THE ENDOGENEITY PROBLEM IN MEDIATION ANALYSIS

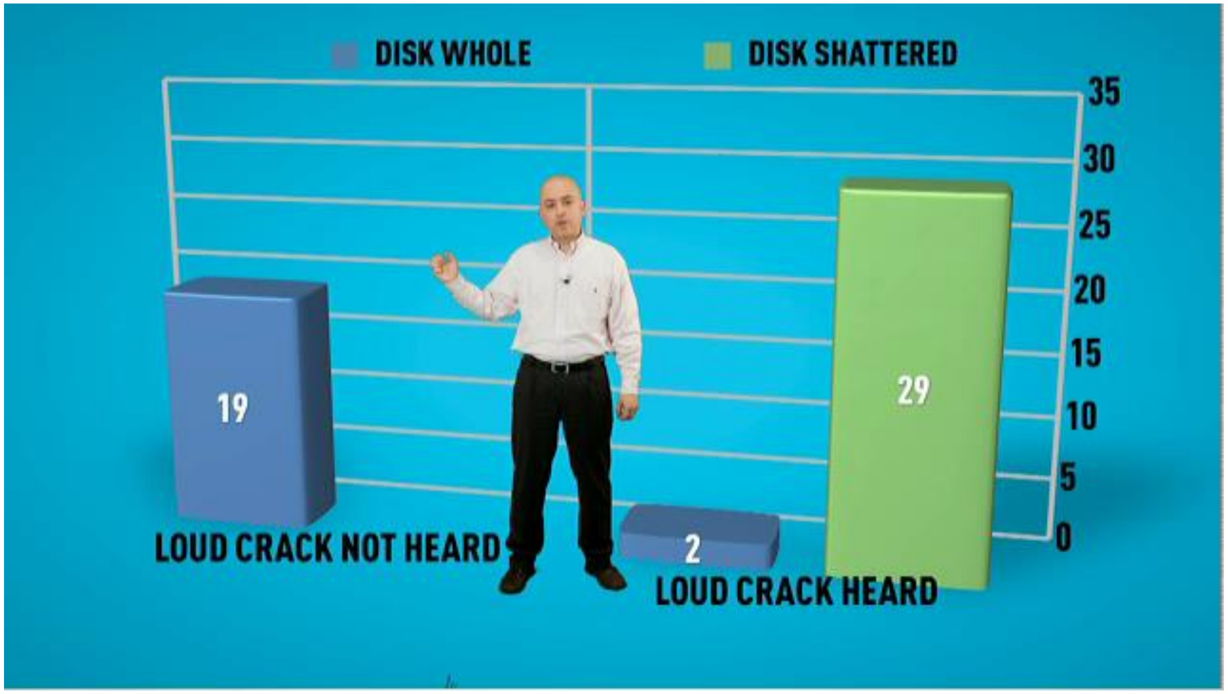
John Antonakis
Faculty of Business and Economics
University of Lausanne

University of Exeter

Research Seminar
4 February 2020

What a basic introduction to Endogeneity?

Available on Youtube as "Endogeneity: An inconvenient truth (full version)"

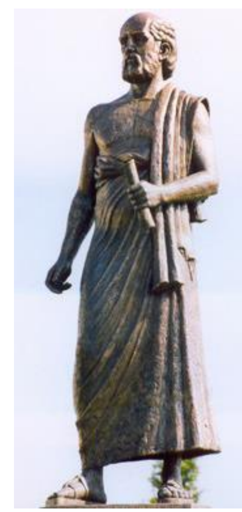


See at least from start to 08:50

Unbeknown to many researchers in management, marketing, and applied psychology endogeneity is a serious problem in mediation. Despite pleas to undertake mediation tests correctly, we are currently in a sorry state of affairs; Shaver (2005), who got it right, has been completely ignored:

Authors	Citations in WOS
Baron and Kenny (1986)	39,722
Shaver (2005)	122
Preacher & Hayes (2004, 2008)	21,008

Remember Aristarchus of Samos and that might is not always right.



Testing mediation in management and applied psychology research is problematic

- Kline (2015) described a "mediation myth," stating "relatively little of the extant literature on mediation is actually worthwhile" (p. 210).
- Mediation is about testing causal questions and ensuing policy implications (Antonakis, Bendahan, Jacquart, & Lalive, 2010); in the presence of endogeneity, policy cannot usually be informed .
- In recent review (n = 110 articles), not **<u>one</u>** study tested mediation correctly (i.e., causally, see Antonakis, et al., 2010)!

Recently we found 3 out of 189 articles that did right (Fischer, Dietz, & Antonakis, 2017).

Table 2
List of Coding Events

	Modeling Multiple Mediation Paths	Avoiding Same Source Sampling	Use of Exogenous Predictor(s)	Use of Instrumental Variable Approach	Use of Time Lags[a]
Yes	69	40	49[b]	3	65[c]
No	120	149	140	186	124
Yes, %[d]	37	21	26	2	34

Note: $N = 189$ quantitative-empirical articles.
[a]We report on the use of time lags in the section about time.
[b]Twenty-two articles use an exogenous experimental manipulation, and 28 articles use stable predictors like IQ.
[c]Forty-three articles use standard measurement of time lags (t1, t2, etc.) in field studies, and 22 articles use sequential measurement in laboratory experiments.
[d]Note that all percentages are rounded to zero-decimal numbers.

Why this sad state of affairs? Most applied researchers don’t know that:

1. “Typical” mediation methods, the ones most often used in applied psychology and management research, assume the *exogeneity* of mediator

2. Violation of above assumption means *biased* and *inconsistent* estimates that will not improve asymptotically or with bootstrapping

3. There is a solution to this problem if the mediator cannot be manipulated: Instrumental variable estimation

The problem is on the radar in psychology; for example, Smith (2012) in a recent editorial in *Journal of Personality and Social Psychology* stated:

(capitalização)

"if the independent variable (X) is manipulated and the mediator (M) and dependent variable (Y) are measured, the usual analysis will be biased if unobserved causes of M are correlated with unobserved causes of Y" (p. 2).

$$x \rightarrow m \rightarrow y$$

Why? There could be a third unmeasured variable, *q*, that predicts both *m* and *y*. Or *m* could be endogenous for <u>other reasons</u> (i.e., measurement errors)

Let's see, more formally, what mediation (i.e., full mediation) entails:

$$m = \gamma_0 + \gamma_1 x + v \qquad \text{Eq. 1}$$

$$y = \beta_0 + \beta_1 m + w \qquad \text{Eq. 2}$$

The key assumptions are made to ensure the test of $\gamma_1 . \beta_1 = 0$ are valid:

1. The disturbance *v* is orthogonal to *x* (*x* is exogenous with respect to *m*)
2. The disturbance *w* is orthogonal to *m* (*m* is exogenous with respect to *y*).

