hypothetical possibility of stepping into the system and fixing *m* to a particular value by some experimental intervention. For instance, one could instruct participants to only report a particular fixed value when asked about *m* (e. g., participants could be asked to always answer "5" on a 7-point scale when asked about *m*). Under the assumption that this "superficial" instruction acts upon *m* only (i. e., it simply decouples *m* from *M* and ε_m), fixing *m* by experimental means does not interfere with the causal pathway from *X* via *M* to *Y* at all. Conditioning on $m = k$ in observed data selects combinations of *X*, U_M, and ε_m that are consistent with $m = k$. However, as a consequence, conditioning on *m* does not fix the value of *M*. Therefore, *X* and *Y* are still connected in the observed data through

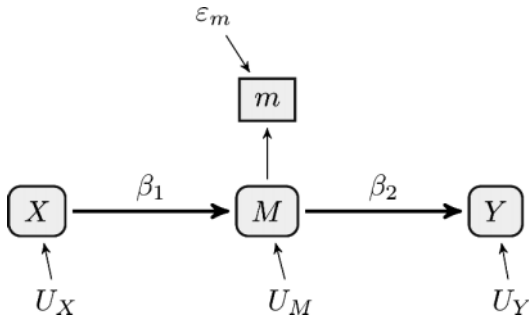


Fig. 10: Mediation with indicator variable m for unobserved mediator *M* (directed acyclic graph)

(residual) variation in M. In addition, regressing Y on m will fail to consistently estimate β_2, because the measurement error increases the variance in m unrelated to Y. Thus, common estimators and tests of indirect effects will yield biased results (see also Pieters 2017). The nature of the bias is such that b is biased towards zero. As a consequence the product $a \cdot b$ is biased towards zero too. Thus, the indirect effect will be underestimated (in absolute terms) and may not be detected at all. At the same time, c^* will be biased upwards for $\beta_1 \cdot \beta_2 > 0$ and biased downwards for $\beta_1 \cdot \beta_2 < 0$ away from the data generating causal direct effect β_3. If the data generating β_3 is equal to zero as in a model with full mediation, the results will misleadingly suggest "partial mediation". However, if one knew about full mediation in this instance *a priori*, a consistent estimate of β_2 could be obtained using IV estimation in this example, or from the ratio c/a (where c is a consistent estimator of the total effect $\beta_1 \cdot \beta_2$ obtained by regressing Y on X, and a a consistent estimator of β_1 obtained by regressing m on X).

In contrast to the models discussed earlier, the model in *Fig. 10* imposes constraints on the set of covariance matrices it can rationalize despite the fact that it features the same number of parameters to estimate. Intuitively, the effect from X to m (via unobserved M) fits $Cov(X, m)$, the product of effects from X to M and from M to Y rationalizes $Cov(X, Y)$. Finally, $Cov(m,Y)$ is a function of the former effects and increasing in $Var(U_M)$. However, because $Var(U_M)$ is bounded from below at zero, it cannot account for the influence of all direct effects from X to Y that are not included in this model. The constraint is most straightforwardly illustrated algebraically in comparison to the the standard IV setup that features unobservables $U_{M,Y}$ connecting M and Y (see Appendix A.3 for the algebraic details).

To illustrate the failure of the model with measurement error in *Fig. 10* to perfectly match the models in Figures 3 and 7, we generate observable covariances $Cov(X, M)$, $Cov(X, Y)$, and $Cov(Y,M)$ as implied by the IV model in *Fig. 3* using the same numerical specifications for parameters as before. We systematically vary the covariance of the error disturbances, ρ, in the interval (–1,1) to investigate whether the measurement error model shown in *Fig. 10* is able to explain the observed covariances for any valid specification of ρ. We find that the measurement error model can only explain the observed covariances for $\rho \in (-0.5,0)$. For $\rho < -0.5$, the measurement error model would require $Var(U_M) < 0$, and for $\rho > 0$ we would need $Var(\varepsilon_m) < 0$. These constraints make full mediation subject to measurement error a testable hypothesis relative to the unconstrained models of Figures 3 and 7.

3. Testing full mediation via conditional independence using Bayes Factors

Our discussion of data generating mechanisms exposes the ambiguity of observed results that are consistent with partial mediation. Such results may be obtained in situations where there is no mediation altogether and in situations where the hypothesis of full mediation actually holds at a causal theory level (i. e., seemingly partial mediation). In contrast, evidence for conditional independence between X and Y given M, of course assuming empirical support for a path from X to M and for a total effect from X to Y essentially constitutes evidence for full mediation, under randomly assigned X. This is because conditional independence would only result in very particular, essentially zero probability circumstances from models where full mediation is not the causal mechanism at work (e. g., a perfect balance between the influence of correlated disturbances U_M and U_Y and a direct effect from X on Y). Thus, strong evidence for conditional independence jointly rules out unobservables $U_{M,Y}$ connecting M and Y, a direct effect from X to Y, and material measurement error in M. In addition, our results suggest that the hypothesis of full mediation at a causal theory level (i. e., the absence of a direct path from X to Y) may be testable in the context of measurement error. Due to the high epistemic value of showing full mediation, we next illustrate Bayes Factors as an advantageous means to probe into the *degree of empirical support* for conditional independence.

