It Takes Four to Tango:

Why a Variable Cannot Be a Mediator and a Moderator at the Same Time

Johann Jacoby and Kai Sassenberg
Knowledge Media Research Center Tübingen

Correspondence address:

Johann Jacoby
Knowledge Media Research Center
Konrad-Adenauer-Strasse 40
72072 Tübingen
Federal Republic of Germany
email: j.jacoby@iwm-kmrc.de
phone: +49 7071 979-246
fax: +49 7071 979-200
Abstract

Statistical mediation and moderation have most prominently been distinguished by Baron and Kenny (1986). More complex models that combine both of these effects have recently received increased attention, namely mediated moderation and moderated mediation (e.g., Muller, Judd, & Yzerbyt, 2005). Presently the focus is on a three variable model that is often claimed to represent an instance of moderated mediation or mediated moderation. More specifically, in this model a single variable is considered to simultaneously mediate and moderate the same effect. We show that this specific model however cannot be considered either an instance of mediated moderation nor of moderated mediation. Also, we argue that this particular model is a priori misspecified. A data pattern that seems to agree with this model is recognized as plausible, but it indicates that the model must be modified in one of two ways to be methodologically sound. We conclude by recommending to not use this three variable model and to consider evidence that seemingly agrees with it as evidence that the three variable model is inadequate.
In 1986, Baron and Kenny published a widely cited and well-known article that popularized the concise distinction between two roles that a third variable $M$ can have concerning an effect of an independent variable $X$ on a dependent variable $Y$, henceforth referred to as the basic effect $X \rightarrow Y$. According to their widely shared distinction $M$ can either explain the basic effect, and thus have the role of a mediator, or it can have the role of a moderator. In this latter case the basic effect is different depending on the value of $M$. Distinguishing these two roles is important for psychological theory, design choices, and statistical analysis. However, Baron and Kenny (1986) have also clarified that mediation and moderation can be present simultaneously within an empirical constellation. In this vein, it is possible that a moderator proves crucial in distinguishing different degrees of mediation (moderated mediation) or that a mediator explains the effect that an independent variable has in interaction with another (mediated moderation).

Mediated moderation and moderated mediation typically feature four variables: two variables constituting the basic effect ($X$ and $Y$), one mediator $M_e$ and one moderator $M_o$. However, one special case of moderated mediation has been discussed, among others, by Preacher, Rucker, and Hayes (2007). In this special case only three variables are involved in a moderated mediation: A variable $M$ appears to simultaneously act as a mediator and a moderator of the basic effect $X \rightarrow Y$. While this particular case of moderated mediation seems to be intuitive and plausible, and data might appear to be consistent with the model, we aim to demonstrate in what follows that it contains serious logical contradictions. These contradictions result in the conclusion that the three variable moderated mediation model is fundamentally misspecified and a priori inadequate. We therefore suggest that the three variable moderated mediation model should routinely be modified so that it contains at least four variables and provide a general recommendation on how this modification can be accomplished.
Simple third variable effects: moderation and mediation

Relative to a basic effect \( X \rightarrow Y \), a third variable \( M \) can take two distinct basic roles. First, it can be a mediator to the extent that it explains the basic effect. Figure 1 visualizes a simple mediation model and introduces the effect or path names for components in this model. The effect of \( X \) on \( M \) is referred to as path \( a \), the effect of \( M \) on \( Y \) is \( b \), the total effect of \( X \) on \( Y \) is \( c \), and the direct effect of \( X \) on \( Y \) is \( c' \): the residual effect of \( X \) on \( Y \) after \( M \) has been partialed out of \( Y \). Further, in this general model, the indirect or mediation effect, is \( ab \): the product of \( a \) and \( b \). The term \( ab \) is equivalent to the difference between the total and the direct effect \( c - c' \) (Hayes, 2009). If this indirect effect is substantially different from zero, then \( M \) can be said to explain the effect \( X \rightarrow Y \) in as far as it "...transmits the effect of an antecedent variable on to a dependent variable, thereby providing more detailed understanding of relations among variables." (MacKinnon & Fairchild, 2009, p. 16)

Secondly, \( M \) can be a moderator, in that the basic effect \( X \rightarrow Y \) is different (in magnitude, direction, or both) depending on the value of \( M \). In Figure 2, the degree of moderation of \( X \rightarrow Y \) by \( M \) is indicated by the interaction effect \( d \). This results in different simple main effects in ANOVA type approaches or different simple slopes in the logic of regression analytic techniques (Aiken & West, 1991).

We will refer to these two constellations of mediation by \( M \) vs. moderation by \( M \) in a model containing the three variables \( X, Y, \) and \( M \) as simple mediation and simple moderation, respectively.

Simultaneous third variable effects

In more sophisticated contexts, mediation and moderation can be considered simultaneously within one constellation of variables. This is the case if either a moderating
A mediator cannot also be a moderator

5

variable allows to distinguish between different degrees of mediation by a mediator (i.e., different indirect effects $ab$ via the mediator depending on the value of the moderator: moderated mediation), or a moderated basic effect is explained by a mediating variable (mediated moderation). These two cases focus on different conceptual nuances, but they are identical with regard to the underlying mathematical model (Muller, Judd, and Yzerbyt, 2005; see also Edwards & Lambert, 2007; and Preacher, Rucker, & Hayes, 2007). We therefore refer to both types of constellations in which moderation and mediation occur as second order constellations, or SOCs for short.