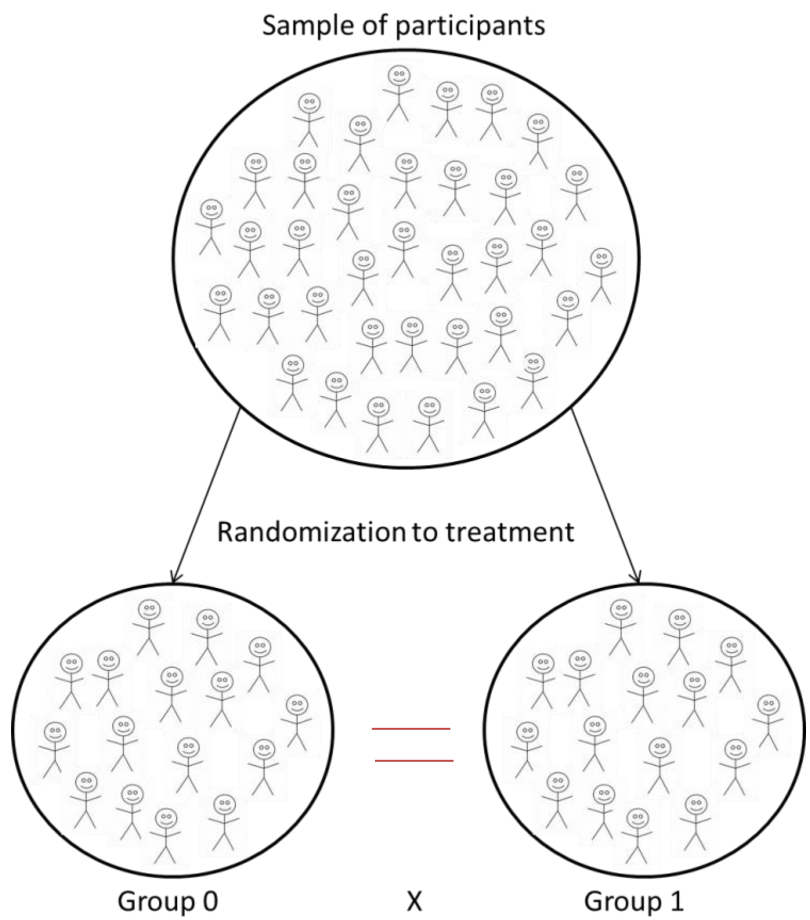
If you manipulated *x* (or *x* is exogenous for other reasons) but not *m*, then you potentially face the beast: Endogeneity. And bootstrapping won't help!

Example: We test if charisma can be trained in managers, and whether increases in charisma impact subordinate satisfaction with the managers. The mediator, *m* (e.g., charisma), is thought to causally channel the effect of a random independent variable, *x* (e.g., receiving training or alternative treatment), on another dependent variable (e.g., subordinate satisfaction), *y*.

Training → *Charisma* → *Satisfaction*

Most theories of mediation present the *m* as an outcome of *x*; thus, *m* by definition is a dependent—or endogenous—variable. How can it not be?

Randomization on *x* assumes that managers in the two treatment groups are approximately equal or interchangeable. Any variable (*q*) that could predict a DV is equally distributed across the groups (Shadish, Cook, & Campbell, 2002).



tylo

With randomization there is no correlation between *x* and *q* (Antonakis, et al., 2010). Thus measuring or not *q* and including it in a regression model will make no difference to the effect of *x* on *m* or any other DV. That is:

$$m = \gamma_0 + \gamma_1 x + v \qquad \text{Eq. 1}$$

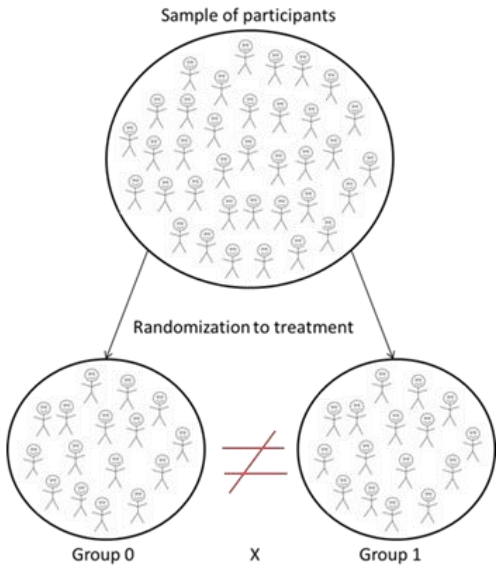
$$m = \delta_0 + \delta_1 x + \delta_2 q + w \qquad \text{Eq. 1.1}$$

Asymptotically: $\delta_1 = \gamma_1$

Although in Eq. 1, *q* is in *v*, *v* is orthogonal to *x*. Even if $\delta_2 \neq 0$, δ_1 is unaffected.

Now, let's screw up randomization for a sec; Suppose:

- q measures extraversion
- extraversion predicts *m* (ratings of charisma)
- there were more extraverted managers in group 1.



If group 1 scores higher on *m* than does group 0, is this due to treatment or extraversion? This answer is unclear. In this case $\delta_1 \neq \gamma_1$ (unless we control for q). Thus, $\delta_1$is affected if $cov(x, q) \neq 0$, which is avoided by randomization. Now, q can be a zillion things (e.g., looks, IQ, etc.)—we cannot know q, which is why we randomize to treatment and *x* is thus exogenous. All clear so far?

How does endogeneity enter into a mediation model?

Even with an exogenous *x* (i.e., proper randomization), we could have a problem if extraversion, good looks, intelligence or any other *q*, predicts both measured outcomes:

- charisma (*m*)
- and satisfaction with the leader (*y*)

In the above, it is not unreasonable to see that *q* predicts both *m* and *y*. The problem we have now is that if we use *m* to predict *y*, is the relation we find due to (a) *q* and/or (b) due to *x'*s effect on *y* via *m*?

