Simpson's paradox, moderation, and the emergence of quadratic relationships in path models: An information systems illustration

Ned Kock
Leebrian Gaskins

Full reference:

Kock, N., & Gaskins, L. (2016). Simpson's paradox, moderation, and the emergence of quadratic relationships in path models: An information systems illustration. *International Journal of Applied Nonlinear Science*, 2(3), 200-234.

Abstract

While Simpson's paradox is well-known to statisticians, it seems to have been largely neglected in many applied fields of research, including the field of information systems. This is problematic because of the strange nature of the phenomenon, the wrong conclusions and decisions to which it may lead, and its likely frequency. We discuss Simpson's paradox and interpret it from the perspective of path models with or without latent variables. We define it mathematically and argue that it arises from incorrect model specification. We also show how models can be correctly specified so that they are free from Simpson's paradox. In the process of doing so, we show that Simpson's paradox may be a marker of two types of co-existing relationships that have been attracting increasing interest from information systems researchers, namely moderation and quadratic relationships.

Keywords: Nonlinear Relationships, Simpson's Paradox, Path Analysis, Moderation Effects, Information Systems.

Introduction

Among phenomena that arise from statistical analyses, few have the potential to cause puzzlement and misinterpretations to the extent that Simpson's paradox does (Pearl, 2009; Wagner, 1982; Westfall et al., 2007). This phenomenon routinely leads to conclusions about causal effects that are the opposite of the corresponding overall true effects – an effect is interpreted as being negative when in reality it is overall positive. While Simpson's paradox is well known to statisticians and has received attention, particularly in specific contexts within certain fields, such as investigations of the effects of medical interventions and social policies (Pavlides & Perlman, 2009), it seems to have been largely neglected in many applied fields of research. This includes the field of information systems (IS).

We conducted detailed reviews of articles published in the 2003-2013 period in two leading academic journals in the field of IS: *Information Systems Research* and *MIS Quarterly*. Among the articles reviewed in this period, at least 14 contained instances of Simpson's paradox in *Information Systems Research*, and 19 in *MIS Quarterly*. These amounted to approximately 9.2 and 15.7 percent, respectively, of all articles published in those journals that employed path analysis or related methods – multiple regression analysis and structural equation modeling. This is arguably a fairly high rate of incidence. We also conducted additional reviews covering articles in those journals since the journals' initial publication dates using various search tools and academic databases. Based on these additional reviews, it appears that Simpson's paradox has never been explicitly checked for or even considered in research published in these journals.

This situation is problematic because Simpson's paradox may lead to misinterpretations of results and consequently to conclusions that are completely disconnected from true relationships among constructs at the population level in IS investigations. For example, Simpson's paradox may lead to the conclusion that the use of a technology causes a decrease in group performance in the context of a particular task when in fact the opposite may be generally true – technology use generally causes an overall increase in group performance. We use the qualifier "overall" here because, as it will be shown, Simpson's paradox is associated with a change in the sign of the slope of a relationship in sub-samples of the same dataset (Pearl, 2009), with the overall slope sign being inconsistent with at least one of the sub-sample slope signs.

Moreover, mathematically Simpson's paradox is associated with relationships that have attracted much interest in the field of IS. As will be demonstrated, when viewed through the lens provided by path models (Anderson et al., 1995; Wright, 1960), Simpson's paradox appears to arise from incorrect model specification (Fiedler et al., 2003). As models are revised and correctly specified, two types of hidden relationships are revealed – moderation and quadratic relationships. Moderation relationships have been the target of much interest in the field of IS (Carte & Russel, 2003; Sykes et al., 2014). Quadratic relationships are common types of nonlinear relationships (Bremmer, 2007), which have been recently receiving increasing attention in the field of IS (Guo et al., 2011; Schmiedel et al., 2014).

We discuss Simpson's paradox, showing why it can lead to confusion, particularly when viewed through the lens provided by contingency tables. We then re-interpret it from the perspective of path models, defining it in precise mathematical terms and showing that it often arises from incorrect model specification. Next we show how models can in many cases be correctly specified so that they are free from Simpson's paradox. In the process of doing so, we put forth the notion that Simpson's paradox may be a marker of two types of relationships that

co-exist: moderation and quadratic relationships. We illustrate our findings based on analyses of true model data and real data, and provide several recommendations for researchers.

Simpson's paradox in contingency tables

Simpson's paradox has received much attention in some fields, particularly medicine. Julious & Mullee (1994) provide an overview of related studies, and a classic illustration of Simpson's paradox based on a kidney stone treatment study. This kidney stone treatment study illustration, discussed in Appendix A, is perhaps the most commonly used to define Simpson's paradox using contingency tables.

Table 1 illustrates a fictitious IS study scenario that is analogous to that in the kidney stone treatment study mentioned above. It shows success rates of business process improvement projects under two technology use conditions, use and no use; and two different management styles, democratic and autocratic. Within parentheses are the numbers of successful projects in each condition (i.e., combination of technology use and management style), divided by the total number of projects in the condition. This fictitious scenario bears some similarities to a real action research investigation of the effects of an asynchronous collaboration technology on the outcomes of business process improvement projects (Kock, 2004). The technology investigated in this action research study was similar to an e-mail distribution list, and it was used in the context of democratically and autocratically managed divisions of the same organization. The numbers used are the same as those in the kidney stone treatment study.

Table 1: IS illustration of Simpson's paradox

	Technology use	**No technology use**
Democratic mgt. style	***Cell 1*** **93%** (81/87)	***Cell 2*** 87% (234/270)
Autocratic mgt. style	***Cell 3*** **73%** (192/263)	***Cell 4*** 69% (55/80)
Both mgt. styles	***Cell 5*** 78% (273/350)	***Cell 6*** **83%** (289/350)

Note: Percentages reflect group success rates in business process improvement projects

Faced with the results of this technology and management styles scenario, we would be pressed to conclude that the success rates of business process improvement projects are higher when technology is used in the context of democratic and autocratic management, when we look at each of the management styles (democratic or autocratic) individually – cell pairs (1, 2) and (3, 4). When these management styles are combined, in cell pair (5, 6), business process improvement projects seem to perform better when technology is *not* used.

Different recommendations would be given to practitioners based on this study. If we know the management style employed in an organization, then we would recommend technology use. If we do not know the management style, we would recommend no technology use. In other words, whether we have knowledge of the management style or not defines the optimal choice. This is an absurd conclusion.

This fictitious IS illustration that shows the absurdity of Simpson's paradox, when viewed through the limited lens afforded by contingency tables, is provided here to set the stage for the conceptual and mathematical discussions in the next sections. An IS illustration based on data

where we know the true model coefficients is presented later, and an additional matched IS illustration based on real data is provided as an appendix.

Mathematical notation used

The mathematical notation employed here is the classic path analysis notation (Duncan, 1966; Kenny, 1979; Mueller, 1996; Wright, 1934; 1960), with a few necessary modifications. The method of path analysis was developed independently from multiple regression analysis, even though these two methods bear many similarities (Wright, 1960). The mathematical notation employed is briefly summarized below through examples.

- p_{ZX} : path coefficient associated with the path connecting two variables *X* and *Z*, namely the path $X \rightarrow Z$. Path coefficients are calculated in the same way as standardized partial regression coefficients in multiple regression.
- $p_{Z\gamma}$: path coefficient associated with the path connecting the error term γ for the variable *Z* and the variable *Z* itself, namely the path $\gamma \rightarrow Z$. In path analysis, differently from multiple regression analysis, path coefficients for error terms are explicitly included in equations.
- m_{ZYX} : path coefficient associated with a moderating effect of a variable *X* on the link between two other variables *Y* and *Z*, namely $X \rightarrow (Y \rightarrow Z)$. Here the three symbols ZYX are used in the subscript because the path coefficient refers to a relationship that involves three variables.
- n_{ZY} : path coefficient associated with the path connecting two variables *Y* and *Z*, namely the path $Y \rightarrow Z$, explicitly modeled as a nonlinear path. In this paper we restrict our analyses to one type of nonlinear relationship, the quadratic relationship. To explicitly model a path as quadratic we need to find the function $F_J(Y) = E(Z|Y) = aY + bY^2$ that best fits the relationship between *Y* and *Z*.

The order of the subscripts in the symbols above follows the classic path analysis notation, where the criterion variable appears before the predictor. For example, the subscript ZX refers to the link $X \rightarrow Z$. The notation above is used in various mathematical equations presented in the following sections.

Simpson's paradox in path models

As Pearl (2009) notes in his seminal exploration of the Simpson's paradox phenomenon within the broader context of causality assessment, one of the reasons for the confusion surrounding Simpson's paradox is the limited view provided by contingency tables. This can be solved by the use of path models, where Simpson's paradox can be defined in precise mathematical terms. In the case of the technology and management styles study, we can do this by creating variables and coding them properly. Let us consider the following variables:

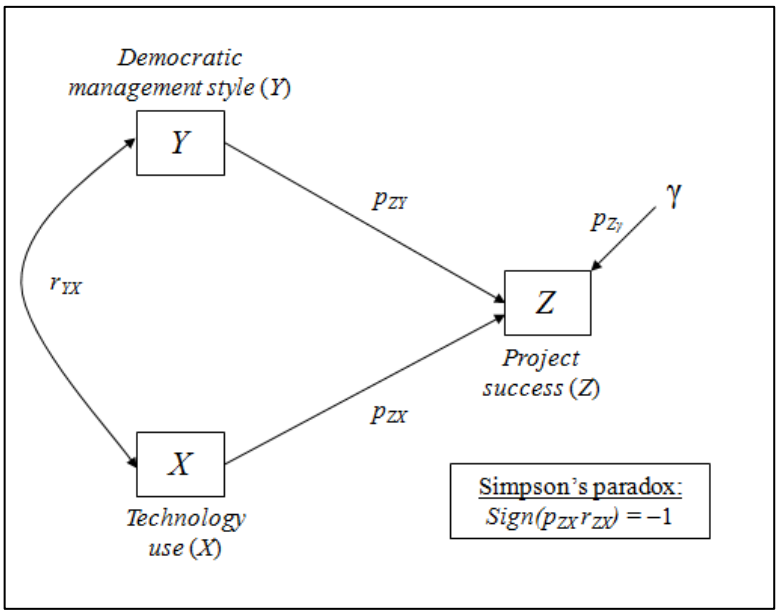
- Technology use (*X*), coded as 0 and 1, for the conditions where technology was not used and was used, respectively.
- Democratic management style (*Y*), coded as 0 and 1, for the non-democratic (i.e., autocratic) and democratic management styles, respectively.
- Project success (*Z*), coded as 0 and 1, for instances in which a business process improvement project was not successful and was successful, respectively.

When we look at the contingency table in the technology and management styles scenario presented earlier, we can see that in the pairs of cells (1, 2) and (3, 4) only one value (0 or 1) of the democratic management style (*Y*) variable is considered. So these pairs of cells refer to the association of the technology use (*X*) variable with the project success (*Z*) variable, when the democratic management style (*Y*) variable is kept constant. Or, stated in multivariate statistics terminology, when we control for the democratic management style (*Y*) variable. In the pairs of cells (1, 2) and (3, 4) the association of the technology use (*X*) variable on the project success (*Z*) variable is clearly positive. When technology is used, success rates go up.

However, when we consider the pair of cells (5, 6) we do not control for the democratic management style (*Y*) variable. Here the association of the technology use (*X*) variable with the project success (*Z*) variable is clearly negative. When technology is used, success rates go down. This means that the associations between the technology use (*X*) variable with the project success (*Z*) variable have slopes with different signs depending on whether we do or do not control for the democratic management style (*Y*) variable.

The path model in Figure 1 illustrates Simpson's paradox consistently with the discussion above. The association between the technology use (*X*) variable with the project success (*Z*) variable, controlling for the democratic management style (*Y*) variable, is measured by the path coefficient p_{ZX}. The association between the technology use (*X*) variable with the project success (*Z*) variable, *not* controlling for the democratic management style (*Y*) variable, is measured by the correlation coefficient r_{ZX}. Simpson's paradox occurs when the path coefficient p_{ZX} and the correlation coefficient r_{ZX} have different signs.

Figure 1: Simpson's paradox in a path model



In this path model, the relationships among the variables *X*, *Y* and *Z* are described by (1). Since Simpson's paradox is a marker of an incorrect specification in terms of causality (Pearl, 2009), both this equation and the path model are likely to be incorrectly specified. In this case, the incorrect specification is expected to be related to the path associated with Simpson's paradox, namely the path pointing from *X* to *Z*.

$$Z = p_{ZY}Y + p_{ZX}X + p_{Z\gamma}\gamma. \quad (1)$$

Mathematically, a Simpson's paradox occurrence can be defined as an instance in a path model that satisfies the following condition:

$$Sign(p_{ZX}r_{ZX}) = -1.$$

The alternative form below will be useful later when we discuss other aspects of the Simpson's paradox phenomenon.

$$p_{ZX}r_{ZX} < 0.$$

Simpson's paradox, like many statistical phenomena involving multiple variables, arises from specific combinations of correlations among variables. In Appendix B we show, using the alternative equation form above in the context of Monte Carlo simulations (Robert & Casella, 2005), that many combinations of correlations among variables give rise to Simpson's paradox. These include combinations where the correlations among variables are not particular high. Our Monte Carlo simulations show that the probability that Simpson's paradox will occur increases as correlations among variables increase. This suggests that Simpson's paradox may be a relatively common occurrence in research studies that has nevertheless been largely ignored in the field of IS and many other fields.

