Journal of Experimental Social Psychology 75 (2018) 95–102

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Unwarranted inferences from statistical mediation tests – An analysis of articles published in 2015☆

Klaus Fiedler*, Chris Harris, Malte Schott

University of Heidelberg, Germany

A B S T R A C T

Recent attempts to improve on the quality of psychological research focus on good practices required for statistical significance testing. The scrutiny of theoretical reasoning, though superordinate, is largely neglected, as exemplified here in a common misunderstanding of mediation analysis. Although a test of a mediation model X ➔ Z ➔ Y is conditional on the premise that the model applies, alternative mediators Z′, Z″, Z‴ etc. remain untested, and other causal models could underlie the correlation between X, Y, Z, researchers infer from a single significant mediation test that they have identified the true mediator. A literature search of all mediation analyses published in 2015 in Sciencedirect shows that the vast majority of studies neither consider alternative causal models nor alternative mediator candidates. Ignoring that mediation analysis is conditional on the truth of the focal mediation model, they pretend to have demonstrated that Z mediates the influence of X on Y. Recommendations are provided for how to overcome this dissatisfying state of affairs.

1. Introduction

A growing number of recent publications are driven by the laudable motive to improve the scrutiny of psychological science. How can we foster solid research findings that are reliable and replicable at the empirical level and well understood at the theoretical level? A glance at the pertinent literature shows that the suggested interventions and the implemented changes in the publication process focus on data sharing, good practices in documentation, and appropriate significance testing. Accordingly, the key to improved science seems to lie in stricter compliance rules for data management and in still more weight given to proper significance testing. There is a conspicuous paucity of discourse on strict theorizing and logic of science (Fiedler, 2017).

In this article, we emphasize the need for proper theorizing and the priority of theoretical reasoning as a major precondition of good science. Research design beats statistical testing, and theoretical reasoning beats research design. Even sophisticated statistical testing is worth nothing if the underlying research design is flawed. And the cleverest and most refined design is useless when applied to a logically inappropriate or undiagnostic hypothesis.

While there are many ways to substantiate this point, the present article concentrates on one issue, namely, the reliance on mediation analysis when drawing scientific inferences. Testing mediation models has become a gold standard for research submitted to prominent journals (Bullock, Green, & Ha, 2010; MacKinnon, Fairchild, & Fritz, 2007). It is supposed to enable rigorous process accounts of otherwise unexciting findings and to allow for causal inferences about what crucial factor mediates the influence of an independent on a dependent variable.

The present article is neither meant to deny the scientific potential of mediation analysis nor to criticize the pertinent statistical methods (cf. Hayes, 2009; MacKinnon, 2008). We are simply concerned with the scientific status of theoretical inferences informed by mediation analysis. Drawing on a universe of 102 articles (126 mediation analyses) solicited by the keyword “mediation analysis” in the internet platform Sciencedirect,[1] we demonstrate that the vast majority of theoretical inferences drawn from such mediation tests are logically unwarranted. Even when state-of-the art statistical procedures for mediation analysis are applied to well designed and carefully conducted experiments, often published in high-ranking journals, most theoretical inferences and practical take-home messages are misleading and logically incorrect.

☆ The work underlying the present article was supported by a grant provided by the Deutsche Forschungsgemeinschaft to the first author (FI 294/23-1).

* Corresponding author.
E-mail address: kf@psychologie.uni-heidelberg.de (K. Fiedler).

[1] The first author's experience as an Associate Editor of this journal, which is covered on the Science Direct platform, originally motivated this critical perspective on mediation analysis (cf. Fiedler, Schott, & Meiser, 2011).

https://doi.org/10.1016/j.jesp.2017.11.008
Received 31 January 2017; Received in revised form 25 September 2017; Accepted 16 November 2017
Available online 08 December 2017
0022-1031/

1.1. What inferences can(not) be informed by mediation tests

Let us illustrate the problem with a thought experiment. Imagine an epidemiologist has found that the emergence of a virus (X) is statistically related to the observation of a disease (Y). The epidemiologist holds a biologically well-founded theory that infection is transmitted by sexual contact (Z); sexual contact is the means by which the virus can infect other people. Indeed, when Z is entered as a third variable in a mediation test, a substantial part of the covariance shared by X and Y is explained by the model X ➔ Z ➔ Y. In this case, a statistical mediation test actually substantiates a causal model, which is reasonable on theoretical grounds.

However, now suppose that the correlations between X, Y, and Z are exactly the same but Z is fever (a symptom of Y) rather than sexual contact (a reasonable infection mechanism). We know, theoretically, that fever is not a mediator, but Sobel tests, regression analysis, or a bootstrapping algorithms do not have causal world knowledge; they are only sensitive to the tri-variate data array but not to the causal surplus meaning of a symptom (fever) versus an infection mechanism (sexual contact). Thus, when fed with the same correlation pattern, the significant test might be mistaken to imply that fever mediates the relation between virus and disease based on the unwarranted assumption that causality can be inferred inductively from a statistical test.

The example highlights the priority of theoretical over statistical reasoning. By the same token, a significant result for sexual contact as a third variable can be reframed theoretically as a moderator rather than a mediator effect. Then, the virus is transmitted only among sexually active people, but not among sexually abstinent people. Choosing between moderator (person groups) or mediator accounts (contagion mechanism) is an essentially theoretical problem that cannot be solved statistically (cf. Fiedler, Walther, Freytag, & Stryczek, 2002).

Although the epidemiological example is clearly relevant to health psychology, it may be worthwhile illustrating the same point with a genuinely social psychological example: The same tri-variate covariance pattern allows for several theoretical interpretations of the role of Z relative to X and Y. For instance, the cognitive responses Z (pro or contra responses) to a persuasive communication in a thought-listing task are often interpreted as a mediator of the impact of a persuasive message X on a changing attitude Y. But Z may be conceived as another measure of the dependent variable, attitude change. Or, it may be framed as a moderator, restricting attitude change to those people who engage in active cognitive responses to the message content.