BTW, endogeneity can enter in many ways, including

- omitted variables (i.e., q's)
- measurement errors in m (which will almost always be the case if m is not perfectly observed)
- simultaneity or reverse causality (Antonakis, et al., 2010; Antonakis, Bendahan, Jacquart, & Lalive, 2014).

In fact, most of the time, *m* <u>will be</u> endogenous, and this simply by definition, because *m* is endogenous to *x* and is not manipulated. How can *m* be exogenous?

A challenge to all in the room: Can *m* ever be exogenous?

How endogeneity biases parameter estimates and why the "usual" estimators for mediation do not work?

Following the previous example of an experiment in which the researcher examines mediation, suppose that the true model is as follows (and this model is reasonable and what we would usually find):

$$m = \lambda_0 + \lambda_1 x + \lambda_2 q + \varepsilon \qquad \text{Eq. 3}$$

$$y = \eta_0 + \eta_1 m + \eta_2 q + \theta \qquad \text{Eq. 4}$$

Where ε and θ are random disturbances, and x and q random variables. Suppose that the researcher has not measured q (remember q is a bunch of causes), and estimates instead, the following:

$$m = \omega_0 + \omega_1 x + u \qquad \text{Eq. 5}$$

$$y = \xi_0 + \xi_1 m + o \qquad \text{Eq. 6}$$

The coefficient ω_1 will remain unaffected and will equal λ_1 in Eq. 3. Thus, there is no problem in Eq. 5 because *x* is random; but *m* is not. The problem is in Eq. 6, because both *m* and *y* correlate with *q* and the effect of *q* is pooled in *o*.

To formally see the problem, we first model the relation between *q* and *m* (they correlate). The direction of the causal relation is not relevant here; we can thus write out *q* as a function of *m* (and omit the constant for simplicity):

$$q = \kappa_1 m + \varphi \qquad \text{Eq. 7}$$

Substituting Eq. 7 into Eq. 4 ($y = \eta_0 + \eta_1 m + \eta_2 q + \theta$) shows:

$$y = \eta_0 + \eta_1 m + \eta_2(\kappa_1 m + \varphi) + \theta \qquad \text{Eq. 8}$$

Multiplying out gives:

$$y = \eta_0 + \eta_1 m + \underbrace{(\eta_2 \kappa_1 m + \eta_2 \varphi + \theta)}_{o} \qquad \text{Eq. 9}$$

Notice, that *o* in Eq. 6 is a "super-disturbance" Rearranging as function of *m*:

$$y = \eta_0 + (\eta_1 + \eta_2 \kappa_1)m + \eta_2 \varphi + \theta \qquad \text{Eq. 10}$$

Thus, the OLS (ordinary least squares) or ML (maximum likelihood) estimate of ξ_1 in Eq. 6 will be biased because (from Eq. 5 and 10):

Bias term

$$\xi_1 = \frac{cov(y,m)}{var(m)} = \eta_1 + \eta_2\kappa_1 \neq \eta_1 \qquad \text{Eq. 11}$$

Recall, the true model is:

$$m = \lambda_0 + \lambda_1 x + \lambda_2 q + \varepsilon \qquad \text{Eq. 3}$$

$$y = \eta_0 + \eta_1 m + \eta_2 q + \theta \qquad \text{Eq. 4}$$

And we estimate:

$$m = \omega_0 + \omega_1 x + u \qquad \text{Eq. 5}$$

$$y = \xi_0 + \xi_1 m + o \qquad \text{Eq. 6}$$

ξ_1 will only equal η_1 if either η_2 or κ_1 are zero (i.e., if q is unrelated to y or to m).

Thus, in the presence of endogeneity, typical estimation procedures used in I-O psychology such as (a) the regression procedures of Baron and Kenny (1986), Preacher and Hayes (2004), or (b) structural equation model (Edwards & Lambert, 2007) will not work because they assume that *m* is exogenous.

When testing full mediation:

$$m = \omega_0 + \omega_1 x + u \qquad \text{Eq. 5}$$

$$y = \xi_0 + \xi_1 m + o \qquad \text{Eq. 6}$$

The indirect (mediated effect) is $\omega_1 * \xi_1$. Because $\xi_1 \neq \eta_1$ the indirect effect is *inconsistent*.

Sometimes researchers estimate a partial mediation model (Eq. 13), which also assumes m to be exogenous.

$$m = \omega_0 + \omega_1 x + u \qquad \text{Eq. 5}$$

$$y = \mu_0 + \mu_1 m + \mu_2 x + i \qquad \text{Eq. 13}$$

Where the indirect (mediated effect) is $\omega_1 * \mu_1$. Both μ_1 and μ_2 will now be inconsistent because the endogeneity bias that affects the coefficient of *m* will also affect the coefficient of *x*, because that *m* and *x* are correlated and the bias is transmitted via that correlation (Antonakis, et al., 2010). Thus, μ_1 (in Eq. 13) $\neq \eta_1$ (in Eq. 4).