Statistical suppression is often equated to Simpson's paradox (MacKinnon et al., 2000; Spirtes et al., 1993, Tu et al., 2008), even though these are phenomena with different defining characteristics. This is due in part to the various definitions of statistical suppression that have been proposed over the years (Cohen et al., 2003; MacKinnon et al., 2000). In Appendix C, we build on Conger's (1974) widely accepted definition of classic statistical suppression to explain the difference between it and Simpson's paradox, as well as to explain why these two phenomena are often equated.

So far we have defined Simpson's paradox mathematically and in terms of an incorrectly specified path model. In the following section, we turn our attention to how to solve Simpson's paradox by correctly specifying the path model.

Simpson's paradox and moderation

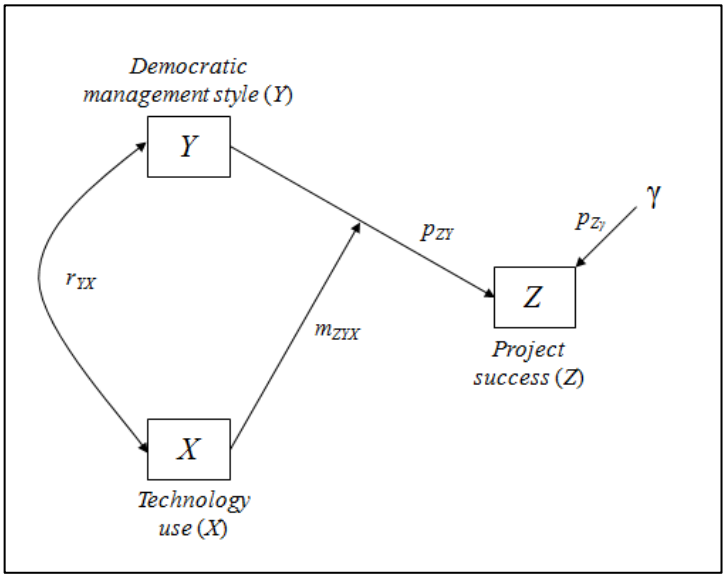
A series of experiments on inductive reasoning conducted by Fiedler et al. (2003) show that a three-variable instance of Simpson's paradox can be seen as a misspecification of a "constrained" correlated moderation relationship as a relationship between two predictors and a criterion variable. In a constrained correlated moderation relationship, a predictor variable moderates the relationship of a correlated predictor with a criterion variable, without affecting the criterion variable through a direct path. We demonstrate the correctness of Fiedler et al.'s (2003) conclusion later, through an illustrative analysis with true model data.

Figure 2 shows how a model can be correctly specified so that the confusion associated with Simpson's paradox is solved through the removal of a direct path and inclusion of a constrained correlated moderation relationship. In it, the path going from *X* to *Z* in the incorrectly specified model is removed, as it is related to an incorrect specification of a causal relationship. Also, a new moderation path is included going from *X* to the path going from *Y* to *Z*.

In the new, correctly specified model, the relationships among the variables *X*, *Y* and *Z* are described by (2). Note that in this equation there is no term representing a direct path going from *X* to *Z*. As it will become clear in the foregoing discussion and illustrative analyses, such a path, if included, would again lead to Simpson's paradox. The absence of the direct path going from *X* to *Z* does not mean that *X* is not associated with *Z*. An indirect association remains, via moderation.

$$Z = p_{ZY}Y + m_{ZYX}XY + p_{Z\gamma}\gamma. \quad (2)$$

Figure 2: Solving Simpson's paradox by correctly specifying the path model



In the correctly specified model above the variables *X* and *Y* are assumed to be correlated but not causally associated via direct links. Because of this we refer to it as a model with a constrained "correlated" moderation relationship. There are two model variations of this correctly specified model, which correspond to two variations of the incorrectly specified model presented earlier. We present first the incorrectly specified model variations, in Figure 3. These variations are generally equivalent to one another, and to the incorrectly specified model presented earlier, for the purposes of the discussion presented here.

The two incorrectly specified model variations lead to two corresponding correctly specified model variations. These correctly specified model variations are shown in Figure 4. The one on the left includes a constrained "mediated moderation" relationship; where the predictor *X* affects a mediator *Y* of its (i.e., *X*'s) relationship with *Z*, and also acts as a moderator of the path going from *Y* to *Z*. The correctly specified model variation on the right includes a constrained "moderated mediation" relationship; where *X* is affected by *Y*, and also acts as a moderator of the path going from *Y* to *Z*.

While these variations are not causally equivalent, the discussion presented in this paper applies to all of them. As a group, they are described by very similar sets of equations. These include (3) and (4). The first equation refers to the correctly specified model including the constrained mediated moderation relationship. The second equation refers to the correctly specified model including the moderated mediation relationship. These equations apply to the

corresponding incorrectly specified models, and are restricted to paths that are in fact correctly specified – i.e., the paths from *X* to *Y* and from *Y* to *X*, respectively.

$$Y = r_{YX}X + p_{Y\delta}\delta. \quad (3)$$
$$X = r_{XY}Y + p_{X\varepsilon}\varepsilon. \quad (4)$$

Figure 3: Variations of previous path model that may lead to Simpson's paradox

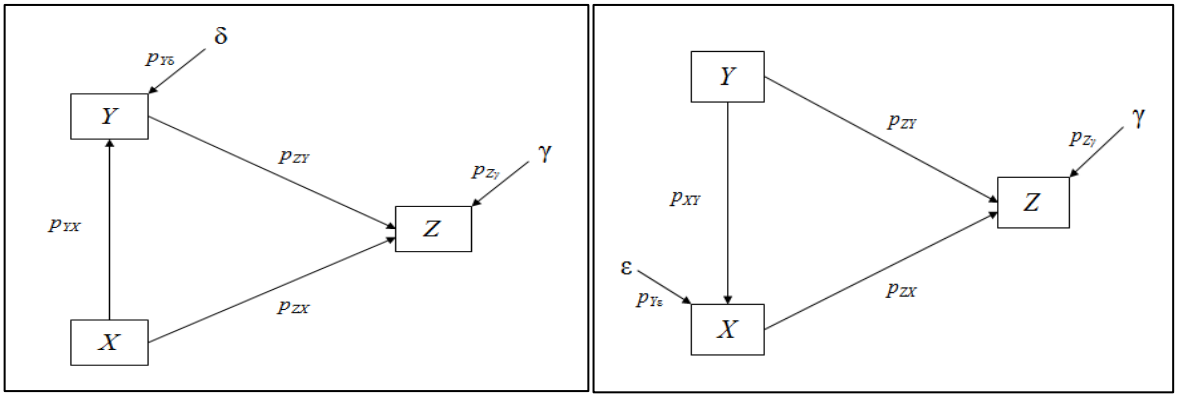
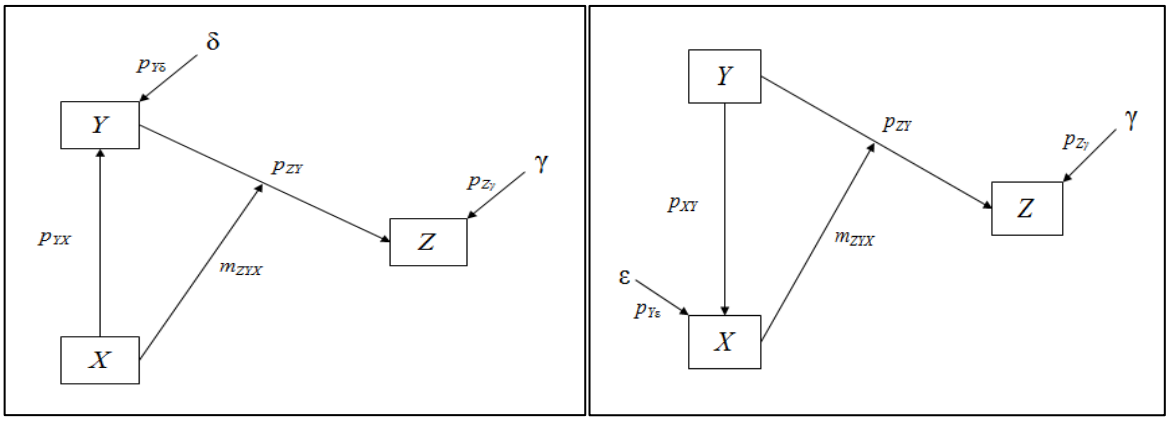


Figure 4: Correctly specified path model variations that solve Simpson's paradox



The equations above are presented as contributing to a full description of the model variations. They nevertheless express a much more general property of correlations. Mathematically, we can always express a linear bivariate association among any pair of variables in a path model, whether the variables are causally linked or not, through a regression equation including a correlation term and an error term (Mueller, 1996; Wright, 1960). This property applies to linear bivariate associations, between pairs of variables, not to multivariate associations.

Next we describe a remarkable property of the correctly specified models. They can give rise to true quadratic relationships, which coexist with moderation relationships as long as the nonlinear relationships are not explicitly modeled. Otherwise the moderation relationships tend to become weaker and disappear, contingently on the strength of the nonlinear relationships.

The emergence of true quadratic relationships

Exploratory analyses where quadratic relationships are explicitly modeled present a key challenge to researchers, which is that "fake" quadratic relationships may arise from error. This is the reason why we use the qualifier "true" to refer to quadratic relationships here; a point that will soon become clearer. As mentioned earlier, we can always express a linear association between any two variables, such as *X* and *Y*, through a regression equation including a correlation term and an error term. For consistency we use the same symbols as before for such an equation expressing *X* in terms of *Y*:

$$X = r_{XY}Y + p_{X\varepsilon}\varepsilon.$$

Combining this with the correctly specified model equation shown earlier, expressing *Z* in terms of *Y* and the product *XY* (Equation 2), leads us to (5) below. Note that this combination is appropriate because of the equivalence of the models to which (2), (3), and (4) refer. State differently, the path coefficients for the links between *X* and *Y*, whether *X* points at *Y* or vice-versa (mediated moderation and moderated mediation), both equal the correlation between *X* and *Y*.

$$\begin{aligned} Z &= p_{ZY}Y + m_{ZYX}Y(r_{XY}Y + p_{X\varepsilon}\varepsilon) + p_{Z\gamma}\gamma \rightarrow \\ Z &= p_{ZY}Y + m_{ZYX}r_{XY}Y^2 + m_{ZYX}p_{X\varepsilon}Y\varepsilon + p_{Z\gamma}\gamma. \end{aligned} \quad (5)$$

The property that we can always express a linear bivariate association between any two variables through a regression equation including a correlation term and an error term also applies to error variables themselves, because they are also variables in the model. Therefore, we can express the error term ε as a function of *X*:

$$\varepsilon = r_{\varepsilon X}X + p_{\varepsilon\theta}\theta.$$

Combining this with (5), leads to an equation that expresses *Z* in terms of a quadratic function of *Y* and the product *XY*:

$$\begin{aligned} Z &= p_{ZY}Y + m_{ZYX}r_{XY}Y^2 + m_{ZYX}p_{X\varepsilon}Y(r_{\varepsilon X}X + p_{\varepsilon\theta}\theta) + p_{Z\gamma}\gamma \rightarrow \\ Z &= p_{ZY}Y + m_{ZYX}r_{XY}Y^2 + m_{ZYX}p_{X\varepsilon}r_{\varepsilon X}YX + m_{ZYX}p_{X\varepsilon}p_{\varepsilon\theta}Y\theta + p_{Z\gamma}\gamma. \end{aligned}$$

Since the error variables θ and γ are assumed to be uncorrelated with Z, the last two error terms on the right in the equation above can be combined into one single error term for a new uncorrelated error variable λ, leading to (6). This is due to the expected values of Z given θ, γ and the product $Y\theta$ being all zero; that is, the variables θ, γ and $Y\theta$ and uncorrelated with Z. In the specific case of $Y\theta$, the pattern of uncorrelated variation in θ with respect to Z is transferred to the product $Y\theta$, leading to $Z \perp Y\theta$. Equation (6) applies to the three correctly specified model variations presented earlier, and expresses the fact that any of the three variations leads to a quadratic component that coexists with a moderation component.

$$Z = p_{ZY}Y + m_{ZYX}r_{XY}Y^2 + m_{ZYX}p_{X\varepsilon}r_{\varepsilon X}XY + p_{Z\lambda}\lambda. \quad (6)$$

In this equation, the quadratic component is made up of two terms: a linear term $p_{ZY}Y$, and a nonlinear term $m_{ZYX}r_{XY}Y^2$. The nonlinear term is the one that gives the relationship between *Z* and *Y* the characteristic shape of a quadratic relationship. The moderation component is represented in the equation by one single term: $m_{ZYX}p_{X\varepsilon}r_{\varepsilon X}XY$.