The special case: The three variable second order constellation

One particular second order constellation is the one among three variables mentioned earlier: the three variable second order constellation, or SOC3. In SOC3, a variable $M$ is hypothesized to both mediate the basic effect $X \rightarrow Y$ (i.e., there is an indirect effect $ab$) and simultaneously moderate the basic effect $X \rightarrow Y$ (i.e., there is an $X \times M$ interaction effect $d$; see Figure 3A). There is thus one single variable that appears be both a mediator and a moderator with regard to the basic effect $X \rightarrow Y$ ('$M$ mediates and $M$ moderates' perspective on SOC3, henceforth referred to as Perspective I). Like other SOC models, this model can also be viewed from a different conceptual perspective: $M$ mediates the basic effect, and $X$ moderates the path $b$ ('$X$ moderates' perspective on SOC3, henceforth referred to as Perspective II; see Figure 3B).

Also, as in other SOC models, while these two perspectives may suggest different psychological phenomena and processes, they are empirically and mathematically indistinguishable. Both perspectives lead to the same set of equations that parsimoniously describe the model:

(1) \[ M = k_1 + aX + e_1 \]

(2) \[ Y = k_2 + bM + c'X + dXM + e_2 \]
where \( k_{1,2} \) are intercepts and \( e_{1,2} \) are random errors. \( X \), \( M \), and \( Y \) are the variables as defined above, and \( a \), \( b \), \( c' \), and \( d \) are the effects mentioned earlier. Because of this equivalence of the two perspectives (Perspectives I and II) any observation about the statistical and empirical characteristics of SOC3 made from one of the perspectives will be true also from the other perspective.

SOC3 appears to be an exception among SOC models because all other second order constellations that are discussed in the literature contain at least four variables, namely a moderator and a mediator as separate variables (Edwards & Lambert, 2007; Iacobucci, 2008; James and Brett, 1984; MacKinnon, 2008; Muller et al., 2005; Preacher et al., 2007). In these four variable models, the variable \( X \) does not act as a moderator and \( M \) does not act as both a mediator and a moderator at the same time. Instead, if \( M \) is the mediator, a fourth variable takes on the role of the moderator, and if \( M \) is the moderator, a fourth variable is the mediator. More precisely, in these common four variable SOC models the indirect effect via \( M \) is moderated, and the interaction of \( X \) and \( M \) is mediated.

The treatments of second order constellations that can be found in the current literature markedly differ, however, in whether SOC3 is included and elaborated on or not. Whereas Muller et al. (2005) or Edwards and Lambert (2007) discuss exclusively second order constellations involving a minimum of four variables, Preacher et al. (2007) or James and Brett (1984) discuss the special case of SOC3. We argue here that, rather than constituting incompleteness on the part of the former authors, omission of SOC3 is sensible and careful. In the following discussion, we present three arguments for the claim that SOC3 is problematic and essentially a logically contradictory model. These arguments are very much related and thus not independent, but they offer different approaches to understand why the SOC3 model must be \textit{a priori} false.
Despite these arguments, it is of course possible that empirical data yield a pattern that seems consistent with SOC3 (and thereby with Equations (1) and (2)). Thus, the three variable second order constellation may seem, on the surface of it, to underlie the data. Our discussion will show however, that especially if the data is consistent with SOC3, this will indicate that the three variable model is wrong, and that a fourth variable in the model must be conceptualized and modeled (i.e., measured or manipulated) if one wishes to understand the actual processes and phenomena among the variables. Afterward we will direct attention to how empirical data that seemingly agree with SOC3 can be understood and finally describe how the extension of the model by a fourth variable could proceed.

The false moderated mediation

We will first outline a concrete but fictitious intervention example in the argument why SOC3 is false. We refer back to it in subsequent sections.

Concrete Example

Suppose that a coaching program for high school students (treatment, \( X \), with values 1 = 'participates in the program' and -1 = 'does not participate in the program/ appropriate control group') successfully teaches self-management skills (\( M \): amount of self-management skills). The self-management skills available to children (as measured at some point in time after or during the intervention) lead to better academic performance (\( Y \): grade point average one year after the program took place). So far, this example depicts a simple mediation: \( X \) causes \( M \) (path \( a \)), \( M \) causes \( Y \) (path \( b \)) and, as a result, \( X \) predicts \( Y \) (path \( c \)), with this latter relationship being substantially reduced once \( M \) is taken into account (i.e., used as a covariate in the prediction of \( Y \) by \( X \), resulting in path \( c' \)). Under common circumstances, conceptually, \((c - c') = ab\) (Baron & Kenny, 1986; MacKinnon, 2008). This model of simple mediation is visualized in Figure 1.
However, in addition to the mentioned relationships among the variables, for those students who participated in the program the relationship between self-management skills and academic performance is positive, but for students who have not participated in the program, there is no relationship (see Figure 4 for a prototypical data pattern). This constitutes a treatment $\times$ self-management skills interaction effect on $Y$. It thus appears that the variable self-management skills mediates the basic effect of the program on academic performance and at the same time is a moderator of this treatment effect. Therefore, this scenario seems to be a case of moderated mediation. However, we will argue first that SOC3 cannot be a case of moderated mediation or mediated moderation from a semantic point of view and then discuss two arguments showing that SOC3 must be a conceptually wrong model.

SOC3 cannot be moderated mediation or mediated moderation

The terms 'moderated mediation' or 'mediated moderation' imply that the indirect effect, and thus the mediation in a model is different for different values of a moderator, or that an interaction effect (i.e., a moderation) is transmitted and thereby explained by a mediator, respectively. SOC3, as we will show next, cannot be considered to be either moderated mediation or mediated moderation in the sense of these terms.