Specifically, a drawback of relying on p-values in the context of mediation analysis is that their distribution across repeated samples under the Null-hypothesis is by construction uniform and independent of the sample size. As a consequence, the probability of rejecting a true $c^* = 0$ based on a sample is always equal to α before sampling the data and does not decrease with increasing sample size.[11] Model comparisons using Bayes Factors do not suffer from this problem but can instead provide stronger evidence in favor of the true Null hypothesis the larger the sample size. From the Bayesian perspective, a test of $c^* = 0$ is based on the comparison between two models – one with a dogmatic prior satisfying $p(c^* = 0) = 1$ (M_0) and another one with a non-degenerate, proper prior for c^* (M_1) in Equation 3. The Bayes Factor measures the relative evidence for M_0 over M_1 given observed data y (see, e. g., Rossi et al. 2005):

$$BF = \frac{p(M_0|y)}{p(M_1|y)} \tag{4}$$

The quantity $p(M|y)$ is proportional to the product of the so called marginal likelihood $p(y|M)$ and the subjective prior probability $p(M)$ assigned to a model. For uniform $p(M)$, the Bayes Factor equals the ratio of marginal likelihoods. The marginal likelihood is defined as the (normalized) likelihood function integrated with respect to the prior distribution. For the linear regression models in Equations 1–3, the Bayes Factor can be easily and reli-

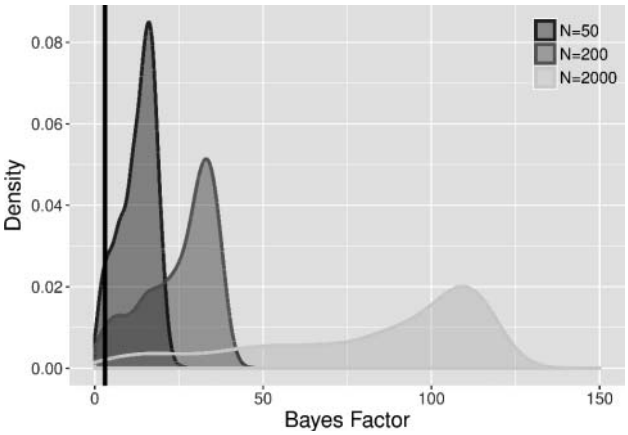


Fig. 11: Distribution of Bayes Factors testing the direct effect of X on Y after controling for M for different sample sizes in the sampling experiment. (The black vertical line at BF = 3 depicts the frontier for positive evidence in favor of M_0)

ably computed using the Savage-Dickey density ratio (see, e. g., Rossi et al. 2005):

$$BF = \frac{p(c^*|y)}{p(c^*)}\bigg|_{c^*=0} = \frac{\int p(c^*|y,\sigma_{y^*})p(\sigma_{y^*}|y)d\sigma_{y^*}}{\int p(c^*|\sigma_{y^*})p(\sigma_{y^*})d\sigma_{y^*}}\bigg|_{c^*=0}$$
$$= \frac{\mathbb{E}_{\sigma_{y^*}|y}[p(c^*|y,\sigma_{y^*})]}{\mathbb{E}_{\sigma_{y^*}}[p(c^*|\sigma_{y^*})]}\bigg|_{c^*=0} \quad (5)$$

Here, $p(c^*|y,\sigma_{y^*})$ is the conditional posterior distribution of c^* that is known in closed form for this model when using the standard conjugate normal/inverse gamma (NIG) priors. $p(c^*|\sigma_{y^*})$ denotes the conditional prior distribution for c^*.

The `bayesm`-package (Rossi 2017) for the statistical software R includes two Bayesian regression routines. The `runireg` function of the bayesm-package (Rossi 2017) works with the fully conjugate prior as alluded to here. With this prior, the marginal posterior $p(c^*|y)$ can be computed in closed form. Alternatively, the `runiregGibbs` function of the same package works with the conditionally conjugate prior, where a closed form solution for $p(c^*|y)$ is not available. No matter which of the two functions one uses, Bayes Factors (BFs) are easily and accurately approximated based on a sample of length R from the posterior distribution as follows: [12]

$$BF \approx \frac{1/R\sum_{r=1}^{R} p(c^*|y,\sigma_{y^*}^r)}{1/R\sum_{r=1}^{R} p(c^*|\sigma_{y^*}^r)}\bigg|_{c^*=0} \quad (6)$$

Intuitively, the ability of the Bayes Factor to increasingly strongly support a true Null $c^* = 0$ in larger samples comes from how the posterior $p(c^*|y)$ concentrates at 0 in larger samples relative to the prior distribution $p(c^*)$. Also note that this Bayes Factor will consistently reject a wrong Null hypothesis (i. e., when $c^* \neq 0$ the Bayes Factor will approach zero with increasing sample size), because the posterior support at the value zero converges to zero in this case. In contrast, the *p*-value is fundamentally asymmetric. It will increasingly reliably reject the Null hypothesis if it is in fact false, as a function of the sample size. However, as already mentioned, the probability of rejecting a true Null hypothesis is constant and equal to α, regardless of the sample size. By convention, Bayes Factors larger 3 count as weak but sufficient evidence in favor of the model in the numerator; Bayes Factors larger 20 count as strong evidence (Kass and Raftery 1995).

Next, we illustrate the differences between the classical and the Bayesian approach in the context of $c^* = 0$ using a sampling experiment. We first consider the case of full mediation as in *Fig. 2*. Accordingly, we set $t_2 = t_3 = 1$, $a = 4$, $c^* = 0$, $b = 0.5$, and $\sigma_M = \sigma_{Y^*} = 1$ in Equations 2 and 3, and generate artificial data sets of different sizes: $N_1 = 50$, $N_2 = 200$ as well as $N_3 = 2000$.[13] We conduct 1000 replications for each data set size.