The term including the product *XY* here clearly has a decreased association coefficient $m_{ZYX}p_{X\varepsilon}r_{\varepsilon X}$ when compared with the association coefficient m_{ZYX} that was present when the quadratic association was not made explicit. The decrease is proportional to the fraction $p_{X\varepsilon}r_{\varepsilon X}$, which is the contribution to the explained variance in *X* by the error variable ε. This contribution is, in turn, inversely related to the correlation between *X* and *Y*; if the correlation between *X* and *Y* goes up, the contribution to the explained variance in *X* by the error variable ε goes down. That is, the greater the correlation between *X* and *Y*, the smaller is the moderation component and the greater is the quadratic component. If this correlation becomes 1, a situation in which *X* and *Y* would be redundant (i.e., *Y* would moderate itself), the relationship between *Z* and *Y* would become a purely nonlinear relationship, where the moderation relationship component would no longer exist:

$$Z = p_{ZY}Y + m_{ZYX}Y^2 + p_{Z\lambda}\lambda.$$

As it will be seen in the illustrative analyses, if the quadratic component is not explicitly modeled, the moderation component retains its original association strength, defined by the association coefficient m_{ZYX}. Nevertheless, the quadratic component is real and remains, even though it does not show up in the statistical results. The quadratic component can in fact be observed through visual inspection of a plot of the relationship between *Z* and *Y* if the correlation between *X* and *Y* is not negligible. This suggests a duality of moderation-nonlinear relationships: depending on how we look at them, they incorporate elements that suggest moderation (cXY) and/or nonlinearity ($aY + bY^2$).

If the quadratic component is explicitly modeled through the inclusion of a nonlinear association coefficient that reflects the fit of the quadratic component with the data, through a quadratic function, the moderation component will lose its strength. As it will be demonstrated later in the illustrative analyses, this loss in strength can cause a moderation association coefficient that was originally statistically significant to lose significance. This does not mean that the moderation no longer exists. It does, but does not show up in statistical results. Again, this characterizes the duality of moderation-nonlinear relationships.

To explicitly model the quadratic relationship in a way that is analogous to what is done for linear relationships in path models we need to replace the quadratic component with the product of a nonlinear coefficient of association n_{ZY} and a quadratic function $F_J(Y)$ that expresses the bivariate nonlinear relationship between *Z* and *Y*. Here the subscript for *F* is denoted by *J* because quadratic relationships are frequently referred to as J-curve relationships (Bremmer, 2007). This takes us to (7):

$$F_J(Y) = E(Z|Y) = aY + bY^2 \rightarrow$$
$$Z = n_{ZY}F_J(Y) + \dot{m_{ZYX}}XY + p_{Z\zeta}\zeta. \quad (7)$$

The coefficient $\dot{m_{ZYX}}$ is represented with a dot accent to differentiate it from the original moderation relationship coefficient without the presence of the nonlinear relationship term (i.e., m_{ZYX}), because under Simpson's paradox conditions our mathematical discussion suggests that $\dot{m_{ZYX}} \ll m_{ZYX}$.

The quadratic function $F_J(Y)$, which expresses the bivariate nonlinear relationship between *Z* and *Y*, can be obtained through a standard quadratic polynomial interpolation (Dodgson, 1997; Franke, 1982). Once this is done, the coefficient n_{ZY}, as well as the other coefficients in the equation, can be obtained through a nonlinear path analysis – essentially a linear path analysis with the values of $E(Z|Y)$ calculated based on $F_J(Y)$ replacing F in the equation below.

$$Z = n_{ZY}F + \dot{m_{ZYX}}XY + p_{Z\zeta}\zeta.$$

So far we have derived a number of important properties arising from Simpson's paradox through conceptual arguments and algebraic operations, but we have not demonstrated these properties based on data. In the next section we present an illustrative analysis that builds on our previous technology and management styles scenario. In this analysis we use true model data. That is, we use data that we created through a Monte Carlo simulation (Paxton et al., 2001) for which we know beforehand the causal directions of the relationships and the true values of the coefficients in the model.

An illustrative analysis with true model data

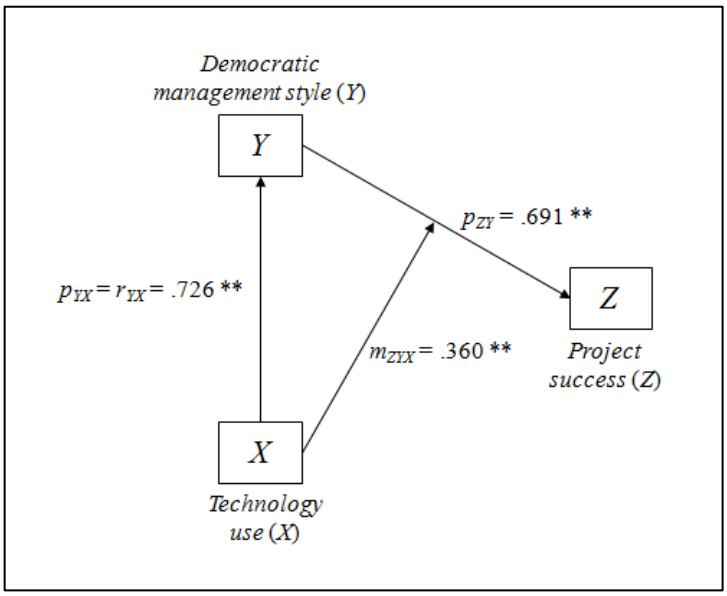
Figure 5 shows the path model including a constrained mediated moderation relationship with true coefficients used in this illustrative analysis, which is inspired by a study with real data discussed later. The data for this model is made up of 250 rows along three columns – for the variables *X*, *Y* and *Z*. It was created through a Monte Carlo simulation (Paxton et al., 2001), as outlined in Appendix E. As expected based on Fiedler et al.'s (2003) analyses and the mathematical discussion above, this model leads to the appearance of Simpson's paradox when incorrectly specified. No relationship between pairs of variables in this model was created to be nonlinear. That is, the only possible source of nonlinearity of the quadratic type in this model is the moderation relationship; this illustrates the duality of moderation and nonlinearity.

We used several tools to create and analyze the data. Featuring prominently among those tools are WarpPLS 4.0 and MATLAB 8.1. WarpPLS is a nonlinear path modeling and analysis tool that supports the use of models with and without latent variables (Kock, 2013). MATLAB is a numeric computing tool with extensive support for statistical analyses and generation of simulated data (Martinez & Martinez, 2008). At several points results were double-checked with other widely used statistical and numerical analysis software tools: IBM SPSS Statistics 22.0, Stata 13, PLS-Graph 3.0, and Microsoft Excel 2010.

We conducted full collinearity tests on both models (true model and incorrectly specified model), checking for both vertical and lateral collinearity (Kock & Lynn, 2012), to rule out multicollinearity as a possible source of bias. This was done in part to address Carte & Russel's (2003) warning that multicollinearity may artificially give rise to nonlinear relationships in

models including moderation. No evidence of multicollinearity was found; all variance inflation factors were lower than the threshold of 3.3 proposed by Kock & Lynn (2012).

Figure 5: Path model with true coefficients



Note: ** P < .01

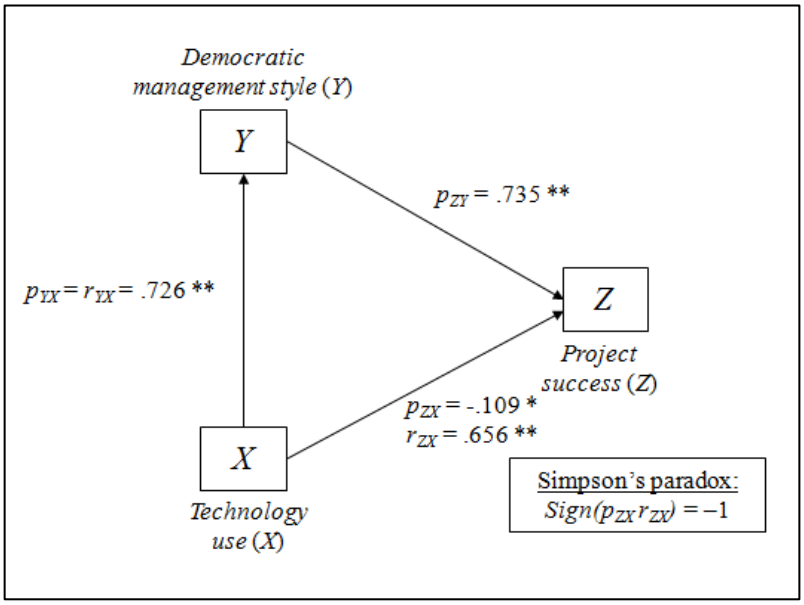
Figure 6 shows the incorrectly specified path model that corresponds to the true model. This model displays an instance of Simpson's paradox occurring in connection with the path going from *X* to *Z*. In an analysis of a model based on real data, this would normally be the first model to be generated if one expected *X* to causally influence *Y* and *Z*, and *Y* to causally influence *Z*. This model is equivalent to a multiple regression model with two independent variables *X* and *Y* pointing at a dependent variable *Z*, augmented by a link between *X* and *Y*; the use of similar models is very common across a number of fields of research, including the field of IS. In this illustration, we know based on our conceptual and mathematical discussion presented earlier, as well as the fact that we departed from a true model, that Simpson's paradox is caused by an incorrect specification of the model including a constrained mediated moderation relationship.

The absurdity of Simpson's paradox becomes clear when we consider the coefficient for the direct path from *X* to *Z* together with the coefficient for the total effect of *X* on *Z*. The total effect of *X* on *Z* is the sum of all of the direct and indirect effects via all paths connecting *X* to *Z*. There are two such paths, one with one segment (the direct path from *X* to *Z*), and the other with two segments (from *X* to *Y* and *Y* to *Z*). The coefficient for the direct path from *X* to *Z* is -.109 (P<.05) suggesting that technology use (*X*) has a negative effect on project success (*Z*). The coefficient for the total effect of *X* on *Z* is .425 (P<.01), suggesting that technology use (*X*) has a positive overall effect on project success (*Z*).

Figure 7 shows the model and its path coefficients when we explicitly model the quadratic relationship. As we predicted based on the mathematical development presented earlier, the moderation relationship coefficient decreased, to the point of losing statistical significance. Since the moderation relationship coefficient is no longer statistically significant, it would be reasonable to consider removing it from the model with the goal of arriving at a more parsimonious version of the model. Note that this model is different from, and arguably

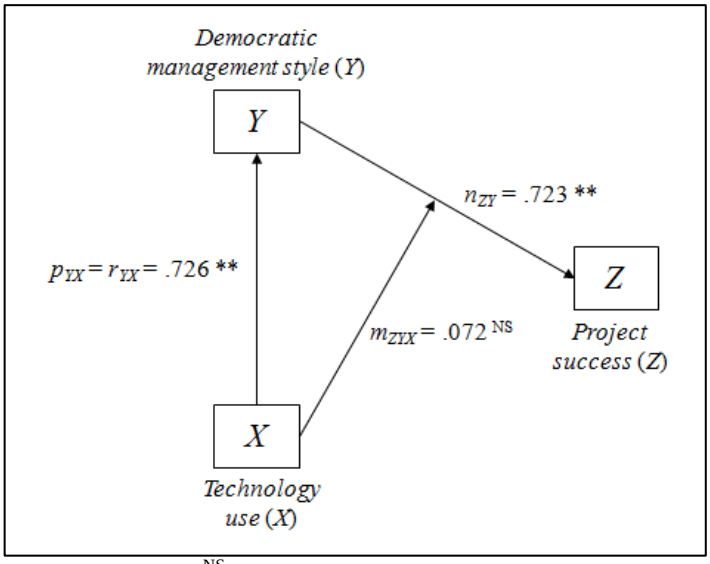
complementary to, the model where moderation is significant. Again, here we explicitly model the nonlinear (quadratic) relationship, which is why the moderation relationship coefficient becomes non-significant.

Figure 6: Incorrectly specified path model with Simpson's paradox



Notes: ** P < .01, * P < .05

Figure 7: Path model including moderation and nonlinear effects



Notes: ** P < .01, NS non-significant

It is important to use a different symbol for the nonlinear relationship coefficient (n_{ZY}) than the usual linear relationship coefficient symbol (e.g., p_{ZY}), as we do here, because these two types of coefficients have different meanings. These two coefficients are similar in that both are a measure of the degree to which a best-fitting function – nonlinear and linear, respectively –

conforms to a distribution of points. However, they differ significantly in their meaning with respect to the slopes of relationships among variables.