SOC3 is not moderated mediation

Perspective I of SOC3 holds that $M$ mediates the effect $X \rightarrow Y$. An inspection of Figure 3A however already reveals that by its interaction with $X$, $M$ does not moderate the indirect effect in the model, but the direct effect of $X$ on $Y$. Thus, different from what the concept of moderated mediation implies, it is not the mediation that is moderated by $M$. Hence, from Perspective I, SOC3 cannot be a case of moderated mediation.
From Perspective II (see Figure 3B), it may appear at first glance as if $X$ indeed moderates its own indirect effect on $Y$ and therefore that the mediation path is moderated by $X$ because it moderates one of the mediation component paths: $b$. But closer inspection reveals that the fact that $X$ moderates $b$ does not indicate that the indirect effect $ab$ is also moderated by $X$, because $a = 0$ at each level of the moderator $X$. The component path $a$ is always zero, because $X$ is trivially a constant among all observations that are homogeneous with respect to $X$. If $X$ is constant, there can be no effect $a$. Thus, at each and any value of $X$, the path $a$ is zero. If $a$ is zero, then the product $ab$, the indirect effect, is also zero. It follows from this that the indirect effect of $X$ on $Y$ via $M$ at any value of $X$ must be zero: there is no mediation within any value of $X$ (which in this case is considered the moderator), and thus there is no moderation of this zero mediation effect by $X$. Hence, $X$ cannot be considered as the independent variable in a mediation model and also as moderating the component path $b$ at the same time. The conclusion of this argument is that from Perspective II, SOC3 can also not be a case of moderated mediation.

Since Perspectives I and II are the only potential ways to interpret SOC3 as a model containing mediation and moderation, SOC3 thus cannot be a case of moderated mediation. SOC3 is not mediated moderation

But could SOC3 be a case of mediated moderation? Inspection of Figure 3A and Figure 3B reveal that it is not, since regardless of which perspective on SOC3 one takes, there is no interaction that is explained by an indirect effect path via a mediator. As Muller et al. (2005) have clarified, mediated moderation is present when the interaction of the independent variable with a moderator predicts a mediator which in turn predicts the dependent variable. Thus, mediated moderation requires an interaction at the origin of the mediation chain. This interaction then unfolds its effect on the dependent variable via a mediator. In SOC3 however, the interaction either occurs by the mediator on the direct effect $X \rightarrow Y$ (Perspective I) or in the second
component path of the indirect effect (Perspective II). Thus, from both perspectives, the interaction and thereby the moderation does not occur at the origin of the mediation chain. Therefore, the moderation cannot be mediated by \( M \). Hence, SOC3 also cannot represent a case of mediated moderation.

In sum, from a purely conceptual point of view, SOC3 represents neither a mediation that is moderated, nor a moderation that is mediated. The two phenomena mediation and moderation do not occur on each other. But even if one disregards the issue of how to name SOC3, this model is conceptually contradictory as we show next using two different approaches.

SOC3 is conceptually contradictory

A different variance that is irrelevant for mediation

The following argument focuses on effects as shared variances and considers various variance portions among the three variables \( X, M, \) and \( Y \) in SOC3 and how they do or do not overlap. An effect as well as a relationship between two variables can be understood as an overlap of the variances of the two variables. For example, the total effect \( c \) of \( X \) on \( Y \) in a simple mediation model indicates that the variances of \( X \) and \( Y \) partly overlap, or that \( X \) and \( Y \) share variance. In other words: If \( X \) shares variance with \( Y \), \( X \) has an effect on \( Y \). In this sense, explaining a variable by another variable means to identify a portion of variance that these variables share. By extension, explaining an effect by a third variable means to identify variance that the third variable shares with the effect of \( X \) on \( Y \). The aim of a test of (simple) mediation by \( M \) is hence to identify variance that is shared by all three variables in the model simultaneously.

We will now explain why the variance that is meaningfully explained by the \( X \times M \) interaction term in SOC3 is not shared by at least one of the three variables involved in the model, and therefore irrelevant with regard to the question of whether \( M \) is a mediator.
The interaction effect in SOC3 indicates that there are systematic relationships between $X$ and $Y$ at specific values of $M$ (from Perspective I as introduced earlier), or systematic relationships between $M$ and $Y$ at specific values of $X$ (from Perspective II). In the terminology of multiple linear regression and ANOVA, these relationships at specific values of another variable, are referred to as simple slopes or conditional effects. From Perspective I however, these conditional effects imply a relationship between $X$ and residuals in $Y$ (after $X$ and $M$, as single predictors, have been partialed out from $Y$) at specific values of $M$. Since the basic effect $X\rightarrow Y$ to be explained (i.e., being mediated via $M$) involves, by definition, not the residuals in $Y$, but that part of $Y$ that is systematic with regard to $X$, the basic effect must be independent of the conditional effects which involve the residuals in $Y$. Hence, conditional effects cannot be relevant to explain the effect of $X$ on $Y$ or, in fact, any effect of $X$.

From the point of view of Perspective II on SOC3, a similar argument applies to the independence of the systematic effect of the single predictor $M$ on $Y$ and the variance in $M$ that relates to $Y$ conditional on $X$ (i.e., effects of $M$ on $Y$ conditional on $X$). Again, the variance that is shared between $M$ and $Y$ as parts of the unique interaction effect, or the conditional effects, involves residuals in $Y$ (after $X$ has been controlled for statistically). Therefore the conditional effects are independent of $X$ and thereby cannot explain any effect of $X$.

Note that we do not claim that the $X\times M$ interaction is not relevant for explaining the variable $Y$ – it most certainly is. However this interaction is irrelevant to explaining the effect $X\rightarrow Y$.

Consider this argument within the example given above: Participating in the coaching program leads to higher self-management skills than not participating ($X \rightarrow M$). Additionally, among participants in the program, self-management skills are positively related to academic performance ($Y$), and among those who are not in the program, skills are not at all related to
performance. These two latter relationships are present within the treatment group and the control group and thus variance in performance overlaps with variance in skills. The latter variance in skills does however not overlap with variance in the treatment (X). Within the treatment group and within the control group, respectively, there are no differences in treatment, because all observations within the treatment group have \( X = 1 \), and all participants in the control group have \( X = -1 \). Thus, the variance that is shared between self-management skills and academic performance in the form of conditional effects (i.e., the interaction effect) are differences that must originate from other factors than the treatment. Such other factors may comprise interindividual differences (e.g., prior academic history, general intelligence), contextual differences (e.g., parenting, number of siblings), and others. Thus, due to their independence of \( X \), the conditional effects, and thereby the interaction effect, are irrelevant for the question of whether \( M \) is a mediator of \( X \rightarrow Y \) as they reflect influences outside the model.