Fig. 11 illustrates the distribution of estimated Bayes Factors over the 1000 simulation replications testing the hypothesis of $c^* = 0$, which follows from conditional independence that in turn is implied by full mediation subject to the additional assumptions discussed earlier.[14] The results in *Fig. 11* verify that the Bayes Factor correctly favors M_0 over M_1 for the vast majority of sampling replications. Importantly, this figure also illustrates that the Bayes Factor provides increasingly stronger evidence for M_0 (i. e., $c^* = 0$) as the sample size increases.

The classical testing framework based on *p*-values fails to measure the strength of evidence in favor of $c^* = 0$. In line with how they are defined, *p*-values are uniformly distributed over sampling replications in the interval of (0,1) (see *Tab. 1*). The probability of observing a *p*-value smaller than the specified significance level α is equal to this level and independent of sample size. In contrast, the probability of obtaining a Bayes Factor larger than 20 in support of $c^* = 0$ increases in N and, for example, approaches one for $N = 2000$ (see *Tab. 2*).

	Prob(p-value>0.01)	Prob(p-value>0.05)	Prob(p-value>0.1)
N=50	0.995	0.958	0.898
N=200	0.986	0.944	0.889
N=2000	0.989	0.944	0.898

Tab. 1: Replicated testing of $c^* = 0$ with 1000 replications using p-values

	Prob(BF>3)	Prob(BF>20)	Prob(BF>100)
N=50	0.943	0.045	<0.001
N=200	0.970	0.708	<0.001
N=2000	0.990	0.927	0.434

Tab. 2: Replicated testing of $c^* = 0$ with 1000 replications using Bayes Factors

Tab. 3: Comparing Bayes Factors and p-values testing the direct effect of X on Y after controlling for M measured with error for different sample sizes

	Bayes Factor	p-Value
N=50	4.660E+00	1.085E-01
N=200	7.982E-01	5.019E-03
N=2000	4.110E-18	2.932E-21

In contrast, when the data generating process implies $c^* \neq 0$, for any of the many reasons discussed earlier, *p*-values and Bayes Factors work essentially in the same way. To illustrate the problem of detecting the true data generating process when $c^* \neq 0$, consider the same data generating process as described in *Fig. 10*, where conditional independence is violated because of measurement error in the mediator. Thus, we assume that the researcher only observes m which measures M up to an error term ε_m ($\sigma_m = 1$). Otherwise we retain the same data generating values as in the earlier example. Thus, this simulation again serves to illustrate an instance, where full mediation at a causal theory level holds despite a lack of conditional independence and therefore $c^* \neq 0$.

Tab. 3 compares Bayes Factors and *p*-values for three artificial data sets with 50, 200 and 2000 observations, each generated based on the data generating process with measurement error presented in *Fig. 10*. Although the data generating mechanism implies full mediation via M, the statistical hypothesis of conditional independence between X and Y given m is rejected using *p*-values or Bayes Factors once the signal in the data as a function of the sample size is sufficiently strong. Only if the sample size (i. e., $N = 50$) and, thus, the power of the analysis is comparably small, both approaches fail to reject the Null hypothesis that $c^* = 0$. That is, the Bayesian approach finds (weak) positive support (indicated by a Bayes Factor larger than 3) for the more parsimonious model that restricts c^* to be zero. Similarly, the classical test approach also fails to reject the Null hypothesis ($p > .10$). However, for larger samples both approaches reject the Null hypothesis of conditional independence (i. e., $c^* = 0$) and point to "partial mediation". A result that per se is too ambiguous to provide theoretical guidance, especially in situations where one would have strongly expected full mediation based on theory. Hence, a drawback of currently popular testing procedures is that they do not distinguish between reasons for failing conditional independence with implications for the validity of the estimated indirect effect.

4. Discussion

Both IV estimation and mediation analysis are heavily, and sometimes mechanically used in empirical research. In fact, mediation analyses have become almost mandatory for successfully publishing papers in top consumer behavior (Pieters 2017) or social psychology journals (Fiedler et al. 2011). While the literature on IV estimation has been emphasizing the underlying (causal) assumptions quite a bit (e. g., Reiss and Wolak 2007), an extensive discussion of causal assumptions for mediation analysis only started more recently (Imai et al. 2010a, 2010b; Pearl 2014; Pieters 2017). More specifically, the current methodological literature on mediation analysis heavily focuses on improving statistical inference for establishing indirect effects (e. g., Hayes and Preacher 2014; Hayes and Scharkow 2013; MacKinnon et al. 2002; Preacher and Hayes 2008; Zhao et al. 2010) without delving into the required assumptions for interpreting these effects as actually indirect, i. e., *causal* effects.

We emphasized that the key to understanding the notion of full mediation at a causal theory level are hypothetical experimental interventions that force the mediator to take a particular value thereby blocking the causal pathway between X and Y (Pearl 2009). The moderation-of-process design introduced by Spencer et al. (2005) essentially implements this idea advocating designs that both manipulate, and randomly assign X and M, i. e., force M to take different values (at least in expectation) by design. While we very much agree with this idea, we believe that measurement-of-mediation designs (Spencer et al. 2005) proposed by Baron and Kenny (1986), i. e., designs in which the hypothesized mediator is not subject to experimental manipulation but simply observed, will continue to play an important role. We then discussed and clarified the relationship between observable aspects of the data, often taken to be diagnostic of the underlying process, and the underlying causal process, and in particular the relationship between conditional independence and full mediation.