The viability of different mediation models can vary strongly on a-priori grounds. Thus, the encoding strategy applied to a persuasive message (e.g., trying to generate few or many arguments or counter-arguments; Tormala, Falces, Briñol, & Petty, 2007) logically affords a more viable candidate for a mediator variable than an enduring personality attribute (e.g., expert knowledge) that existed long before the persuasive attempt (as explained by Tate, 2015).

As a rule, statistical mediation tests are contingent on the validity of the mediation model (Waldmann, Hagmayer, & Blaisdell, 2006) and choosing an appropriate causal model is essentially a theoretical issue, not a statistical one. Therefore, if the causal model makes sense theoretically and logically, convincing and elucidating mediation analyses can be simple and straightforward. For a simple demonstration, take the finding that positive testing mediates the genesis of confirmation biases (Fiedler, Freytag, & Unkelbach, 2007). Most participants in a simulated classroom setting who were asked to test the hypothesis that girls are good in language whereas boys are good in science engaged in positive testing strategies. That is, they asked girls more questions in language classes and boys more questions in science. This difference in sample size was sufficient to subjectively confirm the hypothesis, even though boys and girls did not differ in the relative rate of correct responses in either discipline. The confirmation bias fully disappeared for the minority of participants who did not engage in positive testing search strategies. When the supposed mediator was manipulated experimentally, such that sample size was larger for boys in language and for girls in science, the resulting bias was reversed, thus ruling out a common gender stereotype as an alternative mediator.[2]

However, while a theoretically plausible top-down model can render mediation analysis convincing, bottom-up inferences from statistically significant ad-hoc tests are logically flawed. It is a category mistake to infer from a significant mediation test that "Z mediated the influence of X on Y". Just as it is inappropriate to infer the truth of H_1 from a significant result, or its falsehood from non-significance (Trafimow, 2003), it is particularly wrong to infer the causal status of Z from a significant test result of a mediation model X ➔ Z ➔ Y. Such a model test can happen to be significant for many other reasons than Z being the true mediator.

1.2. Two sources of uncertainty

On the one hand, it is a truism that for every correlation between two variables it is possible to find alternative accounts in terms of several third variables, which can never be identified and controlled exhaustively. Fever comes along with other physiological symptoms (e.g., weakness of the immune system, inflammation) or behavioral correlates (mood states; risk-taking strategies). In persuasion, too, the number of pro and contra responses to arguments is but one possible mediator; cognitive responses come along with experienced fluency, pragmatic inferences, self-perception and demand effects etc. Because it is never possible to include the entirety of all potential mediator candidates (Z, Z′, Z″, … etc.) in a regression model and to decide which one (or two, or three) is the true mediator, it is impossible to identify causes in a statistical bottom-up inference.

On the other hand, given only three variables, X, Y, and Z, of which one (i.e., X) is bound to be exogenous and is therefore never affected by the other two, there is still a variety of 12 different models that might describe the tri-variate causal structure (cf. Fig. 1).[3] The mediation model X ➔ Z ➔ Y is only one, and often not even the most plausible, of all these causal models. For example, fever, or cognitive responses to persuasion, might be consequences rather than mediating conditions of the disease or attitude change, respectively, reflecting reverse-mediation (X ➔ Y ➔ Z, denoted "reflection" in Fig. 1), which is hard to separate statistically from mediation proper (Lemmer & Gollwitzer, 2017; Thoemmes, 2015). An exhaustive bottom-up analysis aiming to identify the true mediator statistically would have to rely on diagnostic tests of mediation against countless other candidates and many alternative causal models. The number of different models increases dramatically when more than one mediator candidate is considered or when bi-directional or non-directional relations are allowed.

All this is by no means novel. Many methodologically-minded researchers and statisticians would pretend that it is actually common sense, asseverating that all serious scientists understand that there is always room for alternative mediators, and that alternative models exist. However, the reality of current behavioral research as it is published in peer-reviewed journals does not justify this disclaimer. In fact, the twofold problem of alternative mediators and alternative causal models is sorely neglected. As documented below, researchers rarely examine more than a single mediator variable, and they virtually never test other causal models than the standard mediation model. Nevertheless, they routinely and confidently infer from significant mediation tests that the arbitrarily chosen variable Z does mediate an effect, and they infer from non-significant tests that Z does not mediate an effect. We further observe that currently there is more of a tendency among statistical experts to facilitate complex analyses by developing

[2] In fact, no standard mediation test was required to substantiate the mediating role of positive testing.

[3] One might argue that mediation analysis is confined to those (six) models, in which X does affect Y. However, whether this condition is met is hardly known beforehand. None of the 12 models can be ignored theoretically.