The linear relationship coefficient represents the standardized partial slope of the relationship between two variables (e.g., *Y* and *Z*), which is constant. For example, if $p_{ZY} = .3$ we can conclude that each standard deviation variation of *Y* is associated with a .3 standard deviation variation in *Z*, across all values of *Y*.

The nonlinear relationship coefficient calls for a different interpretation. In a nonlinear relationship such a slope, which would be the first partial derivative of *Z* with respect to *Y* ($\partial Z / \partial Y$), is not constant. In a quadratic relationship, which is a special case of nonlinear relationship, the slope would be a linear function of *Y*, and thus would not be the same across all values of *Y*.

Figure 8 shows a parsimonious version of the model, without the moderation relationship (which was no longer statistically significant), and where we continue explicitly modeling the quadratic relationship.

Figure 9 contains three plots side-by-side showing three relationships: the linear relationship between technology use (*X*) and democratic management style (*Y*), the moderation relationship between democratic management style (*Y*) and project success (*Z*), and the nonlinear relationship between democratic management style (*Y*) and project success (*Z*). The lines in the moderation relationship plot are the slopes of the relationship between *Y* and *Z* for low and high values of *X*; the data points were sorted by *X* and split in the middle to generate the two datasets of equal size for low and high values of *X*.

We can clearly see the moderation-nonlinear relationship duality when we look at the middle and right plots. Even if the relationship between *Y* and *Z* is not explicitly modeled as nonlinear, the quadratic shape that arises from the moderation relationship is clearly visible in the distribution of points. From the mathematical discussion we can assume that this is a true quadratic relationship, as opposed to a "fake" one arising entirely via capitalization on error.

When the relationship between *Y* and *Z* is explicitly modeled as nonlinear we gain additional insights about it. For example, the nonlinear relationship coefficient (n_{ZY}) gives us a more accurate estimation of the magnitude of the influence of *Y* on *Z*, which would be underestimated via linear modeling. Also, we can see that democratic management style (*Y*) influences project success (*Z*) positively only after *Y* achieves a certain level. Before that, the influence is in fact negative, which is an interesting and counterintuitive result – one that remains largely hidden from view in the analysis with the moderation relationship and without the quadratic relationship being modeled explicitly.

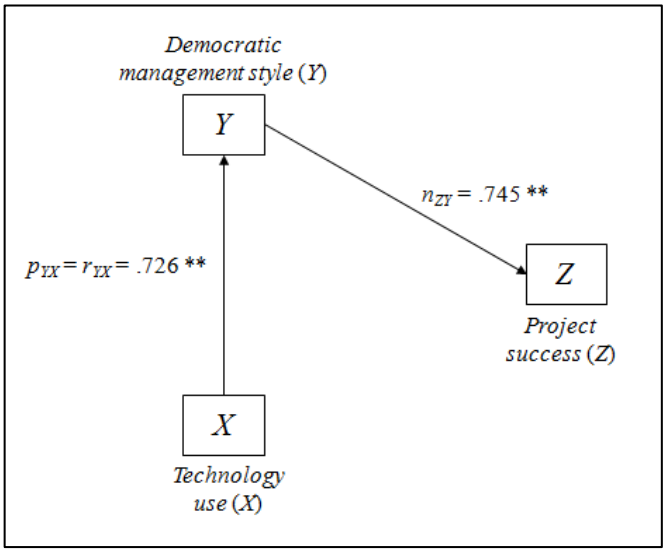
To complement this illustrative analysis based on true model data, we also provide in Appendix F a matched illustrative analysis based on real data. The model used in this analysis of real data depicts relationships among the following variables in a sample of developing countries: Internet diffusion, voice and accountability, and government corruption.

In the context of our discussion, the analysis of real data does not carry the same weight as the analysis of true model data. The reason is that we do not know the underlying model in the real data. Nevertheless, the analysis of real data illustrates the occurrence of Simpson's paradox in an empirical study, and the solution of Simpson's paradox through a model re-specification, both of which are fully compatible with the arguments presented here.

The illustrative analyses based on true model data and real data support Fiedler et al.'s (2003) conclusion, discussed earlier, that a three-variable instance of Simpson's paradox can be seen as a misspecification of a constrained correlated moderation relationship as a relationship between

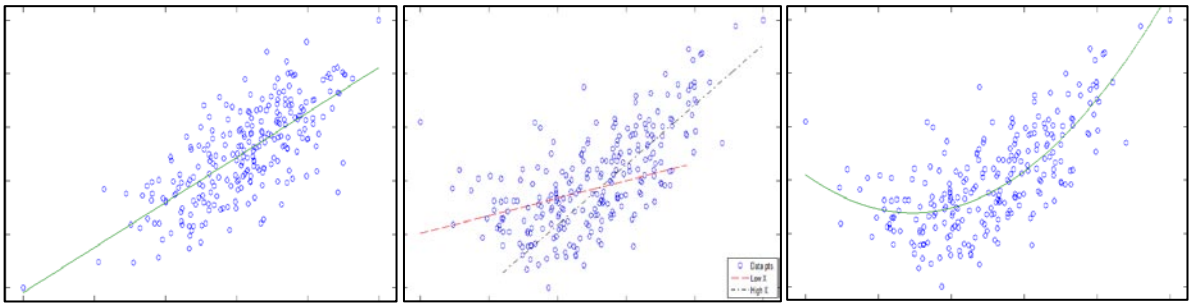
two predictors and a criterion variable. To further test this conclusion by Fiedler et al. (2003), we also created 500 additional datasets employing the same Monte Carlo simulation procedure used to create the true model data discussed in this section.

Figure 8: Parsimonious path model including nonlinear effect



Note: ** P < .01

Figure 9: Plots showing linear, moderation, and nonlinear relationships



Notes:
- Left: linear relationship between technology use (*X*) and democratic management style (*Y*)
- Middle: relationship between democratic management style (*Y*) and project success (*Z*), moderated by technology use (*X*)
- Right: nonlinear relationship between democratic management style (*Y*) and project success (*Z*), where the nonlinearity is due in part to technology use (*X*)

All 500 datasets led to models that contained one instance of Simpson's paradox when they were specified as including linear relationships between two predictors and a criterion variable as in the incorrectly specified model discussed above. When re-specified as models containing constrained correlated moderation relationships as in the correctly specified model discussed above, all models were free from Simpson's paradox. This general result applied to all 500 datasets, with no exception, and provides further decisive support for Fiedler et al.'s (2003) conclusion.

Discussion and recommendations

We have up to this point defined Simpson's paradox in precise mathematical terms, provided a solution to Simpson's paradox, and shown that it is associated with moderation-nonlinear relationships. In this section we synthesize all of this through a focused discussion of several recommendations for researchers.

Simpson's paradox and correct model specification

Our entire discussion so far suggests that it is important to always check for Simpson's paradox in linear path models (i.e., path models where nonlinear relationships are not explicitly modeled), as this phenomenon is likely to be associated with incorrect model specification and the existence of hidden moderation-nonlinear relationships. Simpson's paradox cannot be reconciled with the notion that *X* is a valid direct predictor of *Z* (Geisser, 1974; Pearl, 2009; Stone, 1974), suggesting an incorrect model specification. As such, Simpson's paradox can serve as a causality assessment tool, and perhaps many other purposes yet to be explored.

Checking for the presence of Simpson's paradox in linear path models can be done by looking for instances where one of the two following equivalent conditions is satisfied:

$$Sign(p_{ZX}r_{ZX}) = -1.$$
$$p_{ZX}r_{ZX} < 0.$$

As we can see, unlike identifying Simpson's paradox based on the inspection of contingency tables, it is fairly easy to check for the presence of Simpson's paradox in linear path models. Most multivariate statistical analysis tools that can be used to automate path analyses, with or without latent variables, also report both path coefficients and correlations. If only path coefficients are reported, correlations can be easily calculated with general-purpose statistical analysis tools.

If an instance of Simpson's paradox is found, it should be seen as possible evidence of incorrect model specification. Researchers should consider re-specifying the path that is associated with Simpson's paradox, from a direct path ($X \rightarrow Z$) to a path involving moderation [$X \rightarrow (Y \rightarrow Z)$], if theory supports this decision.

We qualify our recommendation above, as we do with other recommendations, by noting that model re-specification should be conducted "if theory supports" it. The role of theory is still important in the search of the true models that best represent empirical data. When we depart from empirical data, there is no way of knowing the nature of the true models with absolute certainly.

Concurrent moderation and direct relationships

The classical view of moderation relationships, which is in fact the most widely used and considered advisable, is that of a standard linear relationship between variables *X*, *Y* and *Z*, augmented by a moderation relationship between *X* and the link between *Y* and *Z* (Cohen et al., 2003). This view is expressed in the equation below.

$$Z = p_{ZY}Y + p_{ZX}X + m_{ZYX}XY + p_{Z\gamma}\gamma.$$

This classical view requires the inclusion of a direct relationship between *X* and *Z*, represented by the term $p_{ZX}X$, which our conceptual and mathematical discussions suggest to be absent from

a correctly specified model from which a Simpson's paradox instance has been removed. That is, our research suggests a different type of true moderation that differs from the classical view of moderation, which we refer to here as "constrained moderation" – i.e., a relationship that involves moderation [$X \rightarrow (Y \rightarrow Z)$] without a direct relationship ($X \rightarrow Z$). As we have shown, Simpson's paradox can be a marker of this different type of true moderation. This type of moderation characterizes conditional latent growth curve models (Meredith & Tisak, 1990; Simons-Morton & Chen, 2005).

Given this, we recommend that researchers carefully consider whether they should include or not a direct link between the moderator and the criterion variables when conducting a moderation relationship test, as the classical view of moderation would dictate. If the inclusion of a direct link between the moderator and the criterion variables leads to Simpson's paradox in a classic moderation relationship test, this should be seen as evidence that the direct link may not exist, even though the moderation relationship link may still exist. Such direct link should be included only if theory supports it, and thus explicitly tested separately from the main test of moderation.

Our recommendation above differs somewhat from that in Carte & Russel's (2003) seminal article on moderation. There the authors suggest that if both moderation and direct relationships involving the same moderation variable are statistically significant, a type I error (false positive) would be committed, as no conclusions could possibly be drawn regarding the significance of the direct relationship in that context. This is in direct contrast with Cohen et al.'s (2003) perspective, which is that moderation and direct relationships should always be tested together. We take the middle ground here. Our recommendation incorporates the belief that moderation and direct relationships can be tested together, as long as there is theoretical support for both and the model reflecting the theory is free from Simpson's paradox.

Detecting true quadratic relationships

We have demonstrated that relationships that can be modeled through functions of the type $F_J(Y) = E(Z|Y) = aY + bY^2$ arise from: (a) constrained moderation, or moderation [$X \rightarrow (Y \rightarrow Z)$] without a direct relationship ($X \rightarrow Z$), in correctly specified models; and (b) Simpson's paradox, in incorrectly specified models. It is important to note, however, that if the constant b in one such function is below a certain threshold (for example, below .01) the function will take a shape that will be more similar to that of a line than a curve. The reason for this is that b reflects the magnitude of the squared term in the function (i.e., bY^2). This is the term that gives the function a curved, or quadratic, shape. That is, while the relationship may be actually quadratic, it will not appear to be quadratic upon visual inspection, nor will it clearly behave as quadratic from a practical perspective.

One of the main problems of exploratory analyses where quadratic relationships are explicitly modeled is that "fake" quadratic relationships may arise from error. The greater the amount of error involved in a relationship, the greater is also the likelihood that a quadratic relationship will arise by chance. As we have shown in our mathematical analyses, true quadratic relationships can arise from: (a) constrained moderation, or moderation [$X \rightarrow (Y \rightarrow Z)$] without a direct relationship ($X \rightarrow Z$), in correctly specified models; and (b) Simpson's paradox, in incorrectly specified models. This is chiefly demonstrated through our mathematical analyses, and consistently illustrated through the illustrative analysis of true model data. From this we can conclude that both constrained moderation and Simpson's paradox can serve as markers of true quadratic relationships.

Therefore it is recommended that, before explicitly modeling a path ($Y \rightarrow Z$) as being of the quadratic relationship type, researchers should find out if the path is associated with significant constrained moderation [$X \rightarrow (Y \rightarrow Z)$] or gives rise to Simpson's paradox when included in a non-moderated linear model [$(X \rightarrow Z)$ & $(Y \rightarrow Z)$]. Both are indications that the path ($Y \rightarrow Z$) may refer to a true quadratic relationship. These tests should be complemented by a search for theoretical reasons why the relationship should be modeled as quadratic.

The lower the path coefficient associated with a relationship that is modeled as linear, the greater is the amount of error involved in the relationship. Therefore, when faced with non-significant linear paths, researchers should be particularly cautious about explicitly modeling such relationships as quadratic to obtain significant nonlinear path coefficients without first trying to detect true quadratic relationships and search for theoretical reasons why the relationships should be modeled as quadratic.