In sum, an \( X \times M \) interaction effect in the SOC3 model indicates that there are effects relevant to the processes occurring among and around the three variables that however cannot be properly modeled and much less contribute to better understanding of the effects regarding the three variables. The interaction effect does not explain these additional effects, it merely alerts to their existence. It also does not contribute to elucidating the processes within the SOC3 model: In particular, the conditional effects constituting the interaction effect are independent of \( X \) and therefore irrelevant for any simple mediation of \( X \rightarrow Y \) via \( M \).

The trouble with 'main' effects

A second approach to the conceptual content of the model, which, of course, is not independent of the preceding argument, but highlights a different aspect, underscores the difficulties surrounding the \( X \times M \) interaction term in the SOC3 model. It illustrates that SOC3
does not help in clarifying a sophisticated pattern of associations in data, or constitute $M$ being a mediator.

Suppose that, as in the models in Figure 3 and the example data pattern in Figure 4, the analysis of data reveals that indeed $X$ is associated with $M$, $M$ shows a relationship with $Y$ (after partialing out $X$ as prescribed in the original Baron & Kenny, 1986, approach), and $X$ and $M$ interact in predicting $Y$. It is well known from common multiple linear regression as well as ANOVA-type analyses that the effect of one predictor alone cannot be interpreted as a main effect if the estimated model contains a higher order term in which the predictor is involved (i.e., an interaction effect, Aiken & West, 1991; Cohen, Cohen, West, & Aiken, 2003). We can see this problem clearly for the example of the self-management program for students above.

Consider the case from Perspective I. The interaction effect shows that students who are relatively high in self-management skills profit from the program while their equally highly skilled peers' performance in the control group is virtually unchanged by program participation (Figure 4A). The effect on academic performance of the treatment alone, $c'$ or $c$, is thus a conditional effect, namely the program effect for students with values of the self-management skill equal to zero. For example, if the measurements of self-management skills have been properly centered before analysis (see Aiken & West, 1991; Cohen et al., 2003), the effect of the treatment alone on performance will represent how much students who have average self-management skills profit from the program. This program effect will however be different for those with higher or lower skills at the time of measurement. In cases other than the example, referred to as disordinal interaction, the program effect may even have opposite signs for different levels of skills. If there is a possibility of very much lower skills than those in the example and visualized in Figure 4A, it is easy to see that the treatment effect on performance may actually reverse: participants with exceptionally low skills assigned to the treatment group
may then perform worse than their peers with exceptionally low skills assigned to the control group. It is therefore not possible to unambiguously assess the mere treatment effect on performance if the mediator skills interacts with program participation in predicting performance.

From Perspective II, again, a similar problem emerges. If, as the interaction effect shows, students in the treatment group do not only show on average higher performance than students in the control group, but self-management skills are also positively related to academic performance for these students, while there is no skill-performance relationship for the students in the control group, then the prediction of academic performance by self-management skills alone (path b) is a conditional effect. In the present case, b is thus the effect of self-management skills on academic performance in the treatment group coded with zero. If the treatment variable (X) is centered (i.e., coded [1; -1], as Aiken & West, 1991, recommend) then this conditional effect does not even apply to any real group of participants, as no condition of the study is characterized by X = 0.

The ambiguity of single predictor effects in the presence of statistical interaction commonly leads to a general imperative to refrain from interpreting single predictor regression weights or statistical effects associated with a single factor alone in the presence of an interaction. Such single predictor effects are not main effects, but conditional (or, in the case of centered predictors: average) effects (Cohen et al., 2003). If however an interaction effect precludes a single predictor effect to be interpreted as a main effect, then it should also preclude their interpretation as main effect-like component path within an indirect effect (i.e., path b as a component of the indirect effect ab). In order to assess simple mediation however, an unambiguous main effect of M on Y (i.e., b) is necessary, in line with the unambiguous main effect of X on M (i.e., a). Since no such equally unambiguous component paths a and b are available in SOC3, there cannot possibly be an unambiguous indirect effect indicative of mediation. M can therefore not be a mediator in the common meaning of the concept (i.e., there is..
no indirect effect \( ab \) and thus even more not be a mediator and a moderator at the same time. Also, \( X \) may be a moderator of path \( b \), such that \( b \) is different at different values of \( X \), but since, as we have shown above, no effect of \( X \) on \( M \) is possible at a particular value of \( X \) (i.e., \( a = 0 \) for all groups of observations that are homogeneous with regard to \( X \)), there can be no indirect effects \( ab \) (i.e., \( ab \) is always 0 at any value of \( X \)) and thus no moderated mediation.

How then, should the \( X \times M \) interaction be interpreted?

We have shown that SOC3 leads to grave logical contradictions which precludes a straightforward interpretation of \( M \) as a mediator and a moderator at the same time (or \( X \) as an independent variable and moderator of the path \( b \) at the same time). However, the empirical scenario of an interaction effect of \( X \) and \( M \) on \( Y \) is certainly plausible: There may be an \( X \times M \) interaction effect among the three variables in the model with simultaneous substantial paths \( a \) and \( b \). Therefore we now turn to the question of how such an interaction effect should be interpreted.