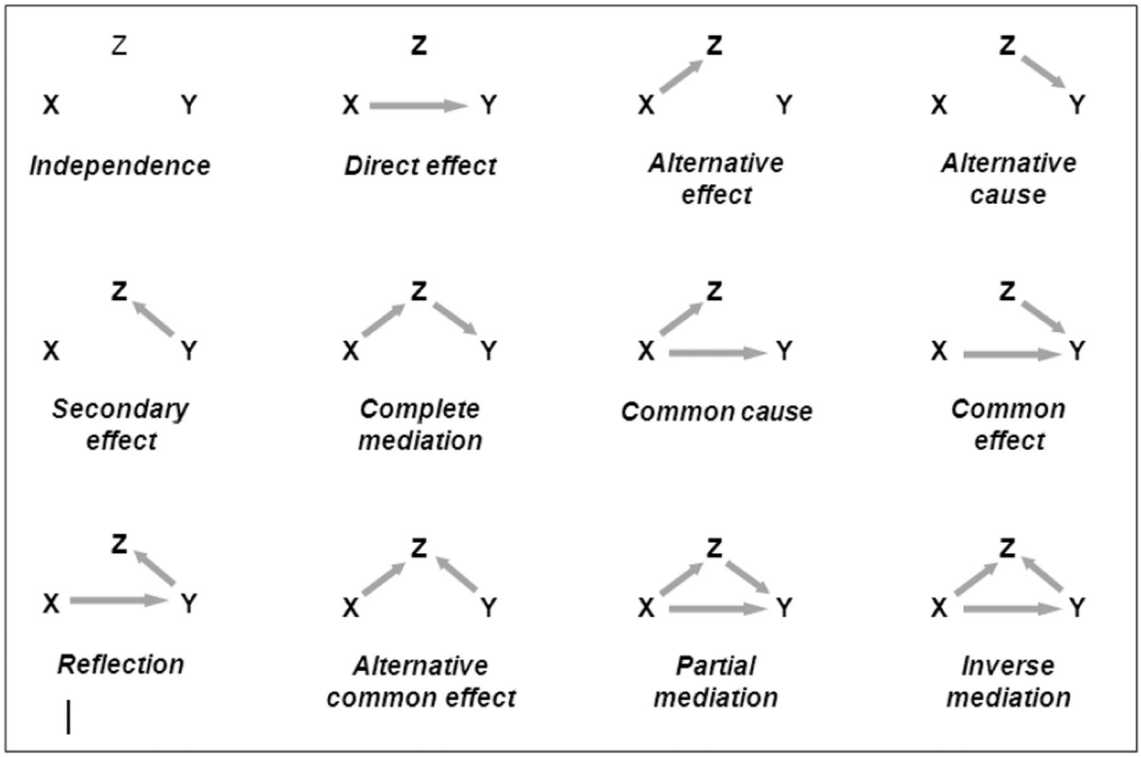

Fig. 1. A variety of different causal models may account for correlations between three variables (X, Y, Z), adapted from Danner, Hagemann, and Fiedler (2015).

statistical tools, than to situate these analyses for research users by emphasizing their theoretical requirements.

1.3. Mediation mimicry

Before we present the results of our inquiry on (inappropriate) theoretical mediation inferences, let us briefly consider a sketch of what we have come to call mediation mimicry (Danner et al., 2015; Fiedler et al., 2011). Monte-Carlo simulation can be used to demonstrate in a simple and straightforward way, what conditional inferences can be drawn or cannot be drawn from statistical mediation tests. If there are good theoretical (a-priori) reasons to believe that Z mediates the influence of X on Y, one can, of course, test the significance and the amount of variance explained by the model X ➔ Z ➔ Y in a specific study context. However, the opposite conditional inference is unwarranted: If a statistical test of a yet to be established mediation model happens to be significant, one cannot infer that Z actually "did" mediate the relation of X and Y (in the present study) or even that Z "does" mediate between X and Y (framed as a general law). As repeatedly noted, neither "*a* mediator" nor "*the* mediator" can be identified or logically inferred from statistical tests of singular variables and models.

In a simulation study by Fiedler et al. (2011), three normally distributed random variables (at a given sample size *n*) were generated in accordance with distinct causal assumptions. For instance, to construct a genuine mediation case, Z was generated to reflect X plus some error variance, and Y was then generated from Z. Thus, the resulting trivariate data set was actually constructed in accordance with an indirect (mediated) path X ➔ Z ➔ Y (with no further constraints imposed on the direct path from X to Y). Likewise, to simulate reverse mediation, Y was generated from X and Z was then generated from Y. In still another condition (not included in Fig. 1), X, Y, and Z were randomly sampled from a pool of homogenously correlated variables (as if all three variables are indices of a single latent construct). The strength of the simulated correlations r_{xy}, r_{xz}, and r_{yz}, and the sample size *n* were varied and all resulting tri-variate data arrays were subjected to a statistical significance test of the assumption that Z mediates the impact of X on Y, regardless of the true causal structure used to generate the data.

As it turned out, it is often only a matter of sufficient sample size and correlation strength to get a mediation test significant, even when the true causal model is different from mediation. From the subset of findings summarized in Fig. 2, it is apparent that the Sobel-test $\mathscr{Z}$-statistic (which provides a convenient summary index) is indeed highest when Z is a genuine mediator.

However, other causal structures also produce significant mediation tests that exceed the dashed horizontal line for a critical Sobel $\mathscr{Z}_{\alpha = 0.05} = 1.96$, especially when samples size *n* increases from 100 (light gray bars) to 200 (dark gray bars). Thus, when Z is generated as a correlate of an unknown true mediator Z′ ($r_{z'z} = 0.70$), the mean Sobel $\mathscr{Z}$ is only slightly reduced. Spurious effects are even stronger when X, Y, and Z represent a causally diffuse cluster of three homogenously correlated indices of the same latent variable. Given reverse mediation (X ➔ Y ➔ Z), mediation mimicry arises at least for a large $n = 200$. Note also that mediation mimicry is generally enhanced in the bottom chart, when the correlation r_{zy} between Z and Y is elevated.

These results were replicated and extended in a more sophisticated simulation approach (Danner et al., 2015) treating X, Y and Z as latent variables (measured by several erroneous indicators) and using structural equation modeling to allow for comparative tests of all 12 models in Fig. 1. While this approach allows one to exclude some of the models as not applicable to the data at hand, it is hardly possible to discriminate between several remaining models with a similar covariance structure (see Lemmer & Gollwitzer, 2017, regarding reverse mediation). In any case, the simulation of mediation mimicry highlights the equivocality of reverse inferences from statistical tests to underlying causal models.