Extending our technical discussion of the merits of full mediation, we would now like to point to the epistemic and applied advantages of a complete process understanding (i. e., full mediation). Specifically, the goal of mediation analysis is to shed light on the process by which experimentally manipulated X brings about an effect on Y. An intuitive understanding of what can be accomplished in principle based on a better process understanding comes, for instance, from medical science. Suppose that a scientist managed to identify the mechanism in the nervous system (M) that translates physical harm (X) into subjective feelings of pain (Y). Such a process understanding would enable the scientist to develop anesthetic drugs that prevent feelings of pain in the presence of physical harm. In other words, the process understanding enables the identification of a moderator that prevents an effect from occurring. Importantly, the anesthetic drugs are likely to work only if the scientist achieved a complete process understanding (i. e., full mediation). If unidentified parallel pathways exist that pass on the signal triggering the pain (i. e., partial mediation), the anesthetic drugs will not work effectively. This example illustrates how an incomplete theoretical under-

standing of the mediating process is considerably less useful when planning interventions. We therefore argue that evidence supporting full mediation at a causal theory level should be the goal of, if not the yardstick for research aimed at establishing process knowledge. Furthermore, we conjecture that the goal of empirically establishing full mediation might have become to appear beyond reach because of an overreaching interpretation of finding $c^* \neq 0$ in empirical applications.

Specifically, we showed that there are data generating mechanisms that are plausibly at work in many empirical settings, where full mediation holds at a causal theory level. However, a test of (full) mediation built on conditional independence between *X* and *Y* given *M* may consistently reject the Null hypothesis that $c^* = 0$, even if full mediation holds at a causal theory level. In addition, standard estimates of indirect effects are biased when applied to data generated from these models. Therefore, the focus on statistically reliable estimation of the (putative) indirect effect fails to resolve the underlying specification problem.

One such model is the DAG shown in *Fig. 3* that motivates IV estimation, where unobserved variables connect the disturbances U_M and U_Y. Imai et al. (2010b) use the experiment by Nelson et al. (1997) as an exemplary illustration. In this experiment, the manipulated *X* variable (message framing in a local news story: Ku Klux Klan rally as a free speech issue vs. Ku Klux Klan rally as a disruption of public order) has an effect on the dependent variable *Y* (tolerance for the Ku Klux Klan) that is mediated by a variable *M* (perceived importance of free speech/public order). However, the mediator *M* and the dependent variable *Y* are likely related over and above a putative causal link transmitting the effect of the experimental manipulation *X* to the dependent variable *Y*. Imai et al. (2010b) refer to general political attitudes or socialization that may connect *M* and *Y* beyond the causal link. Hence, because the mediator is not experimentally manipulated, some of the (residual) variation in *M* after conditioning on *X* (i. e., U_M) likely originates from general political attitudes that are also reflected by the tolerance for the Ku Klux Klan (*Y*). In other words, general political attitudes may act as an unobserved background factor to both *M* and *Y*. Imai et al. (2010b) propose sensitivity analysis of the mediation results to the level of dependence between *M* and *Y* from unobserved joint causes. The need to resort to sensitivity analyses arises because the IV DAG (*Fig. 3*) and the partial mediation DAG (*Fig. 7*) are observationally indistinguishable without additional assumptions.

We note that unobserved substantive background factors may be more likely a priori in some studies than in others. Studies like the experiment by Nelson et al. (1997), where participants have a strong a priori predisposition to answer the questions measuring mediator and dependent variable in a particular way given their opinions, are likely to be affected by an unobserved connection between the disturbances U_M and U_Y. In contrast, studies that use a priori meaningless stimuli like, e. g., abstract visual patterns that need to be evaluated conditional on some experimental manipulation (e. g., Graf and Landwehr 2017), are less likely to suffer from unobserved connections between the disturbances U_M and U_Y. For example, a study that manipulates the frequency of exposure to unknown Chinese characters (*X*) and measures whether the effect of mere exposure on liking of the characters (*Y*) is mediated by processing fluency (*M*) is unlikely to be affected by substantive unobserved background factors connecting fluency *M* to liking *Y* (for an example, see Landwehr et al. 2017).

However, a different reason for rejecting conditional independence despite full mediation at a causal theory level that seems generally important is measurement error in the mediator (see also Baron and Kenny 1986; Pieters 2017). For example, the described study on processing fluency as a mediating variable likely suffers from measurement error in *M*, simply because one cannot measure a cognitive mechanism like processing fluency directly but has to rely on appropriate measurement scales (Graf et al. in press). While measurement error in the mediator will cause the consistent rejection of conditional independence and bias estimates of the indirect effect, it is *not* observationally equivalent to a model with partial mediation or the IV model, as we have shown. We therefore believe that it will be useful to develop tests using the framework of Bayesian model comparisons that quantify the evidence supporting full mediation in the context of measurement error relative to a model that does not constrain observed covariances at all (i. e., true partial mediation or the IV model). Bayesian model comparisons are called for because the model with measurement error in the mediator features the same number of parameters as the IV model or the partial mediation model when the mediator is measured with a single indicator. The constraints embedded in the model with measurement error are ordinal and thus hard to assess using non-Bayesian inference.