A strict interpretation: SOC3 is invalid

A very strict interpretation considers the \( X \times M \) interaction as proof that fundamental assumptions necessary for mediation analysis are violated and therefore simple mediation cannot be established by SOC3 in the first place, much less moderated mediation. The term assumptions must not be taken here in a loose colloquial sense. The assumptions we refer to here are not speculative predictions about empirical states of affairs that could be tested, as an informal everyday use may suggest. Rather, in the original sense of formal logic, these assumptions are premises, conditions \textit{sine qua non}: If they are violated, the entire Baron and Kenny (1986) procedure to test mediation and its later variants that also rely on measured mediator candidate variables (Edwards & Lambert, 2007; Iacobucci, Saldanha, & Deng, 2007; MacKinnon,
A MEDIATOR CANNOT ALSO BE A MODERATOR

Lockwood, Hoffman, West, & Sheets, 2002; Preacher & Hayes, 2004; Shrout & Bolger, 2002) are void.

MacKinnon (2008) identifies as assumptions for mediation analysis – among others – a) additivity of $X$ and $M$, and b) well-behaved residuals as prerequisites for mediation to be tested using regression equations. Prerequisite b) comprises the stipulation that "[t]he residuals in each equation are [...] uncorrelated with the predictor variables in each equation, are independent of each other, and the residuals [...] have constant variance at each value of the predictor variable." (see also Edwards & Lambert, 2007). Both of these prerequisites are blatantly violated in SOC3.

Assumption a) is very clearly violated in SOC3: The interaction effect defies additivity of $X$ and $M$. Strictly speaking, mediation analysis becomes void in the presence of an interaction effect of the independent variable and the proposed mediator.

Assumption b) is also violated in SOC3: There are, of course, always unknown influences on measured variables, unless one is dealing with one of the rare cases in which 100% of the variance of a variable can be completely accounted for by other variables. These unknown influences can sometimes be subsumed under the error variance of the influenced variables, and be ignored at the cost of power but without bias in estimates. This, however, is only safe if the unknown influences are in fact randomly distributed and uncorrelated with residuals of other variables in a model. Randomization in experimental research provides the best known means to ensure these conditions. Without randomization to different levels of a mediator, however, correlated error terms are the rule rather than the exception if a mediator variable candidate is measured (Bullock, Green, & Ha, 2010; Herting, 2002) and pose an immense validity threat.

Most pertinently, as argued above, the $X \times M$ interaction proves that $M$ and $Y$ share variance that cannot be further systematized within the three variable model (i.e., the residuals of $M$ and $Y$ are
correlated). Bullock et al. (2010) provide formal proof that such shared variance systematically biases mediation models in that it can substantially increase type I errors in significance testing (i.e., increase the likelihood of falsely detecting mediation when there actually is none). Conversely, Herting (2002) alerts to the fact that, if the association between residuals of \( M \) and \( Y \) is negative in sign, this can affect the type II error probability, resulting in false rejection of valid mediation models (see also Sobel, 2008; Stone-Romero & Rosopa, 2008). Remedies against such bias from correlated errors in mediation analysis are available (e.g., Bullock et al., 2010; Gennetian, Morris, Bos, & Bloom, 2005; Shrout, 2010) or being developed (Jacoby & Sassenberg, in press), but they clearly contradict SOC3.

In sum, this strict interpretation of the \( X \times M \) interaction effect would only allow the conclusion that it is impossible to assess mediation with common mediation analysis strategies in the presence of this interaction effect. The interaction effect is empirical evidence that fundamental assumptions necessary for mediation analysis are violated. Therefore SOC3 should not be regarded as a valid model that helps advancing understanding of phenomena and processes.

A more constructive approach: Model modification

While the strict interpretation of the interaction in SOC3 would dictate to regard the model as falsified and completely abandon it, one can take the interaction effect in SOC3 (that shows that the model is necessarily contradictory and misspecified) as an occasion to modify the model to resolve this contradiction. After all, the data pattern that seems to agree with SOC3 is instructive and the modification of the model and its new empirical test may bring about deeper understanding eventually (MacKinnon, 2008).
A fourth variable. We have argued that the variance that is shared between $M$ and $Y$ in the form of simple slopes constituting an $X \times M$ interaction effect in SOC3 cannot be properly modeled within SOC3. But it may well be modeled in a modified model that incorporates at least a fourth variable. As noted earlier, the interaction indicates that there are two distinct components of $M$, or independent portions of variance in $M$: One portion that is shared with $X$ and $Y$, which one may refer to as $M_1$, and one which is shared with $Y$, but not with $X$, which one may denote by $M_2$. $M_1$, as already argued, is characteristic for the role of $M$ as a mediator of the effect of $X$ on $Y$. $M_2$, however, is indicative of additional variance that $M$ and $Y$ share. Since these two components of $M$ are obviously statistically independent and therefore have distinct psychological and theoretical significance, they should be considered and treated as two different variables (for a similar idea, see Edwards & Lambert, 2007, Figure 1E). These two variables may be semantically related, and the measurement instrument originally used to assess $M$ may tap into both of these variables. But the $X \times M$ interaction makes it clear that one is in fact dealing with two uncorrelated variables and thereby with two different constructs. These constructs should thus be separated and modeled using two different variables and measurement instruments. Then, the three variable constellation is transformed into a four variable constellation with variables $X$, $M_1$, $M_2$, and $Y$ (see Figure 5). Here, $M_1$ mediates the effect of $X$ on $Y$, and, in addition, $M_2$ and $X$ interact in predicting $Y$. This latter $M_2 \times X$ interaction may qualify an average effect of $M_2$ on $Y$ (see path f in Figure 5). This average effect is, however, equally irrelevant to the effect to be explained between $X$ and $Y$ and therefore of no interest for the question of whether $M_1$ (formerly a component of $M$) mediates this basic effect $X \rightarrow Y$. While this strategy to respond to the falsity of the SOC3 model is more liberal than that from the strict perspective above, it also requires new data collection. In order to actually have $M_1$ and $M_2$ as different variables, a new study must be
conducted in which these two variables must be separately measured, or, in the case of $M_2$, even manipulated.