While significant constrained moderation and Simpson's paradox are markers of possible true quadratic relationships, our research does not allow us to conclude that they are the only markers of such relationships. Particularly noteworthy is the possibility that moderation relationships that are not constrained may lead to quadratic relationships, in which case they may not be associated with Simpson's paradox. However, our mathematical analysis suggests that the squared component of quadratic relationships is proportional to m_{ZYX}, which would normally go down when moderation relationships are not constrained, thus making the curve less "J-like" and more like a line.

Moderation-nonlinear relationship duality

Empirical researchers have been repeatedly warned about the possibility of mistakenly interpreting moderation as nonlinearity and vice-versa (Cohen, 1978), with suggestions that such mistakes are likely to be a byproduct of multicollinearity (Carte & Russel, 2003). Our analyses, however, suggest not only that constrained moderation and quadratic relationships may co-exist, but also that this may be the case in models that are free from multicollinearity. We conducted full collinearity tests as part of our illustrative analyses, and found no evidence of multicollinearity; all variance inflation factors were lower than the threshold of 3.3, which is arguably a conservative threshold in path models without latent variables (Kock & Lynn, 2012).

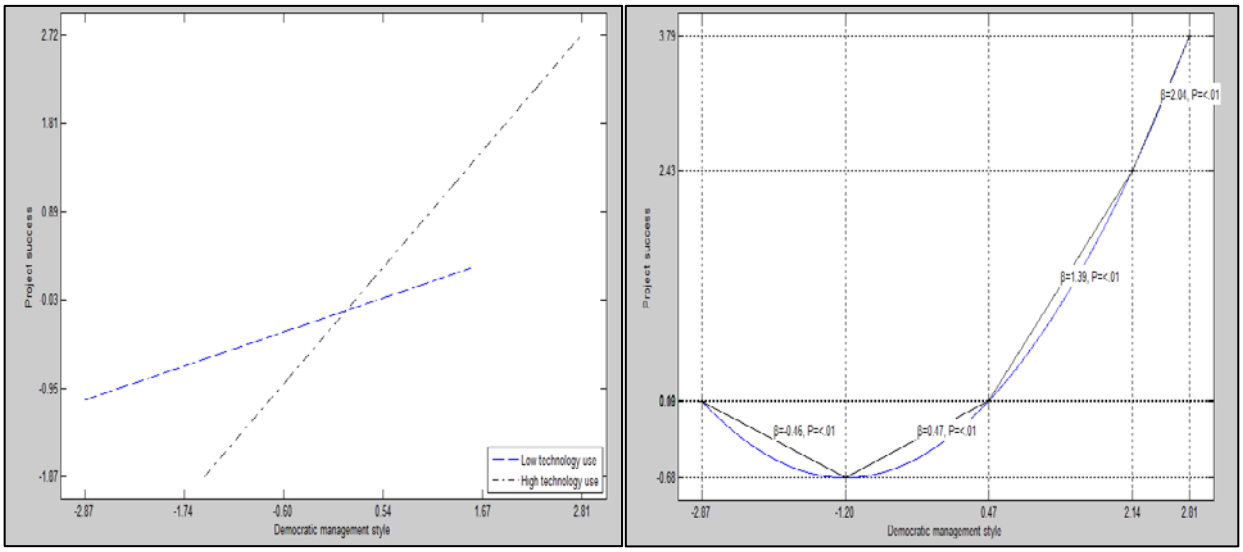
Our conceptual-mathematical discussions and illustrative analysis of true model data suggest that the direct path ($X \rightarrow Z$) did not exist, and that *X* influenced *Z* via *Y*. From a moderation perspective, *X* influenced *Z* via *Y* through the product *XY*. From a quadratic perspective, *X* also influenced *Z* via *Y*, with the relationship between *Y* and *Z* being of the quadratic type and mediating the broader relationship between *X* and *Z*. In both cases, *X* per se cannot influence *Z* without *Y*, even though *Y* can influence *Z* without *X*. From this we can infer that, when moderation and quadratic relationships co-exist, they lead to consistent interpretations.

While leading to consistent interpretations, co-existing moderation and quadratic relationships lead to *complementary* insights for research and practice, as our discussion and analyses have demonstrated. Going back to our illustrative analysis of true model data, the path model with a significant constrained moderation relationship tells us that the effect of democratic management style (*Y*) on project success (*Z*) is significantly stronger at high levels of technology use (*X*). The path model where the quadratic relationship is explicitly modeled ("robbing" the moderation relationship from its strength, so to speak) tells us that the effect of democratic management style (*Y*) on project success (*Z*) conforms to a quadratic pattern and is only positive above a certain level.

Given the complementary nature of these insights, we contend that models with Simpson's paradox should be revised by researchers, with the revision leading to separate models that include constrained moderation and quadratic relationships, so that the complementary insights provided by each model can be used in the development of recommendations for other researchers and practitioners.

The complementary nature of the insights provided by models that include constrained moderation and quadratic relationships is illustrated in Figure 10. The graph on the left shows the relationship between democratic management style (Y) and project success (Z), where the difference in the slopes of the lines is due to moderation of the path $Y \rightarrow Z$ by technology use (X). The graph on the right shows the quadratic relationship between democratic management style (Y) and project success (Z). The scales are standardized.

Figure 10: Plots showing moderation and nonlinear relationships



Notes:
- The scales are standardized.
- Left: relationship between democratic management style (Y) and project success (Z), moderated by technology use (X)
- Right: nonlinear relationship between democratic management style (Y) and project success (Z), where the nonlinearity is due in part to technology use (X)

The graphs on the left and right are similar to graphs shown earlier, but nevertheless depict different ways through which the relationships in the illustrative analysis with true model data can be represented. Both graphs are "focused" graphs (without data points) and display standardized scales. The quadratic graph on the right is segmented; the coefficients shown are approximations of the first partial derivatives of Z with respect to Y ($\partial Z / \partial Y$) for each of the segments. The absolute effect segmentation "delta" is .93. That is, in the quadratic graph on the right each segment has a slope ($\partial Z / \partial Y$) that is approximately .93 plus the previous segment's slope.

Based on the combined insights provided by these graphs, a researcher could make important and relatively precise recommendations to practitioners interested in maximizing project success (Z) at a reasonable cost. The "reasonable cost" constraint means that a generic recommendation such as "maximize technology use (X) and democratic management style (Y)" may not be as

useful as a more specific one such as "increase technology use (*X*) beyond level L_X or democratic management style (*Y*) beyond level L_Y".

What could these levels L_X and L_Y be? Based on the graph on the right, we could say that $L_Y = .47$, beyond which point the standardized association between *Y* and *Z* becomes very strong (equal to or greater than 1.39, rarely seen in linear models in the absence of multicollinearity). Consequently $L_X = .65$, calculated as $.47/.726$, since the relationship between *X* and *Y* is linear and $p_{YX} = r_{YX} = .726$.

Interestingly, the quadratic relationship (graph on the right) provides us with more precise information when it comes to estimating these levels. Based on the graph on the left, where different slopes refer to two sub-samples obtained from splitting the original sample into two sub-samples of the same size, we would only be able to come up with the rough estimation that $L_X = \tilde{X}$, where $\tilde{X}$ is the median for X. The two-dimensional representation of the moderation relationship is intuitively appealing but inherently limited.

The usefulness of these more specific recommendations is based on the reasonable assumptions that the standardized measures provided would be unstandardized, by multiplication by their standard deviations and addition of their means, and that these values would in turn be further translated into actual organizational decisions.

Note that the levels of *X* and *Y* at which project success (*Z*) is minimized are approximately $L_Y = -1.2$ and $L_X = -1.653$, with the latter calculated as $-1.2/.726$. These are not the minimum values of *X* and *Y* in the sample. The values of the technology use (*X*) variable range from -4.09 to 2.48. For the democratic management style (*Y*) variable, the values range from -2.87 to 2.81. In an empirical study with real data, this would suggest that, at fairly low levels of technology use, being autocratic is not necessarily a bad idea. However, after technology use reaches a certain level, being democratic pays off.

These results are consistent with those in Kock's (2004) action research investigation of the effects of an asynchronous collaboration technology on the outcomes of business process improvement projects; an investigation which, as mentioned earlier, served as the inspiration for the technology and management styles scenario for which true model data was created.

Expressing moderation and nonlinear relationship coefficients

There are fundamental differences in meaning among linear, moderation, and quadratic relationship coefficients. If a linear relationship coefficient assumes a certain value, such as $p_{ZY} = .3$, we can conclude that each standard deviation variation of *Y* is associated with a .3 standard deviation variation in *Z*, across all values of *Y*. If a moderation relationship coefficient assumes the same value, $m_{ZYX} = .3$, the meaning is quite different. Here we can conclude that each standard deviation variation of *X* is associated with a .3 increment in the standardized coefficient for the relationship between *Y* and *Z*. For instance, if $p_{ZY} = .4$ this coefficient will go up to $p_{ZY} = .7$ for a one standard deviation variation in *X* if $m_{ZYX} = .3$. Finally, if a quadratic relationship coefficient assumes the same value, $n_{ZY} = .3$, the meaning is entirely different, because in a quadratic relationship the slope (first partial derivative of *Z* with respect to *Y*, or $\partial Z/\partial Y$), is not constant. This applies to all nonlinear relationships, not only quadratic relationships.

The different natures of these coefficients lead us to make the following recommendation. To avoid confusion when reporting research results, researchers should use different symbols to refer to coefficients for linear, moderation, and nonlinear relationships. The following symbols are recommended: linear – p_{ZY}, moderation – m_{ZYX}, and nonlinear – n_{ZY}.

In practice most researchers do not use different symbols to refer to linear, moderation and nonlinear relationship coefficients (Guo et al., 2011; Schmiedel et al., 2014; Sykes et al., 2014). Usually the same symbol is used, with a common choice among IS researchers being the symbol β. This can be confusing to readers, who at first may interpret a nonlinear coefficient, shown with the same symbol as a linear coefficient, as referring to a constant slope – until they see a graph of the relationship where the slope clearly changes.

Even though these three types of coefficients are significantly different in their meanings, they share two key commonalities in that they reflect: (a) the strengths of linear, moderation, and nonlinear relationships; and (b) the extent to which the data points conform to best-fitting lines, three-dimensional surfaces, and curves, respectively.

Simpson's paradox in more complex models

Our conceptual-mathematical discussions and illustrative analysis with true model data addressed models with only three variables, with the configuration of two predictor variables *X* and *Y* and one criterion *Z* leading to Simpson's paradox under certain conditions. Nevertheless, our recommendations extend to more complex path models. These models can always be decomposed into sub-models where two or more predictors point at one criterion, which can then be analyzed separately.

Let us assume that we have a model where three variables *X*, Y_1, and Y_2 point at *Z*. Let us also assume that the path $X \rightarrow Z$ leads to Simpson's paradox. Which of the variables, Y_1 or Y_2, will likely be involved in the constrained moderation and nonlinear relationships, after the model is correctly specified? Based on our analysis, it is clear that the answer is the variable whose correlation with *X* is the strongest, since the strength of both the constrained moderation and nonlinear relationships are a function of that correlation.

This is demonstrated in the illustrative analysis with real data in Appendix F. There 8 variables point at one main criterion variable. Of those 8 variables, 2 are actual structural predictors and 6 are included in the model as control variables. Among the 8 variables, only 1 variable (*I*) led to Simpson's paradox, together with the other actual predictor variable (*V*), which was much more strongly correlated with *I* than with any of the control variables.

The above can also be generalized to situations in which two or more instances of Simpson's paradox occur, where a similar procedure can be employed. Let us assume that the following variables point at *Z* in a sub-model: X_1, X_2, Y_1, Y_2, Y_3 and Y_4. Let us also assume that both paths $X_1 \rightarrow Z$ and $X_2 \rightarrow Z$ lead to Simpson's paradox. Which of the variables – Y_1, Y_2, Y_3 or Y_4 – will be involved in the constrained moderation and nonlinear relationships, after the model is correctly specified? Based on a logic similar to that employed above, the answers are the variables whose correlations with X_1 and X_2 are individually the strongest. These can be easily obtained from a table of correlations among all variables in the model. As discussed earlier, the extent to which nonlinearity will be visibly observable will depend on the strengths of these individual correlations.

Simpson's paradox instances may be associated with higher-order moderation and nonlinear relationships in the special case where multiple variables X_1, X_2 … X_n are involved in 1st, 2nd … nth order moderations of a single link $Y \rightarrow Z$. In this special case we could have equations with terms including the following products and power terms: X_1Y, X_1X_2Y … $(X_1X_2 \ldots X_n)Y$ and Y^2, Y^3 … Y^n. A thorough analysis of this special case and its ramifications is beyond the scope of this paper, and is recommended as future research.