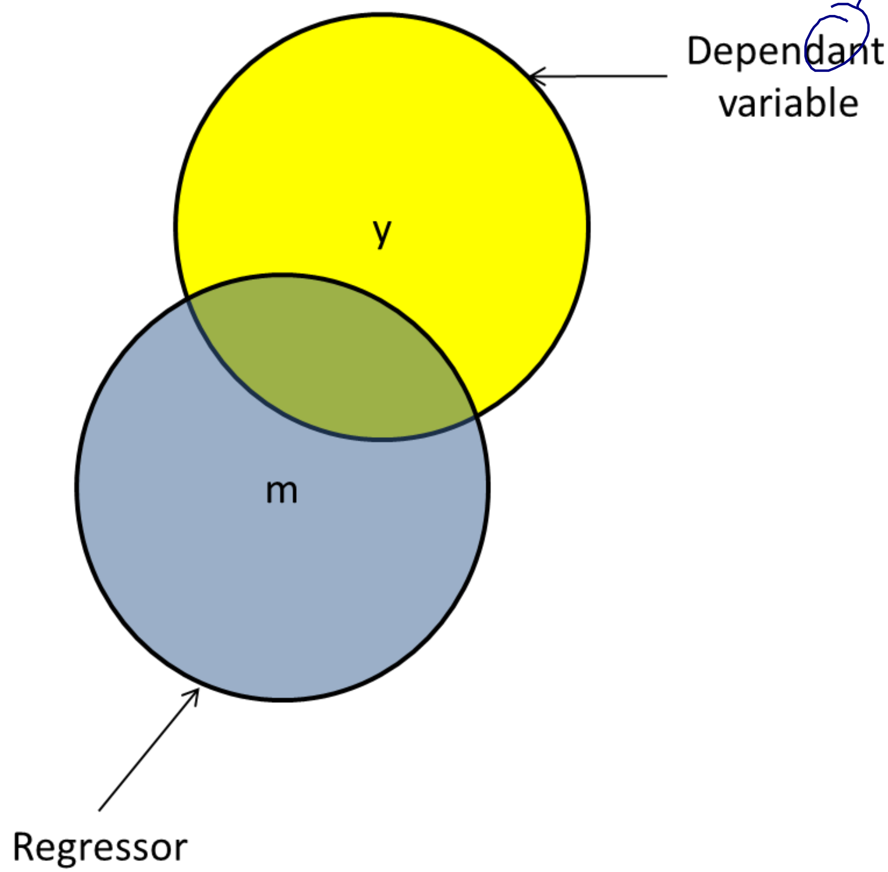
No amount of bootstrapping—as per the Preacher and Hayes guidelines—of an inconsistent estimate will render it consistent. The endogeneity problem and relevant assumptions made in testing mediation *have been acknowledged* by:

1. Edwards and Lambert (2007, p. 19)
2. Baron and Kenny (1986, p. 1177)
3. Hayes (2013, p. 173)

However, the above have not warned researchers of the problem of failing the exogeneity assumption and most applied researchers ignore the issue entirely (Antonakis, et al., 2010).

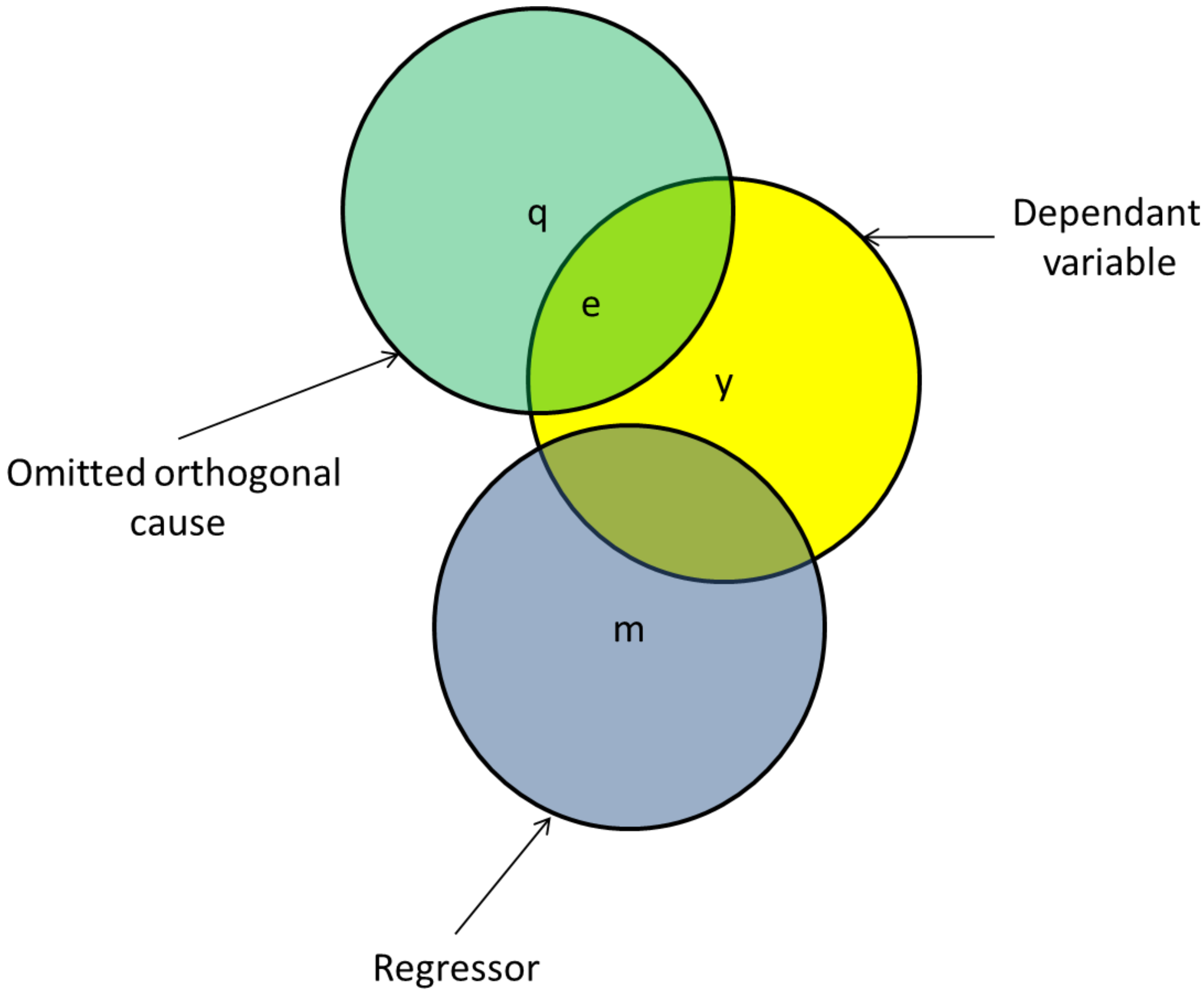
The solution to the endogeneity problem in mediation.