One may argue against the call to split $M$ into two variables and hold that it is difficult to separate components of the variable $M$ in this way because the measurement instruments available in a particular area of research are not accurate enough to warrant such separation. Therefore, according to this argument, one has to manage with calling independent variances by the same name and treat them as one single variable. While we agree with the premises of that argument, we disagree with the conclusion for two reasons. First, if one thinks this conclusion through, it promotes a general practice of simply calling different constructs by the same name, an instance of what Thorndike (1916) has referred to as the 'jingle fallacy' (see also Kelley, 1927). This practice obscures the nature of psychological phenomena and processes rather than clarifying them. We think that such terminological imprecision will preserve and instigate theoretical, experimental and psychometric inaccuracy. It contradicts the general aim of scientific social and behavioral research to be precise and systematic in building a body of knowledge that is at least consistent, if not coherent.

Second, even if the general concerns we have just articulated were void, the specific endeavor of moderation and mediation analysis is – in the first place – to better understand relationships, to obtain a more accurate and complete picture of phenomena in a particular area of investigation. This fundamental aim behind the examination of third variables is clearly simultaneously betrayed if one accepts a model that is a priori false (i.e., of which one knows that it is misspecified even before data are collected and that analysis of the data, especially if it reveals the hypothesized relationships, will additionally confirm to be misspecified). The model which one intends to test in the service of furthering the understanding of phenomena is guaranteed to be false. To employ and test SOC3 thus means to accept its inherent logical
contradictions. Such acceptance is not a pragmatically necessary tolerance of ambiguity, it would be inherently unscientific and irrational. In sum, in contrast to the strict interpretation of the mediation model being void because of the clear violation of assumptions that are necessary to test it, the more constructive approach takes the $X\times M$ interaction as important evidence for the need to modify an SOC3 model and expand it to a four variable model.

A different relationship. Another strategy to modify the three variable model so that it can accommodate the anticipated data pattern but avoid the logical contradictions discussed is based on positing different relationships between the variables than the ones assumed so far. SOC3, in its common form, implies linear relationships $a$, $b$, and $c'$. The interaction, $d$, also essentially describes a linear change in slopes. Mediation analysis is, however, by no means restricted to linear relationships (Hayes & Preacher, 2010). The underlying statistical tool, regression analysis, can easily accommodate higher order relationships, such as quadratic, cubic or higher order trends as well as more complicated functional relationships (e.g., exponential growth or inverse functions). If such trends make conceptual sense and closely correspond to hypotheses about psychological phenomena, linear and non-linear relationships can produce patterns that may look like they are composed of linear associations and strictly linear interactions at first glance. Upon closer inspection, these patterns may however turn out to be more parsimoniously described by straightforward combinations of linear and non-linear relationships. For example, the data pattern hypothesized in SOC3 can be adequately modeled using a linear path $a$, and a linear along with a quadratic component in $b$ without having to invoke the problematic $X\times M$ interaction. Then, $M$ is a simple mediator of a non-linear indirect effect, and no additional moderation hypothesis is necessary.
The logical contradictions surrounding the original SOC3 can thus be avoided by the use of models containing non-linear relationships. However, the methods to test these relationships in a focused manner are just being developed (Hayes & Preacher, 2010) and further elaborating on non-linear mediation models is beyond the scope of the present discussion.

Can structural equation modeling resolve the problem?

There are numerous calls to conduct tests of mediation hypotheses using structural equation modeling (SEM), instead of common linear regression models (e.g., Edwards & Lambert, 2007; Iacobucci, 2008). A particular SEM approach, namely path analysis, provides a specific formalization of SOC3 (Preacher et al., 2007, p. 194, Figure 2, Model 1) and is shown here in Figure 6. It contains the manifest variables $X$, $M$, and $Y$, and the interaction term $X \times M$ along with the paths $a$, $b$, and $c'$ as well as a path $d$ denoting the interaction effect. This model however reveals that accounting for the interaction effect requires the same model modification measures that we have sketched above in more general terms. Since the interaction term in this model is a perfect mathematical function of $X$ and $M$, and $X$ and $M$ are by definition correlated (i.e., path $a$), the model must accommodate the fact that the interaction term inevitably covaries with $X$ and, most pertinently here, with $M$. Since however $M$ is an endogeneous variable the causal logic necessary for consistent SEM models precludes the simple introduction of a covariance parameter in order to account for the covariance of $X \times M$ and $M$. The incorporation of this association must be achieved by separating the variable $M$ into $M$ and an error term $r_M$. Then, while the covariance between $X$ and $X \times M$ is exactly modeled as such, the covariance of the interaction term and $M$ must be modeled as the covariance between $X \times M$ and the independent error $r_M$ that influences measured values of $M$. This practice is common and legitimate in order to preserve identifiability of the model. However, with the manipulated variable $X$ and $M$ as a
measured, one-dimensional variable, the data collected based on SOC3 does not allow for separate estimation of $M$ and $r_M$. Thus, the path-analytic approach to SOC3 yields the same conclusion as our discussion above: In order for SOC3 to be estimated properly and provide grounds for further elucidation of phenomena and processes, $M$ must be split into two separate components.

One important advantage of SEM approaches is that they can accommodate latent variables and therefore often provide opportunities to explicitly model measurement error (i.e., residuals) and thus separate it from a psychological construct proper (Baron & Kenny, 1986). However, even latent variable modeling will not resolve the problem of SOC3. As Iacobucci (2008) shows, the introduction of a fourth variable that has an influence on $M$ or $Y$ into a latent variables simple mediation model can yield substantially distorted estimates of the paths $a$ or $b$. The $X \times M$ interaction term, if it were introduced into the model, would in fact have to be modeled precisely as an influence on $M$ or $Y$. This would thus severely damage inference regarding the simple mediation of $X \rightarrow Y$ via $M$ that is postulated in SOC3: Instead of providing better understanding of the processes occurring among $X$, $M$, and $Y$, this model yields unreliable estimates.