Conclusion

The Simpson's paradox phenomenon leads to conclusions that are puzzling. Perhaps because of this, it has been largely neglected in many applied fields of research, including the field of IS. Puzzling conclusions arising from statistical analyses, such as those in connection with Simpson's paradox, are often ignored or explained away based on bias from measurement error and multicollinearity (Pearl, 2009). We have shown here that the neglect of Simpson's paradox in applied research, beyond discussions of it as a statistical oddity, may have been an unfortunate error of omission. This is so not only because Simpson's paradox can often be solved within the well developed and familiar domain of path analysis, but also because its solution opens the door for the understanding of other complex research issues.

When stated in mathematical terms, Simpson's paradox is clearly associated with relationships that have been the target of increasing interest in the field of IS. Incorrectly specified linear path models, when revised and correctly specified, give rise to moderation and quadratic relationships. Moderation is a methodological topic that has been the target of much interest in the field of IS, reflecting the fact that context is of great importance in IS investigations (Carte & Russel, 2003; Sykes et al., 2014). Quadratic relationships seem to be quite common (Bremmer, 2007), and have, together with other nonlinear relationships, been recently receiving increasing attention in the field of IS (Guo et al., 2011; Schmiedel et al., 2014).

To the best of our knowledge, this is the first time Simpson's paradox is both clearly stated in terms of linear path models and explicitly defined mathematically in terms of the signs of path coefficients and corresponding correlations. This also seems to be the first time a solution to Simpson's paradox is explicitly stated in terms of path model re-specification with inclusion of constrained moderation relationships. Finally, another original contribution of this research has been to show that moderation and quadratic relationships co-exist and show different facets of the same overall relationship depending on how we look at the data – i.e., how we analyze it.

While our discussions and illustrative analyses focus on path models, where variables are measured through single indicators, the related conclusions and recommendations apply to the more general case of structural equation models (Maruyama, 1998; Schumacker & Lomax, 2004). Mathematically, structural equation models can be seen as path models in which structural variables are latent, and thus measured indirectly through more than one indicator (Mueller, 1996). Our conclusions and recommendations also apply to multiple regression models, which can be seen as special cases of path models.

Perhaps one of the main contributions of this research is in connection with the possible detection of true quadratic relationships. While these relationships seem to be fairly common and have the potential to explain important phenomena (Bremmer, 2007), exploratory analyses where they are explicitly modeled can lead to the chance emergence of "fake" quadratic relationships, as error is modeled as though it was a component of an underlying quadratic relationship. Our research leads to an unexpected finding. What seemed to be an annoying problem, namely Simpson's paradox, may in fact be the key to the identification of true quadratic relationships.

References

Anderson, J.C., Rungtusanatham, M., Schroeder, R.G., & Devaraj, S. (1995). A path analytic model of a theory of quality management underlying the Deming management method: Preliminary empirical findings. *Decision Sciences*, 26(5), 637-658.

Bremmer, I. (2007). *The J curve: A new way to understand why nations rise and fall*. New York, NY: Simon & Schuster.
Carte, T.A., & Russel, C.J. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quarterly*, 27(3), 479-501.
Cohen, J. (1978). Partialed products are interactions; partialed powers are curve components. *Psychological Bulletin*, 85(4), 858.
Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, N.J.: L. Erlbaum Associates.
Conger, A.J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34(1), 35-46.
Dodgson, N.A. (1997). Quadratic interpolation for image resampling. *IEEE Transactions on Image Processing*, 6(9), 1322-1326.
Duncan, O.D. (1966). Path analysis: Sociological examples. *The American Journal of Sociology*, 72(1), 1-16.
Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin*, 29(1), 14-27.
Franke, R. (1982). Scattered data interpolation: Tests of some methods. *Mathematics of Computation*, 38(157), 181-200.
Geisser, S. (1974). A predictive approach to the random effects model. *Biometrika*, 61(1), 101-107.
Guo, K.H., Yuan, Y., Archer, N.P., & Connelly, C.E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, 28(2), 203-236.
Harbrecht, H., Peters, M., & Schneider, R. (2012). On the low-rank approximation by the pivoted Cholesky decomposition. *Applied Numerical Mathematics*, 62(4), 428-440.
Hofstede, G. (1983). The cultural relativity of organizational practices and theories. *Journal of International Business Studies*, 14(2), 75-90.
Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: Sage Publications.
Julious, S.A., & Mullee, M.A. (1994). Confounding and Simpson's paradox. *BMJ*, 309(6967), 1480-1481. [Note: This journal was previously British Medical Journal; its name was officially shortened to BMJ in 1988.]
Kenny, D.A. (1979). *Correlation and causation*. New York, NY: John Wiley & Sons.
Kline, R.B. (1998). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press.
Kock, N. (2004). The three threats of action research: A discussion of methodological antidotes in the context of an information systems study. *Decision Support Systems*, 37(2), 265-286.
Kock, N. (2013). *WarpPLS 4.0 User Manual*. Laredo, TX: ScriptWarp Systems.
Kock, N. (2014). Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of e-Collaboration*, 10(3), 1-13.
Kock, N., & Lynn, G.S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.

Kock, N., & Mayfield, M. (2015). PLS-based SEM algorithms: The good neighbor assumption, collinearity, and nonlinearity. *Information Management and Business Review*, 7(2), 113-130.
MacKinnon, D.P., Krull, J.L., & Lockwood, C.M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4), 173-181.
Martinez, W.L., & Martinez, A.R. (2008). *Computational statistics handbook with MATLAB*. Boca Raton, FL: CRC Press.
Maruyama, G.M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage Publications.
Maassen, G.H., & Bakker, A.B. (2001). Suppressor variables in path models definitions and interpretations. *Sociological Methods & Research*, 30(2), 241-270.
Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55(1), 107-122.
Mueller, R.O. (1996). *Basic principles of structural equation modeling*. New York, NY: Springer.
Pavlides, M.G., & Perlman, M.D. (2009). How Likely Is Simpson's Paradox? *The American Statistician*, 63(3), 226-233.
Paxton, P., Curran, P.J., Bollen, K.A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8(2), 287-312.
Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, 26(4), 332-344.
Reinartz, W.J., Echambadi, R., & Chin, W.W. (2002). Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behavioral Research*, 37(2), 227-244.
Robert, C.P., & Casella, G. (2005). *Monte Carlo statistical methods*. New York, NY: Springer.
Schmiedel, T., vom Brocke, J., & Recker, J. (2014). Development and validation of an instrument to measure organizational cultures' support of business process management. *Information & Management*, 51(1), 43-56.
Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.
Simons-Morton, B., & Chen, R. (2005). Latent growth curve analyses of parent influences on drinking progression among early adolescents. *Journal of Studies on Alcohol*, 66(1), 5-13.
Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causality, prediction and search*. Berlin, Germany: Springer-Verlag.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36(1), 111–147.
Sykes, T.A., Venkatesh, V., & Johnson, J.L. (2014). Enterprise system implementation and employee job performance: Understanding the role of advice networks. *MIS Quarterly*, 38(1), 51-72.
Tu, Y.-K., Gunnell, D.J., & Gilthorpe, M.S. (2008). Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon - the reversal paradox. *Emerging Themes in Epidemiology*, 5(2), 1-9.
Wagner, C.H. (1982). Simpson's paradox in real life. *The American Statistician*, 36(1), 46–48.

Westfall, P., Hoffman, J.J., & Xia, J. (2007). Joint analysis of multiple categorical dependent variables in organizational research. *Organizational Research Methods*, 10(4), 673-688.
Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3), 161-215.
Wright, S. (1960). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16(2), 189-202.

Appendix A: Medical illustration of Simpson's paradox

Simpson's paradox has received particular attention in the context of medical investigations because it can lead to wrong decisions regarding medical interventions. These can in turn result in serious health consequences, including death. Julious & Mullee (1994) provided an overview of Simpson's paradox in the context of several studies, including a classic study of actual cases where two kidney stone treatments were compared.

One treatment (A) involved open surgery, while the other treatment was less invasive and required only a small puncture (B). Table A.1 shows the success rates of these treatments, across 6 cells, when different stone sizes were considered. The numbers in the cells are the: (a) success rates, given in percentages, for each pair of treatment and stone size condition; and (b) the numbers of individuals in each pair of treatment and stone size condition for which the treatment was successful, divided by the total number of individuals in the condition.

Table A.1: Medical illustration of Simpson's paradox

	Treatment A	**Treatment B**
Small Stones	***Cell 1*** **93%** (81/87)	***Cell 2*** 87% (234/270)
Large Stones	***Cell 3*** **73%** (192/263)	***Cell 4*** 69% (55/80)
Both stone sizes	***Cell 5*** 78% (273/350)	***Cell 6*** **83%** (289/350)

Note: Percentages reflect treatment success rates

According to the results of this kidney stone treatment study, treatment A is associated with greater success rates for both small and large stones, when those stone sizes are considered individually – cell pairs (1, 2) and (3,4). But when both stone sizes are considered together, in cell pair (5, 6), treatment B is associated with greater success. This leads to a paradoxical conclusion. If a doctor knows the size of the stone, he or she should recommend treatment A. If not, treatment B should be recommended. In other words, knowledge or ignorance about stone size defines the optimal treatment – not the treatment itself!

Appendix B: How common is Simpson's paradox?

We know from the mathematics of path analysis (Mueller, 1996; Wright, 1934; 1960) that the path coefficient p_{ZX} of the incorrectly specified model shown earlier where *X* and *Y* point at *Z* in a strictly linear model can be described in terms of the correlations among *X*, *Y* and *Z* through (B.1).

$$p_{ZX} = \frac{r_{ZX} - r_{ZY}r_{XY}}{1 - {r_{XY}}^2}. \quad \text{(B.1)}$$

Since the occurrence of Simpson's paradox is defined by the inequality $p_{ZX}r_{ZX} < 0$, we are interested in obtaining an equation that expresses $p_{ZX}r_{ZX}$ in terms of the correlations among *X*, *Y* and *Z*, so that we can create an equivalent inequality with this equation. This can be done by multiplying both sides of (B.1) by r_{ZX}:

$$p_{ZX}r_{ZX} = r_{ZX}\frac{r_{ZX} - r_{ZY}r_{XY}}{1 - {r_{XY}}^2}.$$

The Simpson's paradox inequality $p_{ZX}r_{ZX} < 0$ can then be expressed through (B.2). This equation sets the condition for the occurrence of Simpson's paradox in terms of the correlations among *X*, *Y* and *Z*.

$$\frac{{r_{ZX}}^2 - r_{ZX}r_{ZY}r_{XY}}{1 - {r_{XY}}^2} < 0. \quad \text{(B.2)}$$

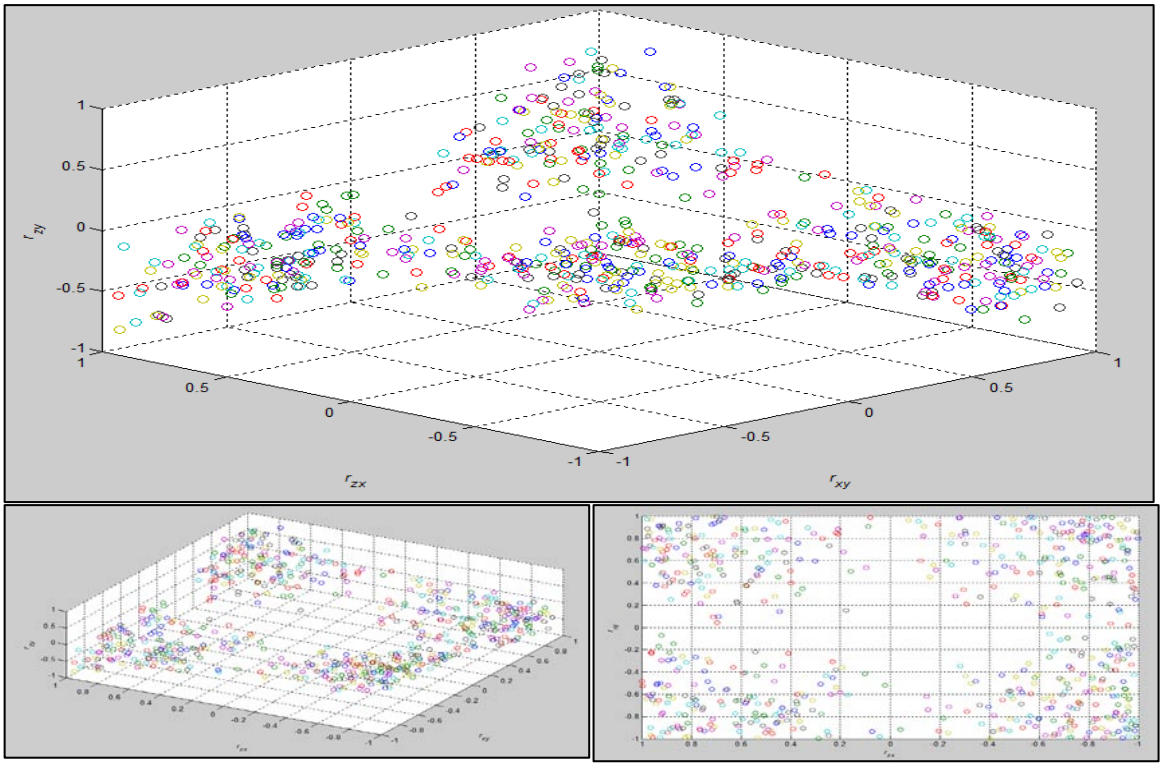
To assess how likely it is that Simpson's paradox will be observed in quantitative analyses involving at least three variables, we generated multiple 5,000 combinations of random values of correlations in linear models where *X* and *Y* point at *Z* (i.e., values of r_{ZX}, r_{ZY} and r_{XY}) through Monte Carlo simulations (Robert & Casella, 2005). We repeated this process 1,000 times (for a total of 5,000,000 combinations), averaging results based on (B.2) and the values of the correlations.