In addition, there is scope for meaningful model comparisons based on theoretically motivated sign constraints. For example, if a researcher or a reviewer names a potential omitted background factor connecting *M* to *Y*, he will very likely (and should) be able to a priori sign-constrain the influence of this background factor (i. e., positive versus negative relationship). Say, the omitted background factor contributes positively to the observed covariance between *M* and *Y*. Upon building the sign constrained correlation between disturbances into the model, the sign-constrained IV DAG, for example, may no longer be able to rationalize a positive direct effect from *X* to *Y*. Remember that the effect from *X* to *M* (β_1 in the DAGs) essentially captures the covariance between these two variables. Further assume that this effect is positive and in line with theory. When we do not allow for a direct effect from *X* to *Y* in the model, the directed connection between *M* and *Y* (i. e., β_2) will adjust to reproduce

the covariance between X and Y. If this covariance is large because of a direct effect that is not accounted for in the model, β_2 will be too large vis-a-vis the observed covariance between M and Y, and only negative dependence between U_M and U_Y could compensate for this. However, negative dependence between U_M and U_Y is ruled out by the a priori assumptions about the relation which in turn makes the absence of a positive direct effect from X onto Y a testable proposition. What is required here is, on the one hand, a priori theory to sign-constrain the directed effects and the potential dependence between U_M and U_Y, which should be easy to come by for researchers that are actually looking to corroborate a process explanation and not just fishing for one. On the other hand, one needs estimation routines that accommodate prior sign constraints and suitable measures for model comparisons. Even though this requires substantial additional development efforts, the Bayesian approach, in principle, accommodates these technical requirements.

Extending this call for the development of more powerful testing procedures, we believe that the practice of mediation analysis that is aimed at establishing positive evidence for a process explanation will benefit from the formulation of additional testable hypotheses. In particular, the empirical observation of "partial mediation" requires further examination of the underlying data generating process instead of instantly triggering a chase for additional mediators because, as we have shown, an empirical pattern consistent with partial mediation can be due to multiple DAGs and incorrect causal inferences are likely to occur.

In this regard, we also believe that the goal for a useful process based explanation is full mediation at a causal theory level. However, while conditional independence between X and Y given M is a powerful affirmative indication of full mediation, at least when obtained in large samples and with X randomly assigned, there are a number of reasons for failing to observe conditional independence, even if full mediation holds at a causal theory level. We have clarified some, we believe plausible, reasons for this to occur. It remains an empirical issue to show that the hypothesis of full mediation can, for example, be supported once measurement error in the mediator is accounted for.

5. Recommendations and Limitations

Given the ambiguous nature of an empirical result consistent with partial mediation, what should a researcher now do different from claiming a significant indirect effect and additional, unobserved mediators in applications that yield a "partial mediation" result? Before the Bayesian model comparisons we advocate in our discussion become available, we suggest that researchers probe into the plausibility of alternative causal structures before advancing their favorite interpretations. There are several ways to do this. A simple approach is to take the observed variance-covariance matrix of $\{X,M(m),Y\}$ (we use the notation $M(m)$ to indicate the likely uncertainty about the measurement quality of the mediator) and to solve for the parameters in alternative models, notably the IV model and the model with measurement error in the mediator, under the assumption of full mediation at a causal theory level ($\beta_3 = 0$), see Appendices A2. and A.3.

In applications where unobservables $U_{M,Y}$ seem likely a priori, it should be useful to see how large the error correlation needs to be for the direct effect to become zero, and how much the estimate of β_2 and thus the conclusion about the indirect effect change. A more elegant way to do this is use structural equation modeling software. For example, Muthén et. al (2016, p. 159) show how to analyze the sensitivity of $\hat{b}$ as an estimate of β_2 and $\hat{c}^*$ as an estimate of β_3 against different levels of error correlations induced by $U_{M,Y}$.

In applications where measurement error in the mediator is a likely concern, the sign of $\hat{c}^*$ relative to the sign of $\hat{a} \cdot \hat{b}$ can be used to heuristically assess the possibility of full mediation subject to measurement error. Because measurement error in the mediator biases $\hat{b}$ towards zero, the bias in $\hat{c}^*$ will be positive for positive $\hat{a} \cdot \hat{b}$ and negative for negative $\hat{a} \cdot \hat{b}$. Therefore, a negative $\hat{c}^*$ in combination with positive $\hat{a} \cdot \hat{b}$, for example, cannot be reconciled with full mediation subject to measurement error. And again, one can use the covariance algebra in Appendix A.3 or structural equation modeling software to solve for the parameters in a model with measurement error under the assumption of full mediation at a causal theory level, and check if i) all variance terms are positive and if ii) the estimated amount of measurement error appears reasonable.

The goal of this paper was to re-emphasize the topic of model specification in the context of mediation analysis and to highlight the ambiguous nature of results that are consistent with partial mediation, both with respect to the existence of a direct causal effect and with respect to correct inference about the indirect effect. While none of our individual results are genuinely new, they are often only discussed on the side or as "special topics" in application oriented discussions of mediation analysis. In contrast, the identifiability of causal ("actual", "real") effects and thus model specification are central topics in the newer, more technical literature on causal inference in mediation analysis. With this article we hope to contribute to the acute awareness of model specification issues in the wider community of researchers that rely on mediation analysis for their substantive research.