2. The logical status of inferences from mediation tests in the published literature

Despite the logical insight that mediation analysis is conditional on an assumed causal model rather than a diagnostic tool to identify the true causal mediator – and despite the asseveration that this is common

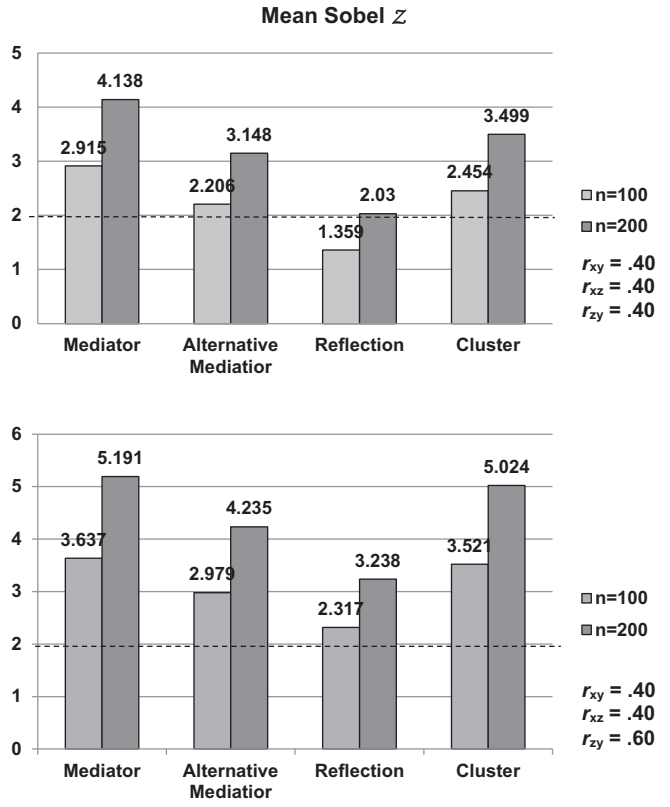

Fig. 2. Mean Sobel $\mathscr{Z}$ statistics obtained in simulation research by Fiedler et al. (2011), when the different causal structures (genuine mediator; alternative mediator; reflection model; cluster of diffusely correlated variables) were used to generate the tri-variate distribution of X, Y, and Z. The dashed horizontal line represents the critical Sobel test statistic $\mathscr{Z}_{\alpha = 0.05} = 1.96$.

sense – the reality of mediation analysis in the published literature testifies to a widespread misunderstanding. The vast majority of mediation tests concentrate on correlates of the dependent and independent variables, which are by no means the only plausible mediator variables. Beyond the observed correlations, there is little reason to assume that the selected variable is playing the role of a causal mediator rather than several other roles a correlated variable can play. The truism that many other variables might exhibit equally strong or even stronger correlations is hardly ever considered. As a consequence, it is common practice to pretend that Z indeed *was* the mediator of the impact of X on Y obtained in a study, or even that Z *is* the mediator in general, simply because a mediation test shows that Z absorbs a significant portion of covariance shared by X and Y. Whether Z is actually the causal mediator, or maybe only a correlate of another mediator Z′ with a clearly distinct meaning, or maybe a confound or correlated index of the dependent variable Y, is mostly a matter of rhetorical framing rather than theoretical cogency or logical necessity.

Let us illustrate this common practice of mediation testing with two examples from the present database. These examples are selected, not because they are particularly distinctive ones, but rather to show how normal it is to publish a mediation analysis without a causal justification. Ge, Brigden, and Häubl (2015) found that participants choosing between pairs of restaurants were significantly more likely to choose the one option they had a chance to unveil before the choice (compared to a condition without the chance to unveil one option). A mediation test was then conducted to demonstrate that the increased choice rate (Y) of one option that was unveiled (X) was mediated by the relative devaluation of the other option (Z). Although both choices (Y) and de-evaluations of rival options (Z) could be plausibly framed as parallel measures of the same choice preference (i.e., the dependent variable Y), the preferred rhetoric was to frame devaluation as a mediator of choices (i.e., as X ➔ Z ➔ Y). Neither the reported statistics nor any theoretical constraints allow us to decide which framing (if any) is correct.

For another telling example, consider Barlow et al.'s (2015) work on intergroup reconciliation. After a series of Australian assaults against Indians, Australian participants were more willing to reconcile with Indians if they were told that an apology of the Australian government was accepted (rather than rejected) by other members of the perpetrator group. A mediation test led to the conclusion that the influence of an accepted apology (X) on willingness to reconcile (Y) was mediated by the resurrection of the perpetrator group's moral image (Z).

Inspection of the verbal measures used in these studies highlights the dependence of this causal interpretation on rhetoric. The items used to assess X (i.e., acceptance of the apology: “The apology made me feel proud.”, “The apology made me feel moral”) tapped on the same optimistic appraisal of the Australian-Indian relationship as the items used to measure Y (i.e., willingness to reconcile: “The apology increases my willingness to express good will toward Indian/Aboriginal people”, “The apology creates a better image of Indian/Aboriginal people in my eyes”, “The apology attests to the good intentions of Australian people”). Similarly, the items used to assess the alleged mediator Z (i.e., moral image: “Other countries will see Australia as a fair people because of the apology”, “The apology restores the moral standing of Australians in the eyes of the international community”, “Other countries will view Australians more positively as a result of reading the apology”) may simply reflect the same optimistic appraisal. Thus, rather than providing cogent evidence for a distinct causal mechanism involving clearly separable theoretical variables, the same data could be paraphrased rhetorically as diffuse cluster of correlations among highly overlapping sets of items reflecting a generally optimistic view. A diffuse cluster of correlations is of course less likely than a mediation model to meet a common standard set by many reviewers and editors.