Our simulations yielded the probability of .128 that Simpson's paradox will be observed. That is, in quantitative analyses involving relationships among three or more variables, one can expect that Simpson's paradox will occur 12.8 percent of the time or more. This is consistent with the frequencies we found in our reviews of articles in IS journals, discussed earlier.

Figure B.1 shows the distribution of approximately 640 combinations (out of 5,000) of values of correlations that lead to occurrences of Simpson's paradox. The plots show the distribution of points from different angles. The two plots at the bottom show the plot at the top viewed from the perspective of someone moving to the left and up. In the bottom-right plot the observer looks from above.

As can be seen from the plots, the probability that Simpson's paradox will occur tends to go up for combinations of high values of correlations, but can also occur at relatively low correlations. The probability goes down and approaches zero as correlations approach zero. All correlations approach zero at the center of the plots.

Figure B.1: Values of correlations that lead to Simpson's paradox (three views from different angles)



Appendix C: Simpson's paradox versus statistical suppression

Simpson's paradox and statistical suppression are different phenomena that are often equated with each other (Maassen & Bakker, 2001; MacKinnon et al., 2000; Spirtes et al., 1993, Tu et al., 2008). This is in part due to the number of different definitions of statistical suppression that have been proposed over the years (Cohen et al., 2003; MacKinnon et al., 2000). In this appendix we focus on a classic definition of statistical suppression to clarify the difference between it and Simpson's paradox.

As pointed out by MacKinnon et al. (2000), the most widely accepted definition of a statistical suppressor has been provided by Conger (1974, pp. 36–37): "… a variable which increases the predictive validity of another variable (or set of variables) by its inclusion in a regression equation". Maassen & Bakker (2001) provide an extensive discussion of the various definitions of statistical suppression that have been proposed over the years; in the context of this discussion, Conger's (1974) definition refers to classic statistical suppression.

The term "predictive validity" in Conger's (1974) definition above refers to the strength of the association between a predictor variable and the criterion variable in a regression equation, after its adjustment to account for the inclusion of other correlated predictors. Typically the predictive validity of a predictor decreases with the inclusion of a correlated predictor. The baseline is the predictor-criterion correlation, or the unadjusted coefficient of association between any predictor and the criterion in a regression equation. In the case of classic statistical suppression, however, the predictive validity of a predictor *increases* with the inclusion of a correlated predictor (Conger, 1974; MacKinnon et al., 2000).

Classic statistical suppression, in the context of a path model where two predictor variables *X* and *Y* point at a variable *Z*, would be characterized, based on Conger's (1974) definition, by an instance in which the absolute value of a path coefficient is greater than that of the corresponding correlation. Mathematically, this would be expressed as follows for the path going from *X* to *Z*:

$$|p_{ZX}| > |r_{ZX}|.$$

This mathematical definition of classic statistical suppression is clearly different from our mathematical definition of Simpson's paradox with respect to the same path going from *X* and *Z*:

$$p_{ZX} r_{ZX} < 0.$$

For example, if $p_{ZX} = .3$ and $r_{ZX} = .2$ we would have an instance of classic statistical suppression but not of Simpson's paradox, with respect to the path going from *X* and *Z*. In path models, Simpson's paradox is characterized by the peculiar situation in which a path coefficient and a correlation have different signs.

Why do some researchers, such as MacKinnon et al. (2000) and Tu et al. (2008), argue that Simpson's paradox and classic statistical suppression are either the same phenomenon or closely related phenomena?

The reason is that often the occurrence of Simpson's paradox with respect to one path, such as the path $X \rightarrow Z$, will cause the occurrence of classic statistical suppression with respect to a competing path $Y \rightarrow Z$. This is a direct consequence of a well known mathematical property of path models, which is that the set of variables that directly cause a variable Z, plus the error term associated with Z, explain all of the variance in Z (i.e., 100 percent of the variance explained, or 1). In a path model with two predictor variables *X* and *Y* pointing a one criterion variable *Z*, this

property would be summarized though the equation below. In this equation, ε represents the error term associated with Z.

$$1 = p_{ZX}r_{ZX} + p_{ZY}r_{ZY} + p_{Z\varepsilon}r_{Z\varepsilon}.$$

We can see from this equation that, if an instance of Simpson's paradox occurs in connection with the path $X \rightarrow Z$ then the product $p_{ZX}r_{ZX}$ will be negative, thus likely leading to the occurrence of classic statistical suppression with respect to the competing path $Y \rightarrow Z$.

The reason for this is that the error term ε is expected to be in most cases uncorrelated with X and Y. Since the Simpson's paradox occurrence in the path $X \rightarrow Z$ leads to a reduction in the sum of terms on the right side of the equation, which must add up to 1, classic statistical suppression in connection with the competing path $Y \rightarrow Z$ would make up for that reduction.

That is, the product $p_{ZY}r_{ZY}$ would go up in value, via classic statistical suppression (as p_{ZY} goes up in value, since r_{ZY} remains the same), thus compensating for the negative contribution of the product $p_{ZX}r_{ZX}$ to the sum on the right side of the equation (which must add up to exactly 1). This is why instances of Simpson's paradox often occur together with instances of classic statistical suppression, even though the two phenomena are clearly different.

Having said the above, it is important to note that Simpson's paradox does not always occur together with classic statistical suppression. If the error term ε is correlated with X or Y, then Simpson's paradox may occur without corresponding classic statistical suppression. Moreover, and for related reasons, classic statistical suppression may occur without Simpson's paradox.

Appendix D: Mathematical properties of path models

The method of path analysis was developed by Sewall Wright, a main contributor to evolutionary biology and one of the founders of the field of population genetics (Duncan, 1966; Kenny, 1979; Mueller, 1996; Wright, 1934; 1960). Path analysis is the foundation of structural equation modeling (Kline, 1998; Mueller, 1996).

This appendix contains well known mathematical properties of linear path models; properties on which path analysis builds, and which are relevant in the context of the mathematical discussions presented earlier. Detailed proofs and discussions of these properties are provided by Mueller (1996) and Wright (1934; 1960).

First law of path analysis

The first law of path analysis states that the correlation between any variable X and a variable Z, where Z is directly and/or indirectly caused by X, is given by the equation below. In it, i is an index that represents a set of variables that directly and/or indirectly cause Z:

$$r_{ZX} = \sum p_{Zi} r_{Xi}.$$

For example, if the variables X and Z are connected through a direct path $X \rightarrow Z$ and through an indirect path with two segments $X \rightarrow Y \rightarrow Z$, then $r_{ZX} = p_{ZY} r_{XY} + p_{ZX} r_{XX} = p_{ZY} r_{XY} + p_{ZX}$. Here the term r_{XX} represents the correlation of X with itself, which is 1.

Coefficient of determination

The coefficient of determination of a variable Z, or the percentage of explained variance in Z, is given by the equation below. In it, the index i represents variables that are members of a set of variables that directly cause Z:

$${R_Z}^2 = \sum p_{Zi} r_{Zi}.$$

For example, if the variables X and Y point at Z, then ${R_Z}^2 = p_{ZX} r_{ZX} + p_{ZY} r_{ZY}$. This is the case regardless of whether X and Y are causally linked or correlated. Having said that, if X and Y are not correlated then $p_{ZX} = r_{ZX}$ and $p_{ZY} = r_{ZY}$, thus making ${R_Z}^2 = {r_{ZX}}^2 + {r_{ZY}}^2$.

Contribution of error in relation to the coefficient of determination

The individual contribution of the error term associated with a variable Z, in relation to the coefficient of determination of Z, is given by:

$${R_{Z\varepsilon}}^2 = 1 - {R_Z}^2.$$

This also means that:

$${R_{Z\varepsilon}}^2 = 1 - \sum p_{Zi} r_{Zi},$$

where i represents a set of variables that directly cause Z; and that:

$$1 = \sum p_{Zi} r_{Zi} + p_{Z\varepsilon} r_{Z\varepsilon}.$$

That is, the set of variables that directly cause Z, plus the error term associated with Z, explain all of the variance in Z.

Appendix E: True model data creation

The true model data used in our illustration was generated through a Monte Carlo simulation (Paxton et al., 2001). Researchers interested in generating similar simulated datasets can employ two main procedures (Robert & Casella, 2005; Reinartz et al., 2002; Reinartz et al., 2009). One of these procedures involves using the covariance matrix associated with the model, provided in Table E.1, as an input for data generation functions available from multi-purpose covariance-based structural equation modeling tools such as Mplus, LISREL, and EQS. In this case, the data would normally be generated via a Cholesky decomposition (Harbrecht et al., 2012).

Table E.1: Covariance matrix for true model data creation

	X	*Y*	*Z*
X	1.000		
Y	.726	1.000	
Z	.425	.656	1.000

The other procedure for generating similar simulated datasets involves creating variable scores directly by iteratively solving the system of equations below. This can be done with general numeric computing tools such as MATLAB. In these equations $Randn()$ denotes a random number generator that yields different normally distributed random values each time it is invoked, $Stdz()$ the standardization function, and $Corr()$ the correlation function.

$$r_{XY} = .726.$$
$$p_{ZY} = .691.$$
$$m_{ZYX} = .360.$$
$$\delta = Stdz(Randn()).$$
$$\gamma = Stdz(Randn()).$$
$$X = Stdz(Randn()).$$
$$r_{Y\delta} = Corr(Y, \delta).$$
$$r_{Z\gamma} = Corr(Z, \gamma).$$
$$r_{ZY} = Corr(Z, Y).$$
$$r_{ZM} = Corr(Z, M).$$
$$p_{Y\delta} r_{Y\delta} + {r_{XY}}^2 = 1.$$
$$p_{Z\gamma} r_{Z\gamma} + p_{ZY} r_{ZY} + m_{ZYX} r_{ZM} = 1.$$
$$Y = r_{XY} X + p_{X\delta} \delta.$$
$$M = Stdz(XY).$$
$$Z = p_{ZY} Y + m_{ZYX} M + p_{Z\gamma} \gamma.$$

Typically the data created based on the procedures above will initially be standardized. Unstandardization can be easily accomplished through multiplication of the variable scores by a specified standard deviation and addition of a specified mean. Additionally, rounding can be used if data on a Likert-type scale is desired.

Appendix F: An illustrative analysis with real data

This appendix briefly discusses an illustrative analysis based on real data through a model that depicts relationships among the variables Internet diffusion (*I*), voice and accountability (*V*), and government corruption (*C*). This analysis is presented here only for illustration purposes, and is not meant to be an empirical contribution. The data used in this illustrative analysis covers 47 countries; 24 in Latin America and 23 in Sub-Saharan Africa. The data spans 5 years (2006 to 2010), adding up to a total sample size of 235 (47 x 5).

As done in the rest of this paper, our discussion here focuses on a path model, where variables are measured through single indicators. Nevertheless, it applies to the more general case of structural equation models in which latent variables are measured indirectly through more than one indicator, whether the measurement is reflective or formative (Kock & Mayfield, 2015). Our discussion assumes the existence of latent variable scores in structural equation models, which would have been obtained through an appropriate algorithm prior to the considerations discussed here.

Internet diffusion was measured by the number of Internet users per 100 inhabitants in a country, obtained from the World Bank (http://www.worldbank.org). Voice and accountability was measured through the eponymous index, also from the World Bank. Government corruption was measured through the Corruption Perceptions Index published by Transparency International (http://www.transparency.org). The Corruption Perceptions Index scores were reversed to properly reflect the degree of corruption of countries and facilitate interpretation of the results; the original index measures the lack of corruption.

Figure F.1 shows the incorrectly specified path model. It would be reasonable to expect a researcher to hypothesize this model based on theory and past empirical research. This model is similar to the incorrectly specified model discussed before in the illustrative analysis with true model data, with the difference that this model includes control variables.

This model displays an instance of Simpson's paradox occurring in connection with the path going from *I* to *C*. No Simpson's paradox instances occurred in connection with any of the control variables, which were: year in which the data was collected, gross domestic product (GDP) per capita, and the four main country-specific cultural dimensions from Hofstede's (1983; 2001) model – power distance, uncertainly avoidance, long-term/short-term orientation, and individualism/collectivism.