In a more general sense, the problem in SOC3 we have discussed is not one of a particular estimation method chosen, it is a problem of model identification. In essence the issue of model identification is concerned with the question of how much reasonably accurate information one can obtain from a given number of variables. SEM approaches reflect this concern in the stipulation that models should be ideally overidentified (i.e, have a positive, non-zero number of degrees of freedom). Overidentification means essentially that the information one can retrieve from data exceeds the conclusions one potentially wants to draw. If the overshoot of information
in the data still agree with the hypothesized model, this lends solid credibility to the model. Occasionally, just identified models are considered in research, even though, in a strict sense, it cannot be tested whether the empirical data fit such models. Such models have 0 degrees of freedom. The original three variable model of simple mediation (see Figure 1) is one such just identified model: It has 0 degrees of freedom and is thus not testable as modeled. Recovering the paths \( a \), \( b \), and \( c' \) is the absolute maximum of reasonable knowledge that can be drawn from that simple mediation model with three variables. Adding the interaction of \( X \) and \( M \) to the model with the same measurements does not solve this problem, but exacerbates it because the \( X \times M \) term is a perfect function of other variables in the model (\( X \) and \( M \)). The number of degrees of freedom is thus actually further reduced. Testing the interaction effect along with the paths in the original simple mediation model amounts to modeling four variables with three measures.

In sum, since model identification is at least as important in SEM as in regression analysis, and misspecification of the model underlying data collection is as problematic for SEM as it is for regression analysis-based approaches, the contradiction of trying to recover more information from three variables than possible also becomes obvious within SEM approaches.

Additional notes

Why are computational methods for SOC3 questionable?

There are methods to test second order constellations that specifically target SOC3. Most prominently, Preacher et al. (2007) provide a macro that assesses moderated indirect effects in this constellation using bootstrapping procedures to estimate confidence intervals (p. 194ff., Model 1). Why does this procedure (and others) yield meaningful results if the three variable second order constellation is logically contradictory? The bootstrapping procedure for SOC3 in Preacher et al. (2007) essentially assesses the product of the path \( a \) estimated across all
observations, and different paths $b$ for subgroups of cases with equal values of $X$. Mathematically these products have a sampling distribution and are amenable to estimation by bootstrapping procedures, but they are psychologically questionable as they compute \emph{conditional} indirect effects (i.e., different products $ab$) using an \emph{unconditional} effect $a$.

The MacArthur approach

Similar to our conclusion here, the recently proposed MacArthur approach to mediation and moderation (Kraemer, Wilson, Fairburn, & Agras, 2002; Kraemer, Kiernan, Essex, & Kupfer, 2008) also rejects the notion of a variable that could simultaneously mediate and moderate an effect $X \rightarrow Y$. This approach, however, also discusses the $X \times M$ interaction effect in SOC3 and offers the interpretation of this interaction effect as an independent indicator of the mediating role of $M$. Importantly, Kraemer et al. (2008) claim that the interaction effect is indicative of $M$ playing a mediating role because "$M$ plays a specific role in determining the effect size [total variance explained in $Y$] and thus fulfills the conceptual definition of a mediator of $T$" (p. S104) where $T$ is an independent variable that we refer to as $X$ here. The MacArthur approach thus proposes that \emph{by virtue of} the $X \times M$ interaction effect a variable $M$ is a mediator (see Kraemer et al., 2008, p. S106, Table 2). Thus, while the discussed perspectives on SOC3 ("$M$ mediates and $M$ moderates" and "$X$ moderates") revolved around the question of how to interpret an $X \times M$ interaction with regard to the hypothesis of a second order constellation, the MacArthur approach declares the interaction effect in SOC3 to be an indicator of simple mediation. The present discussion has however shown that the statistical model comprising the interaction effect is contradictory, independently of possible differences in terminology. The presently proposed modification of SOC3, the separation of two different components $M_1$ and $M_2$, one of which moderates and one of which mediates $X \rightarrow Y$, makes it obvious that there is a mediation in the
common sense, and a moderation in a common sense. The new definition of mediation in the MacArthur approach therefore not only relies on a contradictory model, but it also unnecessarily redefines terminology.

Research strategic implications

Based on the foregoing considerations we propose three practical implications:

First, when in the course of planning the testing of a mediation hypothesis, the model guiding data collection comprises a variable $M$ that acts both as a mediator and a moderator of a particular basic effect $X \rightarrow Y$, researchers should make sure to theoretically recover from the model the separate variables $M_1$ and $M_2$ and collect data accordingly. Otherwise, their model will be guaranteed to be false.

Second, if data based on a three variable model have already been collected and $X$ turns out to interact with $M$ in predicting $Y$ during analysis, the three variable mediation model is definitely false. In this case, researchers should refrain from the simple conclusion that $M$ is a mediator of the effect of $X$ on $Y$, and especially from the conclusion that $M$ is a mediator and a moderator of $X \rightarrow Y$. Similarly, $X$ should not be considered as moderating its own direct effect via $M$.

Third, the $X \times M$ interaction effect should be routinely subjected to statistical testing in simple mediation models. Only if this interaction proves to be far from being conventionally significant the common test of simple mediation should be performed. Since one tests for instead of against the null hypothesis in this case, the traditionally used decision criterion, the $p$-value, should be rather high, everything else being equal, to ensure adequate power (i.e., rigor in testing the null hypothesis, Hager, 2000). The most accurate and sensible decision criterion will result from an a priori power analysis securing a rather small $\beta$ error in a study (Cohen, 1988; Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). If the interaction
effect emerges as substantial, researchers should theoretically identify the different components of $M$ in order to arrive at a more sophisticated model comprising $M_1$ and $M_2$ and collect data to test this refined four variable model.