These two sample studies are in several ways typical of the current reality of mediation testing. Authors are eager to point out that their statistical mediation analyses were based on up-to-date software tools and thousands of bootstrapped samples of the model in question. By contrast, they hardly care about whether their study design allows for an unequivocal assessment of X, Y, and Z; and they largely neglect the problem of alternative mediator variables and alternative causal models. Most importantly, hardly any test of a mediation model relies on a well-established a-priori causal argument that imposes strong theoretical constraints on the mediation mechanism – of the kind illustrated by the role of sexual contact in transmitting a disease or sampling strategies mediating judgment biases (Fiedler & Kutzner, 2015). As a result of this neglect in rigorous theorizing and logic of science, mediation analysis often serves the function of a questionable methodological tool that justifies researchers to draw unwarranted inferences and to ignore the perils of correlational evidence.

To gain a more systematic picture of this state of affairs, we engaged in the following assessment of the reality of causal inferences from mediation analysis in the published literature.

3. Methods

3.1. Literature search

Our assessment was based on a literature search, conducted on June 22nd, 2016, of all articles included in the 2015 database of sciencedirect.com, using the search term “*mediation analysis*” in the search field Keyword and the search term *psychology* in the search field Journal/book title. Sciencedirect.com, operated by Elsevier, is known to cover research published in high-ranking journals. The search field Keyword applies the search term to an entire article, chapter, or abstract, excluding the references, and was thus the best choice for a broad search of mediation analyses. A reference period of an entire year promised to provide us with a useful sample to check on the status of scientific inferences derived from mediation analysis. Although the

Table 1
Frequency counts out of 126 mediation tests (and corresponding percentages) of coded "yes" responses to distinct questions concerning current practices in mediation analysis.

Coding question	Studies in JESP	Studies in JCP	Studies in other journals	Across all journals
Explicit conclusion that the influence of the independent on dependent variable was mediated by the tested mediator candidate?	47	42	37	126 (100%)
Conclusion in past tense	34	28	30	92
Conclusion in present tense	13	14	7	34
Delineation of a-priori theoretical argument explaining the causal function of the mediator?	5	1	12	18 (14%)
Any alternative mediator candidate tested, in addition to a single focal candidate?	6	2	4	12 (10%)
Any alternative causal model tested, in addition to the ubiquitous mediator model?	4	2	5	11 (9%)
Was reverse mediation the only alternative model tested?	2	–	–	2 (2%)

resulting set of 102 relevant articles, including 126 mediation tests, is far from being exhaustive, it does represent a reasonably large reference set to point out an existing problem, the precise prevalence of which is hard to estimate in general.

3.2. Coding criteria

The full set of references and coding results of the 126 mediation analyses reported in the 102 articles can be found under the following link: https://doi.org/10.1016/j.jesp.2017.11.008.

Each mediation analysis was coded for the following aspects:

(1) The first and foremost question pertained to the logical form of the theoretical inference drawn from a mediation test: Do the authors explicitly infer from a significant test that the focal mediator variable actually was the mediator between independent and dependent variable? That is, do the authors make an unwarranted backward inference from a significant mediation test to the causal status of the focal mediator? Whether this criterion is met or not can be decided with virtual objectivity, based on manifest linguistic form, as evident from the sample inferences in the Appendix A (providing the first item per initial letter of author names) and from the exhaustive list of all verbatim conclusions provided in the supplements. The causal meaning of the quoted inferences (e.g., "was fully mediated"; "is mediated"; "had a significant indirect effect"; "impacted indirectly through") is simply a matter of straightforward language comprehension, quite independent of subjective interpretation. The provision of all verbatim inferences makes the coded evidence maximally transparent. We believe that this format is more informative than reporting a kappa coefficient for multiple coding, which is hardly appropriate when estimating the validity of logical inferences (as distinguished from subjective meanings).
The table in the Appendix A (like the exhaustive table in the supplements) also indicates whether the authors' conclusions are presented in past tense ("Z did mediate the influence of X and Y"), restricting the inference to the current study, or in present tense ("Z does mediate"), raising the finding to a generalizable law. Moreover, we also coded whether the inferences are excerpted from the results section or from another article section.

(2) Unwarranted causal inferences from a significant mediator test are most conspicuous when there is no a-priori theory established prior to the reported study, from which the focal mediator's causal status can be derived. We also coded whether such an a-priori theoretical argument was provided, beyond the ad-hoc hypothesis that the mediator was at work. Whereas present causal inferences can be coded in a straightforward manner, coding the absence of a-priori theories may not be fully unequivocal. However, because this aspect is only subsidiary, we avoided the gigantic extra work of having several coders read and code all articles by this theory criterion. As apparent from the next section, though, the paucity of a-priori theorizing is so overwhelming that the imperfect reliability of the data provided by a single coder (the second author) can have hardly obscured the true state of affairs.

(3) Whether or not a mediation test was motivated theoretically, we coded whether or not the mediation analysis included a statistical tests of at least one alternative mediator.

(4) Likewise, with regard to the variety of different causal models, we coded whether or not any alternative model was tested along with the mediation model.

4. Results and discussion

For convenience, the present article is confined to summary statistics that offer clear-cut answers to the questions guiding our inquiry. Interested readers who want to see the results in more detail only have to click on the aforementioned link to get a more complete picture of the study sample and the coding data.

The synopsis of results in Table 1 is organized by coding aspects (rows) and sources (columns). While the first two columns refer to the two most frequent outlets in the reference set, Journal of Experimental Social Psychology (JESP; 47 mediation tests) and Journal of Consumer Psychology (JCP; 42 mediation tests), the third column pools the remaining 37 mediation tests gathered from all other journals. The overall statistics across all articles are provided in the rightmost column. A glance at the table reveals little variation between sources or disciplines (e.g., social vs. consumer psychology); too extreme is the one-sided skew in all coded aspects.

Obviously, drawing the explicit conclusion from a significant mediation test that the tested mediator candidate did actually mediate the influence of the independent on the dependent is virtually the norm. All 126 mediation tests (100%) led to such a conclusion. The vast majority of these conclusions explicitly use the terms "mediation" or "to mediate" in the predicate (cf. Appendix A and exhaustive list in the supplements), clearly asserting what variable is pretended to mediate what basic effect. In a few cases, the same causal inference is implicitly conveyed as a causal chain (e.g., saying that X influences Z, which in turn influences Y).