Finally, we have deliberately concentrated on the simplest case of three variables X, M (or m), and Y in this paper to keep the arguments as simple and straightforward as possible. However, model specification is at least equally important in the context of, for example, multiple mediators, moderated mediation, or multiple indicator measurement.

Notes

[1] Some authors argue that a total effect different from zero is not required for mediation analysis to be valid (see e. g., Rucker et al. 2011; Zhao et al. 2010). At a conceptual level, we disagree because perfect cancellation between multiple paths is an event of measure zero, i. e., has zero prior probability under any non-degenerate continuous prior distribution for parameters. However, if a researcher hypothesizes the possibility of a near zero effect because of, say, two indirect paths of opposing signs a priori, and the small total effect fails statistical significance in a given sample, the researcher should of course proceed with the analysis. In contrast, the results of a search for mediators after finding a non-significant total effect should be interpreted cautiously, unless full mediation can be established in this search.

[2] Furthermore, evidence for $c \neq 0$ and for $a \neq 0$ implies that $b \neq 0$ under these assumptions. This is because both c and a are measures of causal effects and $b = 0$ would contradict the existence of an effect already established under this set of assumptions.

[3] In the limit of an infinite amount of data, the estimate of c^* will only converge to exactly zero under full mediation. The only alternative process that yields $c^* = 0$ in the limit features M as a joint cause of X and Y without another connection between X and Y. This process is ruled out a priori, when X is experimentally manipulated.

[4] Because of the focus on mediation and to avoid confusion, we keep with the variable names (*X,M,Y*) throughout, instead of (*Z,X,Y*) that appears to be more common in the context of IV estimation.

[5] In the typical application of IV estimation to observational data, the assumption of independence between U_X and $U_{M,Y}$ required for X to be a "valid instrument" has to be defended based on theory.

[6] Note that conditioning on M of course also yields dependence between U_X and U_M in *Fig. 2*. However, because U_M and U_Y are independent in *Fig. 2*, conditional independence between X and Y is preserved.

[7] Zhang et al. (2009) propose to use so called "latent instruments" in the context of mediation analysis. Thus, they propose to replace M by its instrumented estimate $\tilde{M}$ where the instrument is not X but another variable. A drawback of their approach, and any other approach that conditions on thus instrumented estimates $\tilde{M}$ of M in mediation analysis, is that estimates of direct effects will be biased away from zero. Therefore, while instrumenting M using an additional variable yields an unbiased estimate of β_2 and thus an unbiased estimate of the indirect effect, it biases the estimate of the direct effect.

[8] Truly categorical mediators pose an exception to this perfect observational equivalence (see Pearl 1995) that we omit from our discussion.

[9] The path connecting M to m in this DAG is fixed to one for identification.

[10] Note that an analogous problem occurs when the actual (unobserved) mediator is continuous but only observed categorically, e. g., on a rating scale.

[11] Conditional on an observed sample, the p-value is obviously just a function of the data.

[12] Appendix A.4 illustrates how to run the test of $c^* = 0$ in **R** for a simulated data set based on *Fig. 2*.

[13] X_i is drawn from a uniform distribution for each $i \in \{1, \ldots, N\}$.

[14] Here we use `runireg`, i. e., the fully conjugate *NIG*-prior. We use standard weakly informative prior settings (see Rossi et al. 2005) and use 5000 draws from the posterior.

A. Appendix

A.1. Mediation according to Baron and Kenny (1986) in R

```
# NOTE: We first simulate data assuming full mediation.

# Set seed (i. e., make results reproducible).
set.seed(77)

# Data generating function: M fully mediates X
simMediation = function(beta_1,beta_2,int_1,int_2, sigma_M, sigma_Y,N,X) {
  eps_M = rnorm(N)*sigma_M^.5  # generate errors for M (independent)
  eps_Y = rnorm(N)*sigma_Y^.5  # generate errors for Y (independent)
  M = int_1 + beta_1 * X + eps_M  # generate latent mediator M
  Y = int_2 + beta_2 * M + eps_Y  # generate dependend variable
  list(X = X, M = M, Y = Y)
}

# Set up data generating parameters
N = 2000      # number of observations
sigma_M = 1   # error variance M
sigma_Y = 1   # error variance Y
beta_1 = 4    # beta_1 (i. e., parameter a according to Baron & Kenny)
beta_2 = .5   # beta_2 (i. e., parameter b)
int_1 = 1     # intercept for equation M on X
int_2 = 1     # intercpt for equation Y on M
X = runif(N)  # generate random X from a uniform distribution
```

```
# Generate data based on parameters
datsimM = simMediation(beta_1,beta_2,int_1,int_2,sigma_M,sigma_Y,N,X)

# Put data into dataframe
df <- data.frame(y = datsimM$Y, m = datsimM$M, x = datsimM$X)


# Mediation according to Baron and Kenny (1986)
#
# 1st equation: Show correlation between causal variable X and outcome Y.
summary(lm(y ~ x, data=df))

##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0716 -0.7185 -0.0052  0.7095  3.7251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.47036    0.04918   29.90   <2e-16 ***
## x            2.01812    0.08533   23.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.089 on 1998 degrees of freedom
## Multiple R-squared:  0.2187, Adjusted R-squared:  0.2183
## F-statistic: 559.3 on 1 and 1998 DF,  p-value: < 2.2e-16

#
# 2nd equation: Show correlation between causal variable X and mediator M.
summary(lm(m ~ x, data=df))

##
## Call:
## lm(formula = m ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0976 -0.6509 -0.0039  0.6512  3.6503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00954    0.04462   22.63   <2e-16 ***
## x            4.01341    0.07742   51.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9885 on 1998 degrees of freedom
## Multiple R-squared:  0.5736, Adjusted R-squared:  0.5734
## F-statistic:  2687 on 1 and 1998 DF,  p-value: < 2.2e-16

#
# 3rd equation: Show that mediator M affects outcome Y (i. e., effect is non-zero)
#               and show that M completely mediates X-Y relationship (i. e., effect
#               of X on Y controlling for M is zero, conditional independence).
summary(lm(y ~ x + m, data=df))

##
## Call:
## lm(formula = y ~ x + m, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2762 -0.6682 -0.0221  0.6705  3.3963
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.96985    0.04924  19.696   <2e-16 ***
## x            0.02834    0.11674   0.243    0.808
## m            0.49578    0.02203  22.506   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9733 on 1997 degrees of freedom
## Multiple R-squared:  0.3768, Adjusted R-squared:  0.3762
## F-statistic: 603.7 on 2 and 1997 DF,  p-value: < 2.2e-16
```