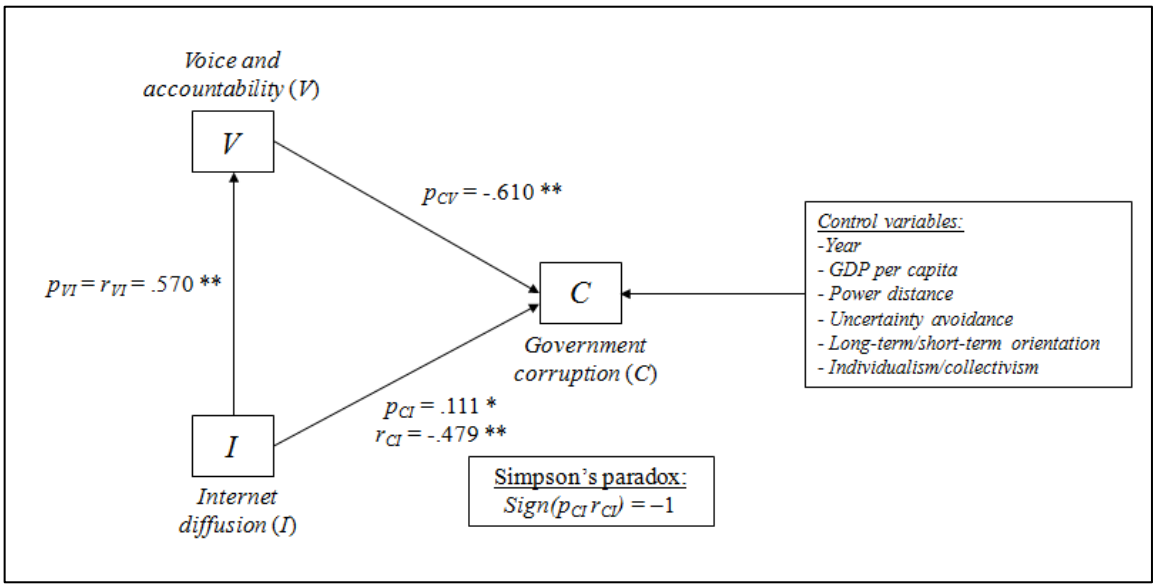
Here we see the same type of absurdity arising from Simpson's paradox as in the illustrative analysis of true model data, when we consider the coefficient for the direct path from *I* to *C* together with the coefficient for the total effect of *I* on *C*. The coefficient for the direct path from *I* to *C* is .111 (P<.05) suggesting that Internet diffusion (*I*) contributes to an increase in government corruption (*Z*). The coefficient for the total effect of *I* on *C* is -.237 (P<.01), suggesting that Internet diffusion (*I*) contributes to an overall decrease in government corruption (*Z*).

Figure F.2 shows what our previous discussion would suggest to be the correctly specified path model including a constrained mediated moderation relationship. This model does not include any instance of Simpson's paradox. The Simpson's paradox occurrence in the incorrectly specified model and its disappearance here suggest the possible existence of a quadratic relationship, which is not yet explicitly modeled here.

As we have done earlier in the illustrative analysis with true model data, we conducted full collinearity tests on both models (true model and incorrectly specified model), checking for both

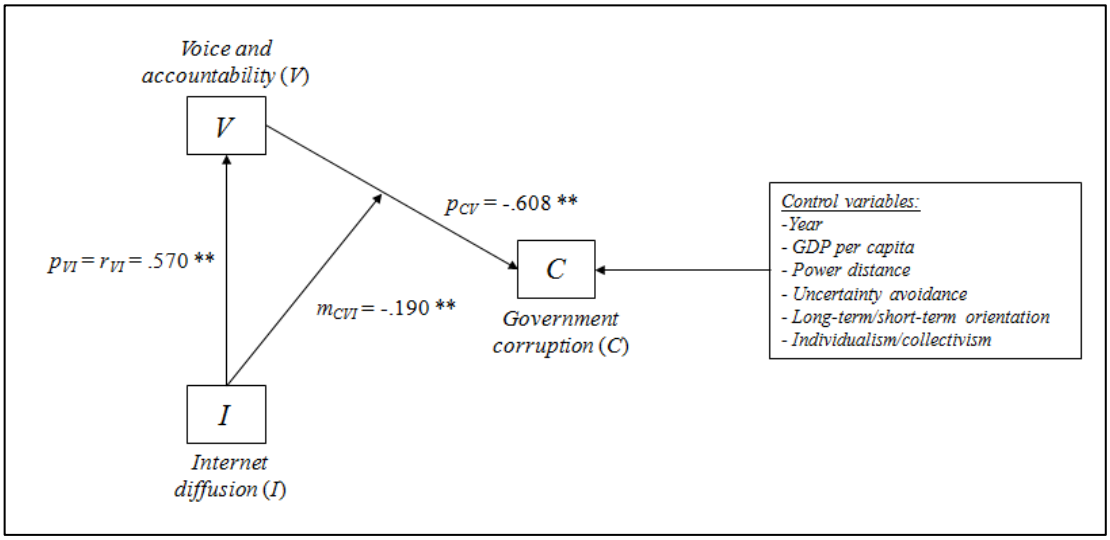
vertical and lateral collinearity. We found no evidence of multicollinearity, as all variance inflation factors were lower than the threshold of 3.3 proposed by Kock & Lynn (2012).

Figure F.1: Incorrectly specified path model with Simpson's paradox



Notes: ** P < .01, * P < .05

Figure F.2: Correctly specified path model without Simpson's paradox

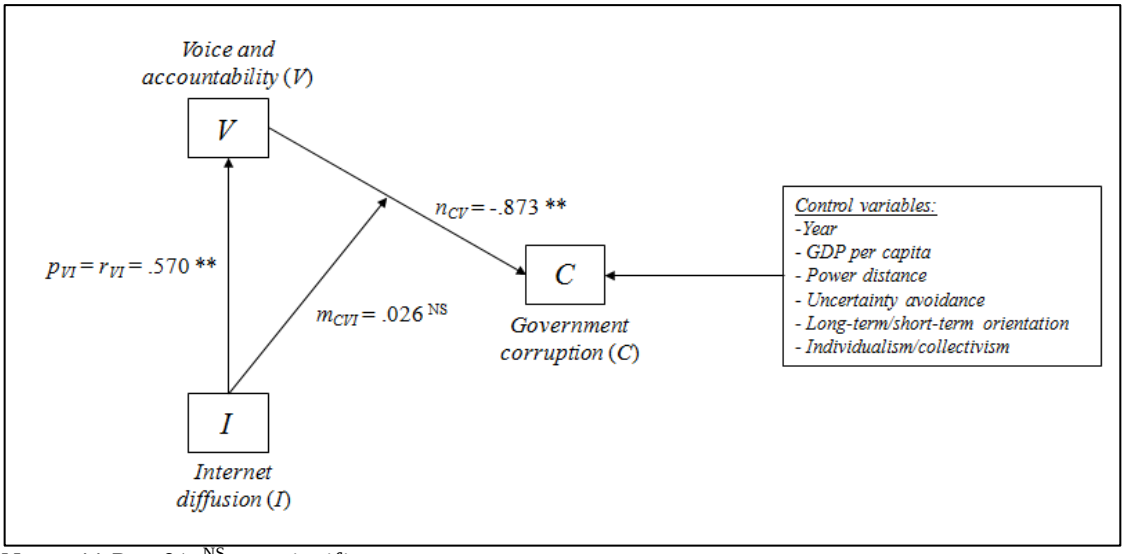


Note: ** P < .01

Figure F.3 shows the model and its path coefficients when we explicitly model the quadratic relationship. Consistently with our mathematical development and previous illustration based on true model data, the moderation relationship coefficient decreased to the point of losing statistical significance.

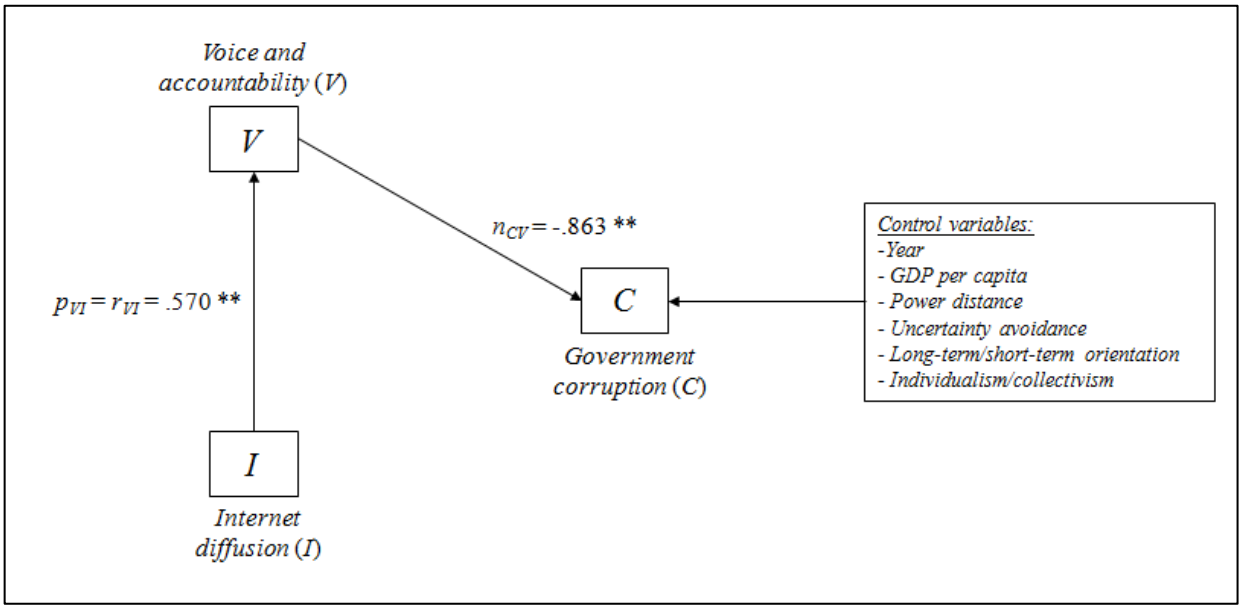
Figure F.4 shows a parsimonious version of the model, without the moderation relationship that is no longer statistically significant. Here we continue explicitly modeling the quadratic relationship.

Figure F.3: Path model including moderation and nonlinear effects



Notes: ** P < .01, NS non-significant

Figure F.4: Parsimonious path model including nonlinear effect



Note: ** P < .01

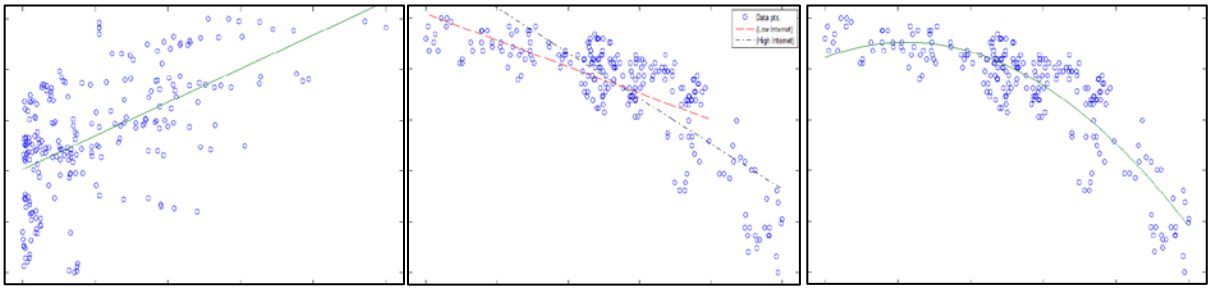
Figure F.5 contains three plots side-by-side showing three relationships: the linear relationship between Internet diffusion (*I*) and voice and accountability (*V*), the moderation relationship between voice and accountability (*V*) and government corruption (*C*), and the nonlinear relationship between voice and accountability (*V*) and government corruption (*C*).

As with the previous illustrative analysis with true model data, we can clearly see here the moderation-nonlinear relationship duality.

In contrast to the previous illustrative analysis with true model data, here we do not actually know the true nature of the underlying relationships. Therefore, we also explored the possibility

that the relationship between Internet diffusion (*I*) and voice and accountability (*V*) was nonlinear, by explicitly modeling it as a quadratic relationship. The results of this exploratory analysis did not support nonlinearity, as the resulting shape was very close to that of a line.

Figure F.5: Plots showing linear, moderation, and nonlinear relationships



Notes:
- Left: linear relationship between Internet diffusion (*I*) and voice and accountability (*V*)
- Middle: relationship between voice and accountability (*V*) and government corruption (*C*), moderated by Internet diffusion (*I*)
- Right: nonlinear relationship between voice and accountability (*V*) and government corruption (*C*), where the nonlinearity is due in part to Internet diffusion (*I*)