In 34 cases, the causal inference from the significant result that a particular mediator was actually at work was even expressed as a present-tense statement, suggesting that it reflects a general law that goes beyond the particular finding obtained in the present study. The remaining 92 inferences were expressed in past tense, conveying a more modest conclusion confined to the internal validity of a specific variable mediating a causal influence in the study at hand.

While such inferences from a significant test of selective variables embedded in selective models is generally unwarranted, because it is logically impossible to rule out the entirety of all alternative mediators and models, it is still of interest to examine to what extent mediation

analysts engage in explicit theorizing (beyond purely statistical hypothesis testing). The answer provided by the present frequency count is disillusioning. In no more than 18 out of 126 studies did authors engage in distinct theoretical reasoning or in a literature search for prior evidence in favor of their mediation model. (As this coding criterion involves some subjective judgment of a-priori theorizing, the interested reader is invited to cross-check our coding decisions.)

Unwarranted inferences would reflect less severe violations of scientific reasoning if researchers were not fully disregarding alternative mediators and alternative causal models. As evident from Table 1, at least one alternative mediator candidate was tested in no more than 12 studies. A similar small number of 11 studies did consider at least one alternative model besides the ubiquitous mediation model. Only 2 of these 11 exceptional studies focused on reverse mediation, which we had expected to be considered quite often, because X ➔ Y ➔ Z suggests itself as a plausible theoretical alternative account of X ➔ Z ➔ Y.

5. Discussion

Apparently, then, the practice of mediation analysis is subject to a wide-spread collective mistake. Authors of journal articles – and of course editors and reviewers alike – seem to share the assumption that a causal mediator of an effect is identified when a single mediation test that focuses on an arbitrarily selected variable turns out to be significant, even though other potential mediators and other causal models are simply ignored.

To be sure, the database of around one hundred coded studies is restricted and there may be variation between paradigms and disciplines in the rigorousness and carefulness with which mediation models are tested and interpreted. However, our convenience sample is large and prominent enough to raise the problem of theoretical scrutiny, which is superordinate to statistical scrutiny. And, those who are regularly involved in peer reviewing will probably agree that the problem is not peculiar to the Sciencedirect platform or to the 2015 publication year.

6. Positive recommendations for appropriate mediation analysis

Let us finally turn from the critical appraisal of unwarranted practices into a positive and constructive discussion of how the problem might be overcome and what scientific rules should be established – in the journal review process and in methods trainings – to render mediation analysis a sound and useful instrument. Elucidating the mechanisms and intervening process steps that can explain observed empirical relations is, of course, at the heart of all scientific inquiry. Therefore, the bottom line of our critical note cannot be to refrain from mediation analysis. The crucial question is, rather, what can be done to exploit the potential of valid scientific inferences from the empirical world.

Developing a comprehensive answer to this question is a major goal for future research. For the moment, we would like to suggest the following tentative set of maxims that should be rather easy to implement, monitor, and control.

6.1. Keep in mind what a statistical mediation test can (not) do

Logically, what a statistical mediation test can do is test that, IF a causal model is assumed, THEN a prediction derived from that causal model can account for a substantial part of the variance in a certain study context. As a principle, *statistical tests are conditional on the validity of the model being tested*; no statistical test can ever identify the true causal model from the entire set of all logically possible models. So, empirical scientists should simply refrain from such conclusions as "Z was shown to mediate the influence of X on Y". Even an interpretation like "our data are consistent with the model X ➔ Z ➔ Y" is misleading and biased because it distracts from the truism that many other models and alternative mediator might also be consistent with the data. Good science relies heavily on precise language and logically sound inference making.

6.2. Ideally, mediation tests should be based on well-established theories or empirical laws

Granting that mediation analysis is conditional on the causal models being tested, the burden is on good theorizing or, to put it in Bayesian terms, theoretical priors (Fiedler, 2017). Mediation analysis is on safe ground if the causal mechanism is well understood. The epidemiologist may know, or may have strong data to assume, that contagion is mediated by sexual contact just as the social psychologist can reasonably assume that encoding strategies are reasonable mediators of persuasion processes. Based on such solid theoretical knowledge, the researcher may then rely on statistical mediation analysis to estimate the amount of variance explained by sexual contact, to make inferences about the need to postulate other mediators etc. Thus, in the ideal case, well-established theories and laws that imply a mediation process (e.g., conditioning accounts of attitude learning; sampling accounts of overconfidence) can be tested in a straightforward way.

6.3. Beware of causal-temporal constraints on mediation models

When no well-established causal model exists on a-priori grounds, the creation of novel mediation hypotheses must be subject to distinct logical and psychological constraints. According to the Hyman-Tate criterion (Tate, 2015), the causal ordering of a mediation hypothesis (X ➔ Z ➔ Y) implies a conceptual time ordering from predictor (X) to mediator (Z) to criterion (Y). That is, the mediator must refer to a causal condition that emerges prior to the criterion outcome but not prior to the predictor. By this criterion, for example, an enduring personality trait acquired long before the causal predictor is temporally too remote to qualify as a reasonable mediator; it is more likely to represent another predictor or moderator. Conversely, an arbitrary temporal ordering imposed by the experimental procedure on the assessment of two self-report measures of simultaneously existing psychological states, Y and Z, is constrained by the procedure and therefore precludes statistical inferences about the causal ordering. Although there may be other constraints – for instance, spatial constraints in neuro-anatomy – the causal-temporal constraints of the Hyman-Tate criterion afford a highly useful heuristic to planning logically sound mediation analyses.