A.2. Properties of the partial mediation and IV models

We first show how the model of partial mediation in *Fig. 7* rationalizes the covariance structure between manipulated X and observed M and Y. We can write down the linear equations corresponding to *Fig. 7* as follows:

$$M = \beta_1 X + U_M$$
$$Y = \beta_2 M + \beta_3 X + U_Y$$

We can then match the coefficients with three observed covariances:

$$Cov(X,M) = \beta_1 Var(X)$$
$$Cov(X,Y) = (\beta_2\beta_1 + \beta_3)Var(X)$$
$$Cov(Y,M) = \beta_2(\beta_1^2 Var(X) + Var(U_M)) + \beta_1\beta_3 Var(X), \textit{ since } Cov(U_Y, U_M) = 0$$

Note that $Var(U_M)$ is directly identified since $Var(M)$ is observed, i. e., $Var(U_M) = Var(M) - (\beta_1)^2 Var(X)$.

The system of linear equations corresponding to the IV model in *Fig. 3* is given as follows

$$M = \beta_1 X + U_M$$
$$Y = \beta_2 M + U_Y,$$

with $U_{M,Y} = \begin{pmatrix} U_M & U_Y \end{pmatrix}'$. We can match the coefficients with the observed covariances in the same fashion as before:

$$Cov(X,M) = \beta_1 Var(X)$$
$$Cov(X,Y) = (\beta_2\beta_1)Var(X)$$
$$Cov(Y,M) = \beta_2(\beta_1^2\ Var(X) + Var(U_M)) + Cov(U_Y, U_M)$$

with $Cov(U_Y,U_M) \neq 0$ in IV applications. Both in the IV model and the partial mediation model observed $Cov(X,M)$ together with observed $Var(X)$ identify β_1. The remaining two covariances need to identify two additional parameters in both models, β_2 and β_3 in the partial mediation model, and β_2 and $Cov(U_Y,U_M)$ in the case of the IV model. In the absence of prior constraints on these parameters, both models are just identified and will fit any valid covariance matrix of X, M, and Y perfectly.

A.3. Properties of the model with measurement error

The set of linear equations is given as follows for the model with measurement error in the mediator represented by *Fig. 10*:

$$M = \beta_1 X + U_M$$
$$m = M + \varepsilon_m$$
$$Y = \beta_2 M + U_Y$$

The coefficients are then again matched with three observed covariances:

$$Cov(X,m) = \beta_1 Var(X)$$
$$Cov(X,Y) = (\beta_2\beta_1)Var(X)$$
$$Cov(Y,m) = \beta_2(\beta_1^2\ Var(X) + Var(U_M))$$

Note that in the case with measurement error $Var(U_M)$ is no longer directly determined by the data because observed $Var(m)$ is a function of both $Var(U_M)$ and measurement error $Var(\varepsilon_m)$, i. e., $Var(m) = (\beta_1)^2\ Var(X) + Var(U_M) + Var(\varepsilon_m)$. However, the model is relatively more constrained than the IV model or the partial mediation model because both unobserved variance terms need to be positive.

A.4. Bayes Factors in R

```
# NOTE: we use the parameters and same simulated data (datsimM)
#       that we use in Appendix A.1.

# add an intercept for X (needed by runireg)
data_runireg = list(y = datsimM$Y, X = cbind(rep(1,N),datsimM$X,datsimM$M))


# Run MCMC (stay with default priors)
Mcmc = list(R = 10000)
out_runireg = runireg(Data = data_runireg, Mcmc = Mcmc)

# load functions to compute Compute Bayes Factor (BF)
# see Appendix A.5
source("Compute_BF_Savage_Dickey.R")
constraint = c(0,1,0) # indicate which parameter is tested to be zero
# Bayes Factor
BF_savage_dickey(out_runireg, data_runireg, constraint)

## [1] 112.0747
```