Spelling

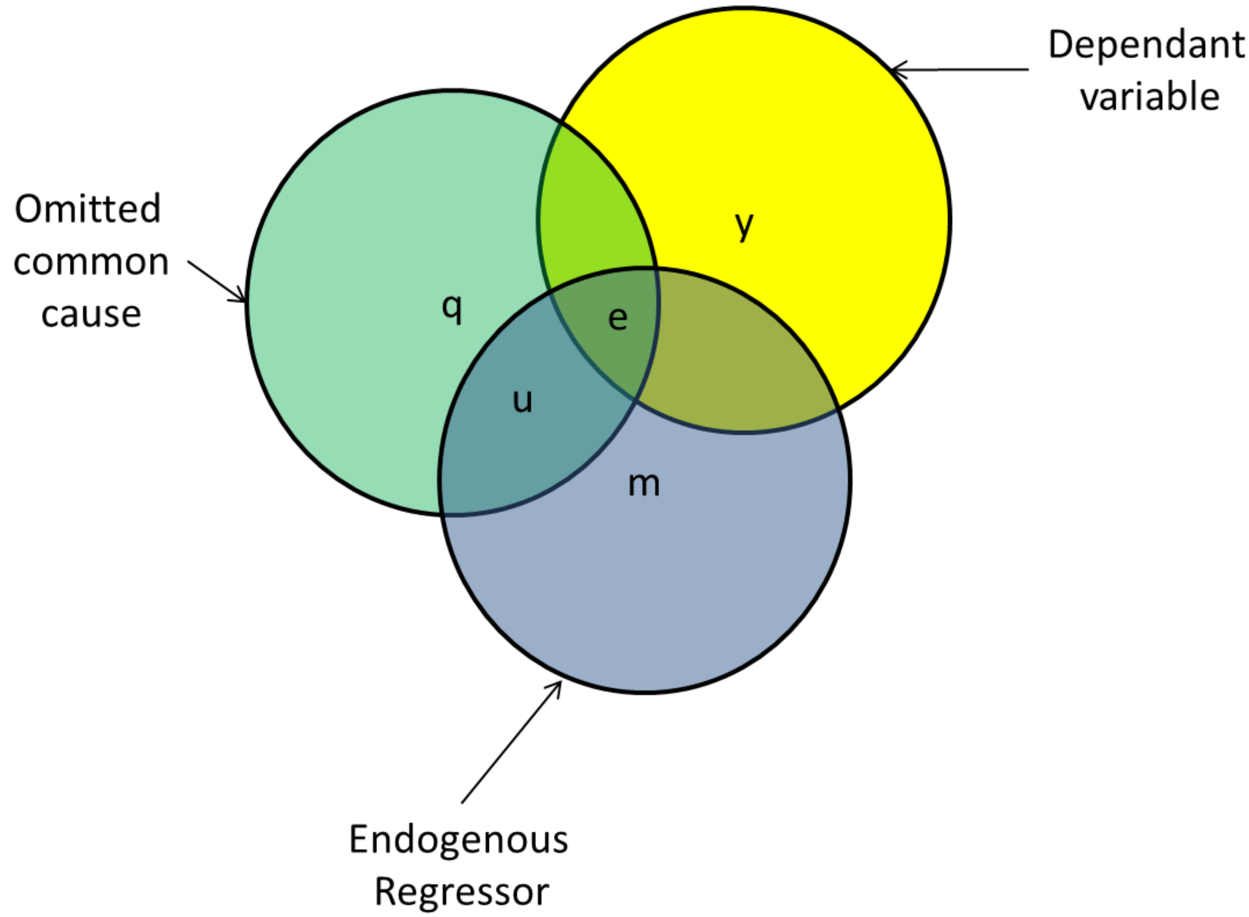


The overlap in ballentines is the variance that the two variables share.

?



Here we have no problem of endogeneity (*e* does not overlap with *m*).

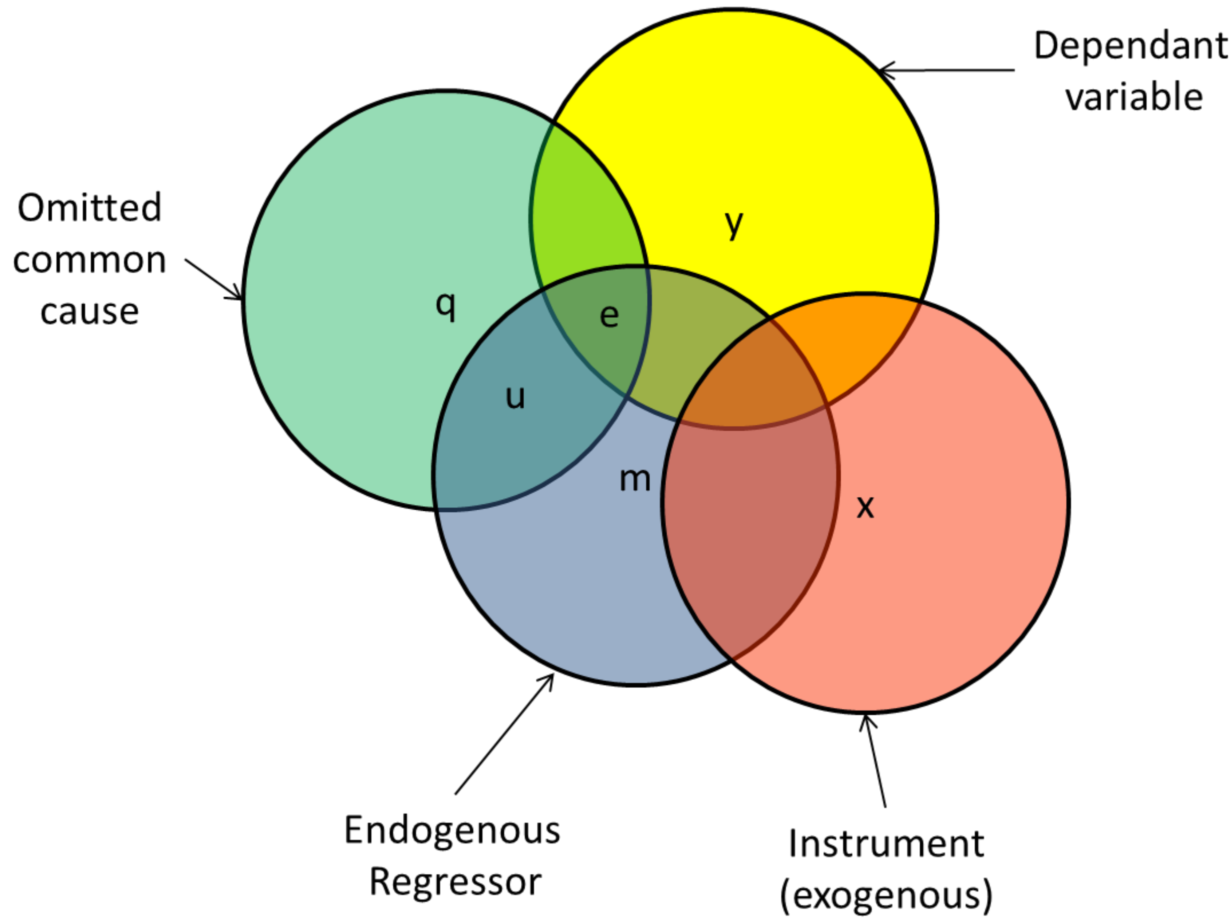


This specification is problematic, *m* correlates with *e*. Thus, if we estimated the yellow-blue-green overlap, the coefficient will not be consistent.