6.4. Open-minded theorizing allows for more than one mediator

Every non-trivial empirical relation between X and Y can be explained by more than one mediator. It is therefore essential to test or at least to consider alternative mediation hypotheses in an open-minded fashion.

6.5. Beware of alternative causal models

By the same token, one must take the possibility into account that another causal model than a mediation model may explain the relation between three variables X, Y, Z. Even when statistical testing cannot discriminate between (X ➔ Z ➔ Y), reverse mediation (X ➔ Y ➔ Z), and several other causal model in Fig. 1 (Danner et al., 2015; Lemmer & Gollwitzer, 2017; Thoemmes, 2015), this should not prevent one from explicit theorizing about alternative causal models and from specifying their testable constraints. Non-statistical constraints (like the Hyman-Tate criterion) are then required to identify the most reasonable causal chains (Spencer, Zanna, & Fong, 2005) from the variety of all possible causal models. This is not meant to ban statistical methods from theorizing. For instance, structural-equation modeling may be used to find out what subsets of possible causal models are most compatible with the given correlation data (Danner et al., 2015). Importantly, though,

one should not draw unwarranted reverse inferences from the significance of a statistical test. Just as there is no rationale to infer the truth of any H_1 from a significant result (Trafimow, 2003), or the truth of H_0 from an insignificant test, mediation analyses can never imply that "Z mediated the influence of X on Y", or else that "X was a common cause of both Y and Z".

6.6. Propositional form of logically permissible inferences

Granting the basic insight that a statistical test can never establish a causal model, on which it is conditionalized, it should be clear that neither direct inferences (about Z mediating the relation of X and Y) nor indirect inferences that imply a mediating path (X affects Z, which in turn affects Y) are permissible. The ultimate question then arises as to how the results of a mediation analysis should be formulated positively.

Here is a sample of appropriate phrases. One might write "Conditional on the model assumption X ➔ Z ➔ Y, our statistical test shows that Z can account for a significant portion of variance"; it might be fair to add that other models cannot be excluded. Or, it would be appropriate to write "the pairwise correlations between X and Z and between Z and Y are strong enough to [partially] account for the [full] relationship between X and Y, consistent with a mediation model, but also with several other models." Still another way to report the finding would be to state "if Z were to be included in a regression model, it would absorb a significant part of the variance shared between X and Y. This is consistent with a mediation model but not exclusively. It is also compatible with several other models, which can only be distinguished through sound theorizing and clever experimentation." Or, a minimal viable formulation would be to say that "the obtained significant results can be predicted if the assumption of a mediation model is correct." Technically, it may generally appear justified to report that "our test of a mediation model was significant", but this apparently justified summary statement does not mention that many other models might have been tested and provided significant results as well.

7. How can such recommended rules be implemented?

How can these recommended rules be implemented? What can be done to effectively improve on the current status of mediation analysis? The primary answer that suggests itself points to the journal review process. It would be so easy to instruct and sensitize expert reviewers to the perils and misuses of mediation testing, which have been criticized consensually in various recent articles (Giner-Sorolla, 2016; Kline, 2015; Tate, 2015; Thoemmes, 2015). Without extra training or formal instruments, reviewed studies can be judged by the same criteria as in the present survey of published studies. The Hyman-Tate criterion affords a very useful and easily applicable heuristic to evaluate mediation models; alternative causal models and alternative mediators should be considered anyway, and a theoretically compelling mediator model is easy to distinguish from an arbitrary, merely rhetorical mediator hypothesis. Before too long, it should be possible to impose more rigorous theoretical and logical constraints on published articles.

At a stage earlier than article writing, intriguing advanced seminars, workshops, and graduate curricula could be enriched with appropriate training programs on mediation analysis and related problems in logic of science. The training goal would be to educate young scientists to specify the mechanisms and the functional constraints of their theoretical approaches, and to discriminate clearly spelled out (mediation) theories from unconstrained speculation.

Last but not least, we need to reach a minimal degree of consensus about the problem of mediation analysis and its potential to undermine the scientific value of published psychological research. Establishing such a minimal consensus is the major aim of the present paper.

8. Concluding remark

Recent suggestions about how to improve the quality of behavioral science have emphasized the need to comply with good practices and statistical standards. The present article focuses on deficits in theoretical reasoning and logic of science. The wide-spread habit to draw unwarranted inferences from mediation tests is but one prominent example of a broader class of theoretical weaknesses. Other examples that deserve to be monitored more critically include the development of theoretically appropriate manipulation checks (as distinguished from superficial instruction memory checks), the failure to distinguish universal laws and existence proofs, or violations of the logic of conditional reasoning. A more comprehensive analysis of the literature may reveal that deficits in theoretical reasoning maybe constitute the most severe obstacle on the way to good science – more severe than subordinate issues of research design and statistics.

Appendix A

Sample of inferences from mediation analyses (first item per initial letter of author names) drawn from the full list provided in the Supplemental materials.

Authors	Causal inference	Present tense	Place in paper	Verbatim quotation of the corresponding interpretation
Achtziger et al.	Yes	No	Abstract	Detailed analyses revealed that the link between self-control and debts *was fully mediated* by compulsive buying.
Bailey et al.	Yes	No	Abstract	national differences in knowledge of fraction concepts *were fully mediated* by differences in knowledge of fraction procedures
Carlston et al.	Yes	Yes	Highlight	*Memory partially mediates* assimilation effects and may fully mediate contrast effects
Dimofte et al.	Yes	Yes	Theory	Study 3 shows that the loss of collective *self-esteem mediates the impact of aspirational ads* on product attitudes.
Egan et al.	Yes	No	Discuss	we found that mood-induced changes in working memory performance *were driven by changes* in perceived mental depletion
Fennis et al.	Yes	No	Results	The analysis confirmed that need for order *acted as a significant mediator* of the relationship between perceiving a disordered environment and motivation in goal pursuit
Galindo et al.	Yes	No	Results	Results from the combined mediation analysis indicated that both math proficiency and indicators of the home learning environment in kindergarten *partially mediated* the relation between SES and math achievement when both sets of variables were included in the model