A.5. Savage-Dickey Bayes Factor estimator in R

```
# Contents of file 'Compute_BF_Savage_Dickey.R' (cf. Appendix A.4)


# Function to evaluate density at constraints given draws of sigma
# 'constraint' indicates which entries are set to zero
den_atconstraints_BF <- function(draws, btilde, constraint, X, A) {

    R = dim(draws$betadraw)[1]
    draw_den = array(0, dim = c(R, 1))
    constraint_ind = seq(1, length(constraint)) * constraint
    constraint_ind = constraint_ind[constraint_ind != 0]
    Factor_sigma = t(X) %*% X + A
    Factor_sigma = chol2inv(chol(Factor_sigma))

    for (r in 1:R) {
        sigma_updated = Factor_sigma * draws$sigmasqdraw[r]
        sigma_updated = sigma_updated[constraint_ind, constraint_ind]
        draw_den[r] = lndMvnvec(rep(0, sum(constraint)), btilde[constraint_ind],
            solve(chol(sigma_updated)))
    }
    draw_den
}

# Function to draw from prior
den_atprior_BF <- function(R, constraint, nu, ssq, A, betabar) {

    draw_den = array(0, dim = c(R, 1))
    constraint_ind = seq(1, length(constraint)) * constraint
    constraint_ind = constraint_ind[constraint_ind != 0]

    for (r in 1:R) {
        sigmasq_prior = (nu * ssq)/rchisq(1, nu)
        sigmasq_prior_updated = chol2inv(chol(A)) * sigmasq_prior
        sigmasq_prior_updated = sigmasq_prior_updated[constraint_ind, constraint_ind]
        draw_den[r] = lndMvnvec(rep(0, sum(constraint)), betabar[constraint_ind],
            solve(chol(sigmasq_prior_updated)))
    }
    draw_den
}
```

```r
# Define functions to compute the Bayes Factor (BF)
lndMvnvec <- function(x, mu, rooti)
# a vectorized version of lndMvn x is of dimension p times R, mu is a vector
# of length p, rooti is the inverse of the cholesky factor of the p times p
# covariance; the function returns a vector of normal densities on the
# log-scale of length R
{
    z = t(rooti) %*% (x - mu)
    return(-(length(x)/2) * log(2 * pi) - 0.5 * colSums(z * z) + sum(log(diag(rooti)
)))
}

# Final function to compute Bayes Factor a la Savage-Dickey
BF_savage_dickey <- function(draws, Data_xm, constraint) {
    # Do computations
    nvar = dim(draws$betadraw)[2]
    A = 0.01 * diag(nvar)
    betabar = rep(0, nvar)
    RA = chol(A)
    X = Data_xm$X  #cbind(x,datsimM$m)
    W = rbind(X, RA)
    Z = c(Data_xm$y, as.vector(RA %*% betabar))
    IR = backsolve(chol(crossprod(W)), diag(nvar))
    btilde = crossprod(t(IR)) %*% crossprod(W, Z)

    # Evaluate numerator
    BF_numerator = den_atconstraints_BF(draws, btilde, constraint, X, A)

    # Denominator now
    nu = 3
    ssq = var(Data_xm$y)
    betabar = c(rep(0, nvar))

    BF_denominator = den_atprior_BF(dim(draws$betadraw)[1], constraint, nu,
        ssq, A, betabar)

    # Bayes Factor
    BF = mean(exp(BF_numerator))/mean(exp(BF_denominator))

    # Return results
    return(BF)
}
```

References

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

Conley, T. G., Hansen, C. B., & Rossi, P. E. (2012). Plausibly exogenous. *The Review of Economics and Statistics*, *94*(1), 260–272.

Demming, C. L., Jahn, S., & Boztug, Y. (2017). Conducting mediation analysis in marketing research. *Marketing ZFP*, *39*(3), 76–98.

Ebbes, P., Papies, D., & van Heerde, H. J. (2016). Dealing with endogeneity: A nontechnical guide for marketing researchers. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of Market Research*. Heidelberg: Springer International Publishing.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*(6), 1231–1236.

Graf, L. K. M., & Landwehr, J. R. (2017). Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking. *Frontiers in Psychology*, *8*(15).

Graf, L. K. M., Mayer, S., & Landwehr, J. R. (in press). Measuring processing fluency: One versus five items. *Journal of Consumer Psychology.*

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*, 451–470.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis. *Psychological Science*, *24*(10), 1918–1927.

Imai, K., Keele, L., & Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

Imai, K., Keele, L., & Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Landwehr, J. R., Golla, B., & Reber, R. (2017). Processing fluency: An inevitable side effect of evaluative conditioning. *Journal of Experimental Social Psychology*, *70*, 124–128.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using mplus*. Los Angeles: Muthén & Muthén.
Nelson, T. E., Clawson, R. A., & Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, *91*(3), 567–583.
Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 435–443). San Mateo: Morgan Kaufmann.
Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge, England: Cambridge University Press.
Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459–481.
Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, *44*(3), 692–716.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods*, *36*(4), 717–731.
Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891.
Reiss, P. C., & Wolak, F. A. (2007). Structural economic modeling: Rationales and examples from industrial organization. In *Handbook of Econometrics* (Vol. 6A, pp. 4277–4415). Elsevier.
Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*.
Rossi, P. E. (2017). *Bayesm: Bayesian inference for marketing/micro-econometrics*. https://CRAN.R-project.org/package=bayesm
Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. John Wiley & Sons.
Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, *5*(6), 359–371.
Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845–851.
Wooldridge, J. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.
Zhang, J., Wedel, M., & Pieters, R. (2009). Sales effects of attention to feature advertisements: A bayesian mediation analysis. *Journal of Marketing Research*, *46*(5), 669–681.
Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering baron and kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197–206.

Keywords

Mediation Analysis, Instrumental Variables, Bayes Factor, Causal Inference, Consumer Psychology.