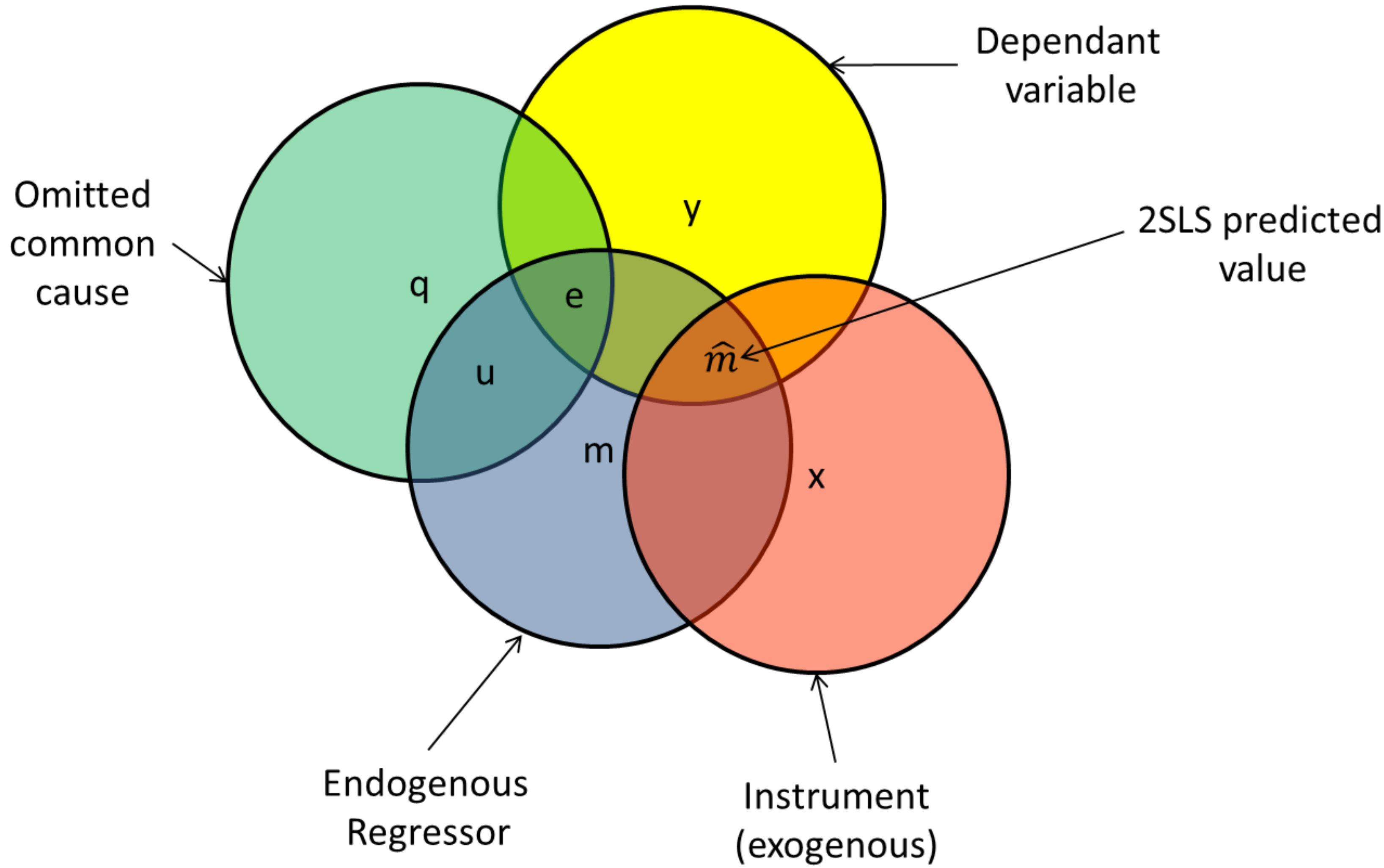
Now, what if we don't know what common-cause variables are potentially omitted?

- One way is to manipulate *m*; however, sometimes this is not practical to do
- Another way to get around the problem of endogeneity is to find an instrument/s:
  - is an exogenous source of variance (thus, by definition it does not correlate with *e*).

Graphically (note, the instrument here is called *z* and not *x*):