Hagtvedt et al.	Yes	Yes	Abstract	Study results indicate that under low arousal, questions have a more favorable influence on product evaluation than statements do; this influence *is mediated by* the perceived interestingness of the phrase. Under high arousal, the influence is reversed, and it is mediated by perceived clarity.
Kahn et al.	Yes	No	Abstract	The relationships between phenotypic racial stereotypicality condition on organizational attractiveness and diversity perceptions *were mediated by* identity-related trust.
Levontin et al.	Yes	No	Results	Mediation analyses (Hayes, model 4) using bootstrapping (e.g., Preacher, Rucker, & Hayes, 2007) with 5000 replications confirmed that perceived resource abundance *mediates the effect* of emptying (vs. filling) on participants' allocation of their year-end bonus (95% CI = [− 36.02, − 0.68]), providing unambiguous support for our hypothesis.
Malhotra et al.	Yes	No	Discuss	Mediation analysis revealed that a higher propensity for conscious motor processing positively *influenced performance early in practice by specifically reducing variability* of impact velocity and putter face angle at impact.
Ng et al.	Yes		Results	Thus, the results provide *evidence of the mediational role of regret*.
Pereira et al.	Yes	No	Results	This indicates that the effect of expectancy violation on support for collective punishment *is entirely mediated* by the second serial mediator: group value.
Rakoczy et al.	Yes	No	Results	The results revealed that children's assessment of advisor expertise *was indeed a significant mediator*.
Salim et al.	Yes	No	Abstract	Two coping strategies *were found to mediate* this relationship: emotional support and positive reframing.
Tiefenbach et al.	Yes	Yes	Results	In summary, our results show that (i) 3–11 had a substantial direct negative effect on SWB in Japan, (ii) this negative effect *is mediated by* the positive effect on donations by about 31% (1 (0.189/0.274)), which (iii), still leads to an overall negative impact of 0.189 points experienced after 3–11.
van Bree et al.	Yes	Yes	Abstract	Path analyses showed that habit *significantly mediates the relationship* between prior and later PA, after ASE/TPB variables were taken into account.
Van de Vyver et al.	Yes	No	Results	Sequential mediation analyses showed that positive appraisals and then elevation significantly and *sequentially mediated the effect* of the elevation-inducing video on donations
Wester et al.	Yes	No	Abstract	Mediation analysis showed that *women felt greater discomfort because of higher levels of pathogen disgust sensitivity*.
Xu et al.	Yes	No	Results	These results confirmed that the perceived *importance of money mediated the pain-buffering effect* of social support.
Zaleskiewicz et al.	Yes	No	Results	indicating that amount of money sent to the Receiver *was a significant moderated mediator* of the relationship between mortality salience and Proposer satisfaction

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2017.11.008.

References

Barlow, F. K., Thai, M., Wohl, M. A., White, S., Wright, M., & Hornsey, M. J. (2015). Perpetrator groups can enhance their moral self-image by accepting their own intergroup apologies. *Journal of Experimental Social Psychology, 60*, 39–50.

Bullock, J., Green, D., & Ha, S. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology, 98*, 550–558.

Danner, D., Hagemann, D., & Fiedler, K. (2015). Mediation analysis with structural equation models: Combining theory, design, and statistics. *European Journal of Social Psychology, 45*(4), 460–481.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*, 46–61.

Fiedler, K., Freytag, P., & Unkelbach, C. (2007). Pseudocontingencies in a simulated classroom. *Journal of Personality and Social Psychology, 92*(4), 665–677. http://dx.doi.org/10.1037/0022-3514.92.4.665.

Fiedler, K., & Kutzner, F. (2015). Information sampling and reasoning biases. *The Wiley Blackwell handbook of judgment and decision making* (pp. 380–403). .

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

Fiedler, K., Walther, E., Freytag, P., & Stryczek, E. (2002). Playing mating games in foreign cultures: A conceptual framework and an experimental paradigm for inductive trivariate inference. *Journal of Experimental Social Psychology, 38*(1), 14–30. http://dx.doi.org/10.1006/jesp.2001.1483.

Ge, X., Brigden, N., & Häubl, G. (2015). The preference-signaling effect of search. *Journal of Consumer Psychology, 25*(2), 245–256.

Giner-Sorolla, R. (2016). Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology, 65*1–656. http://dx.doi.org/10.1016/j.jesp.2016.01.010.

Hayes, A. F. (2009). Beyond baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420.

Kline, R. B. (2015). The mediation myth. *Basic and Applied Social Psychology, 37*(4), 202–213. http://dx.doi.org/10.1080/01973533.2015.1049349.

Lemmer, G., & Gollwitzer, M. (2017). The 'true' indirect effect won't (always) stand up: When and why reverse mediation testing fails. *Journal of Experimental Social Psychology, 69*, 144–149. http://dx.doi.org/10.1016/j.jesp.2016.05.002.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* New York, NY: Erlbaum.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology,* 58593–58614. http://dx.doi.org/10.1146/annurev.psych.58.110405.085542.

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. http://dx.doi.org/10.1037/0022-3514.89.6.845.

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology, 37*(4), 235–246. http://dx.doi.org/10.1080/01973533.2015.1062380.

Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology, 37*(4), 226–234. http://dx.doi.org/10.1080/01973533.2015.1049351.

Tormala, Z., Falces, C., Briñol, P., & Petty, R. (2007). Ease of retrieval effects in social judgment: The role of unrequested cognitions. *Journal of Personality and Social Psychology, 93*(2), 143–157.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110*(3), 526–535. http://dx.doi.org/10.1037/0033-295X.110.3.526.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science, 15*(6), 307–311.