Conclusions

We have argued that a particular model of moderated mediation, namely that in which the mediator of an effect seems to also moderate this effect leads to contradictions. Therefore it does not represent a case of what Muller et al. (2005), or Edwards and Lambert (2007) have defined and formalized as moderated mediation. Instead, the interaction effect between an independent variable and a mediator turns out to indicate that the model is inherently misspecified. The misspecification should not be mistaken for an issue of empirical corroboration or estimation method, it rather reflects an implicit aim to extract more information from given data than is possible and thus a genuine conceptual problem.

We recommend to expand any hypothesized mediation model with three variables that foresees an interaction between the independent and mediator variables, to become a four variable model. If an interaction effect unexpectedly emerges from analyses of collected data, we recommend to modify the model in the same way and to refrain from incorrectly claiming that a variable acts as a mediator and a moderator at the same time. When it comes to moderation and mediation within the same model, it takes at least four variables to tango. Three variables will probably only allow for a clunky waltz.
References


(Ed.), *Learning more from social experiments. Evolving analytic approaches.* (pp. 75-114). New York, NY: Russell Sage.


Footnotes

1 Reiterating the exact steps of the analysis strategy recommended by Baron and Kenny (1986) and the characteristics of the variables and paths is beyond the scope of the current discussion. We assume here that most readers are meanwhile well acquainted with the general terminology explicit in Figure 1.

2 We will assume for now that all requirements for a state of the art study design and data collection have been met, in particular, that $X$ has been manipulated. We do so because our argument does not rest on a study compromised by common flaws that have extensively been discussed in the past. Also, the exact way of how mediation is assessed (e.g., by a sequence of regression models as prescribed by Baron & Kenny, 1986, the test developed by Sobel, 1982, or resampling methods as introduced by Preacher & Hayes, 2004) is of little importance in the current context.

3 This third argument that we present does not depend on whether one considers $c$ or $c'$ as the single effect of $X$ on $Y$. The problems arising from the interaction effect for the interpretation of the individual predictor effects is present regardless of whether one considers $M$ to moderate $c$ or $c'$.

4 When estimating models containing interaction effects, there are good reasons to center dichotomous and continuous predictors (Aiken & West, 1991) even though centering is not necessary if one carefully interprets coefficients associated with lower order predictors.

5 Other approaches to test mediation, such as structural equation modeling (e.g., Iacobucci, 2008) or bootstrapping (e.g., Preacher & Hayes, 2004) can, for current purposes, be considered as more sophisticated approaches to mediation effect estimation than common regression equations (Baron & Kenny, 1986). Therefore, the assumptions discussed here are equally applicable to or at least have an equivalent within these more advanced methods.
These two different variables may also be two measurements of the same construct at different consecutive points in time $t_1$ and $t_2$. In such a special case of a longitudinal design, they should however be modeled as separate variables instead of being collapsed into a single variable. If this is adequately done, it is obvious that the same variable cannot both mediate and moderate the same basic effect $X \rightarrow Y$. Rather, for example, the measurement at $t_1$ acts as the moderator (measured before $X$), and the measurement at $t_2$ as the mediator (measured after $X$ and before $Y$). The two measurements must then be conceptualized as two variables and at least consistently with the temporal implications of the modeled variables $X$ and $Y$. 

Figure captions

Figure 1. The simple mediation model.

Figure 2. The simple moderation model.

Figure 3. The basic three variable second order constellation SOC3 in its two conceptual versions.

Figure 4. Fictitious data pattern for a study of the effects of a coaching programme ($X$) on academic performance ($Y$) via self-management skills ($M$).

Figure 5. The formerly three variable second order constellation after separation of $M$ into the independent components of $M_1$ and $M_2$.

Figure 6. The three variable second order constellation after Preacher et al. (2007, p. 194, Model 1).
Note. $X$ = independent variable; $M_e$ = mediator variable; $Y$ = dependent variable. $c$ is the total effect of $X$ on $Y$ when $M_e$ is ignored, $c'$ is the direct effect of $X$ on $Y$ while controlling for $M_e$. The indirect effect, or mediation effect, is $(c - c') = ab$. 

Figure 1
Figure 2

Note. $X$ = independent variable; $M_o$ = moderator variable; $Y$ = dependent variable. $d$ denotes the $X \times M_o$ interaction effect on $Y$ and represents the degree of moderation.
A MEDIATOR CANNOT ALSO BE A MODERATOR

Figure 3

Note. $X =$ independent variable; $M =$ third variable; $Y =$ dependent variable. $d$ denotes the interaction effect among $X$ and $M$ in predicting $Y$. Panel A shows Perspective I, Panel B shows Perspective II (see text for details). Empirically and statistically, the two perspectives in A and B are identical.
Figure 4

Note. Panel A shows the pattern with $X$ as a moderator of the effect of $M$; Panel B shows the pattern with $M$ as the moderator of the effect of $X$; the lines represent simple slopes at arbitrary, but sensible values of $M$ (e.g. -1 SD and +1 SD).
Note. $e$ denotes the interaction effect of $X$ and $M_2$ on $Y$, $f$ represents the average effect of $M_2$ on $Y$ which is not pertinent to moderation or mediation in this model.
Note. $X = \text{independent variable}, \quad M = \text{mediator variable}, \quad Y = \text{dependent variable}$. $XM$ is the product interaction term of $X$ and $M$, $d$ denotes the size of the interaction effect among $X$ and $M$ in predicting $Y$. $r_M$ denotes an error that influences the measurement values of $M$ and necessarily covaries with the product term $XM$. 