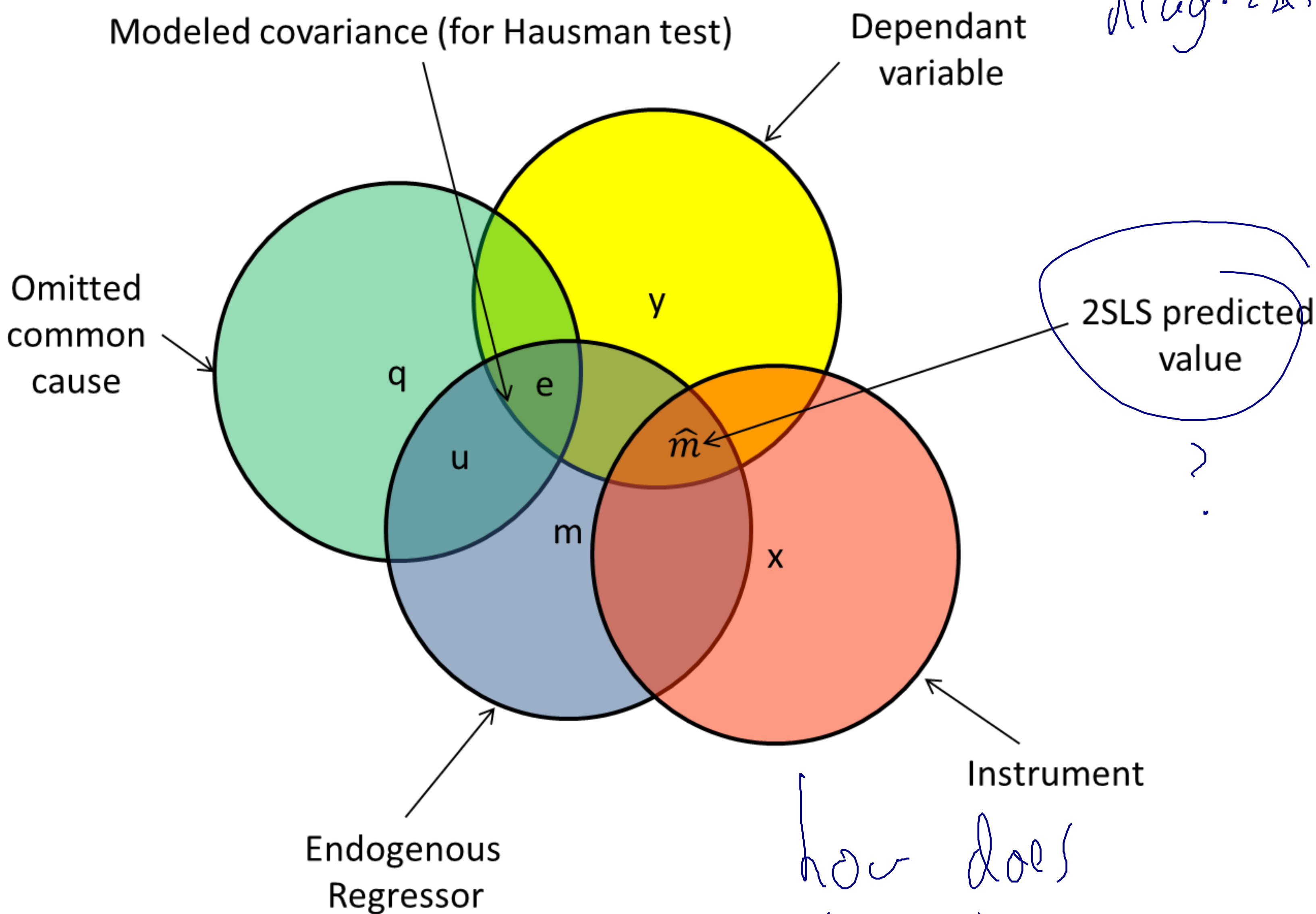


Because *x* does not correlate with *e* we can use the portion of variance that *x* predicts in *m* (that overlaps with *y*; it *must* if *m* is a predictor of *y*) to predict *y*.



$\widehat{m}$ uses less information; however, the estimate is consistent, that is, the slope is correct (even though less efficient, i.e., with higher SEs).

confusing diagram



?

how does the diagram show exclu?

Let's see the algebra. Recall, the true model is:

$$m = \lambda_0 + \lambda_1 x + \lambda_2 q + \varepsilon \qquad \text{Eq. 3}$$

$$y = \eta_0 + \eta_1 m + \eta_2 q + \theta \qquad \text{Eq. 4}$$

The "reduced form" model for the direct effect of *x* on *y* can be written as:

$$y = \beta_0 + \Pi x + w \qquad \text{Eq. 14}$$

Where $\Pi = \eta_1 * \lambda_1$. Thus, if we harness the exogeneity of *x* (called the "instrument"), the predicted value of the coefficient of η_1 is (see Angrist & Pischke, 2008):

$$\hat{\eta}_1 = \frac{\hat{\Pi}}{\hat{\lambda}_1} = \frac{Cov(y,x)/Var(x)}{Cov(m,x)/Var(x)} = \frac{Cov(y,x)}{Cov(m,x)} \qquad \text{Eq. 15}$$

Voilà: This is the general formula for an instrumental-variable estimate (e.g., see Bollen, 2012):

$$\hat{\eta}_1 = \frac{Cov(y, x)}{Cov(m, x)}$$

Thus, *x*, the <u>instrument</u>, *must* correlate with *y* and *m* if *m* is a true cause of *y* (and if *m* is endogenous)*;* else, the estimate of Eq. 15 is either zero or undefined, which shows up another fallacy in mediation—and a widely believed one too—that *x* need not correlate with *y* to establish mediation (Shrout & Bolger, 2002; Zhao, Lynch, & Chen, 2010); this belief is only true if *m* is exogenous!

clarify this

Note, from Eq. 15 we know (and substituting Eq. 4, $y = \eta_0 + \eta_1 m + \eta_2 q + \theta$, into the numerator of Eq. 15):

$$\hat{\eta}_1 = \frac{Cov(y,x)}{Cov(m,x)} = \frac{Cov(\eta_0+\eta_1 m+\eta_2 q+\theta,x)}{Cov(m,x)} \qquad \text{Eq. 16}$$

Dropping constants and the disturbance, and expanding gives:

$$\hat{\eta}_1 = \frac{Cov(\eta_1 m,x)+Cov(\eta_2 q,x)}{Cov(m,x)} \qquad \text{Eq. 17}$$

$$\hat{\eta}_1 = \frac{\eta_1 Cov(m,x)+\eta_2 Cov(q,x)}{Cov(m,x)} \qquad \text{Eq. 18}$$

$$\hat{\eta}_1 = \eta_1 + \frac{\eta_2 Cov(q,x)}{Cov(m,x)} \qquad \text{Eq. 19}$$

Thus, if $cov(q,x) = 0$, then Eq. 15, $\hat{\eta}_1 = \eta_1$ (from Eq. 4).

Instrumental variable estimation can be achieved in basically two ways; recall:

$$m = \omega_0 + \omega_1 x + u \qquad \text{Eq. 5}$$

$$y = \xi_0 + \xi_1 m + o \qquad \text{Eq. 6}$$

1. Close-form solutions, like two-stage least squares, which uses the predicted value of *m* from Eq. 5 (stemming from *x*), $\hat{m}$, as the regressor in Eq. 6;

2. ML estimation, whereby the disturbances of Eq. 5 and Eq. 6, that is u and *o* are correlated (Antonakis, et al., 2010; Antonakis, Brulhart, & Lalive, 2017).

An empirical example: Instrumental-variable estimation vs "wrong way"

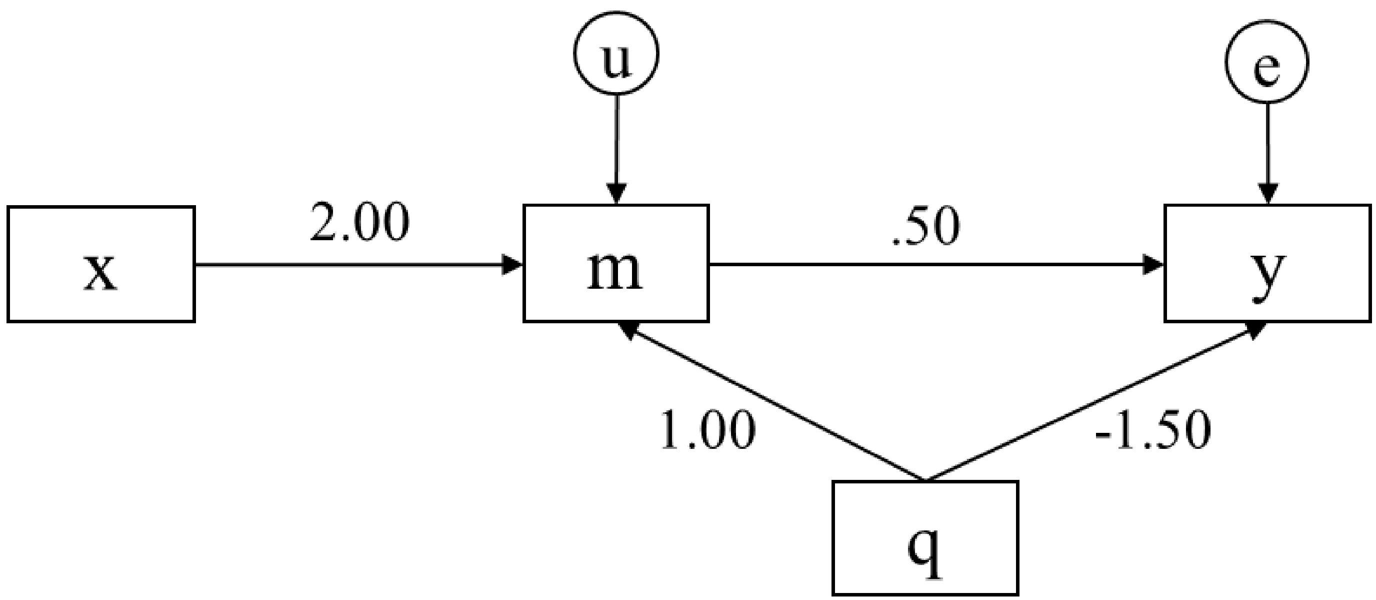
Assume the following dataset and correlation matrix on the below diagonal (n = 1,000); underlined entries on the top diagonal are the variance-covariance matrix:

	Mean	Std. Dev.	*x*	*q*	*m*	*y*
x	0.522	0.500	<u>0.250</u>	<u>-0.014</u>	<u>0.488</u>	<u>0.261</u>
q	10.000	1.037	-0.027	<u>1.075</u>	<u>1.045</u>	<u>-1.071</u>
m	11.022	1.780	0.549	0.566	<u>3.169</u>	<u>0.098</u>
y	5.509	1.659	0.314	-0.623	0.033	<u>2.753</u>

The data can be downloaded here:

http://www.hec.unil.ch/jantonakis/siop2015mediation.xlsx

Included below is the true model, as well as different estimated models. I list estimates with SEs in parentheses:



True model parameters

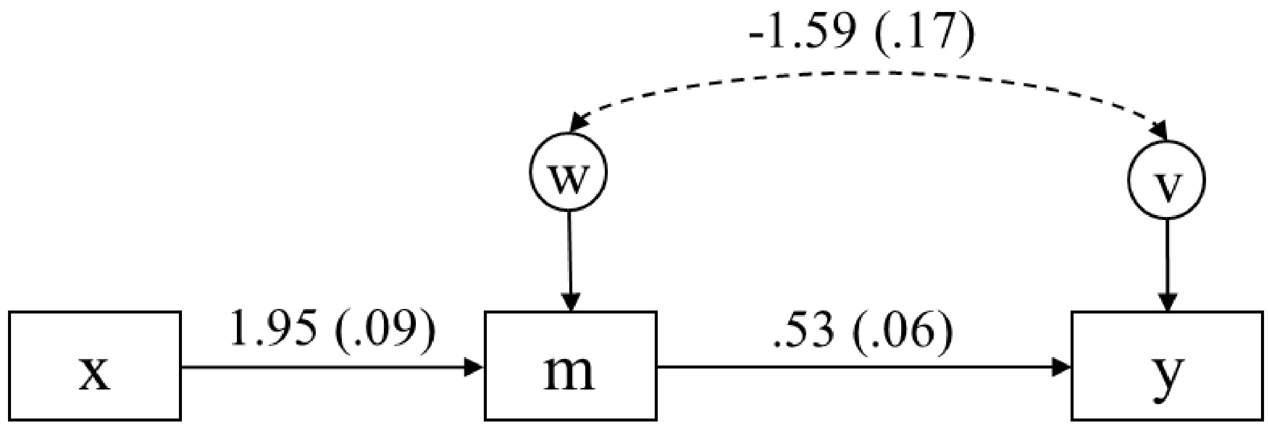
True model (for data generation)

The indirect effect is thus 2.00*.50 = 1.00. For estimation we will omit q.

Note, here is the reduced form estimate:

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    109.35
       Model |  271.554254         1  271.554254   Prob > F        =    0.0000
    Residual |  2478.36475       998  2.48333141   R-squared       =    0.0987
-------------+----------------------------------   Adj R-squared   =    0.0978
       Total |    2749.919       999  2.75267167   Root MSE        =    1.5759

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.043228   .0997627    10.46   0.000      .847459    1.238996
       _cons |   4.964435   .0720781    68.88   0.000     4.822993    5.105877
------------------------------------------------------------------------------
```

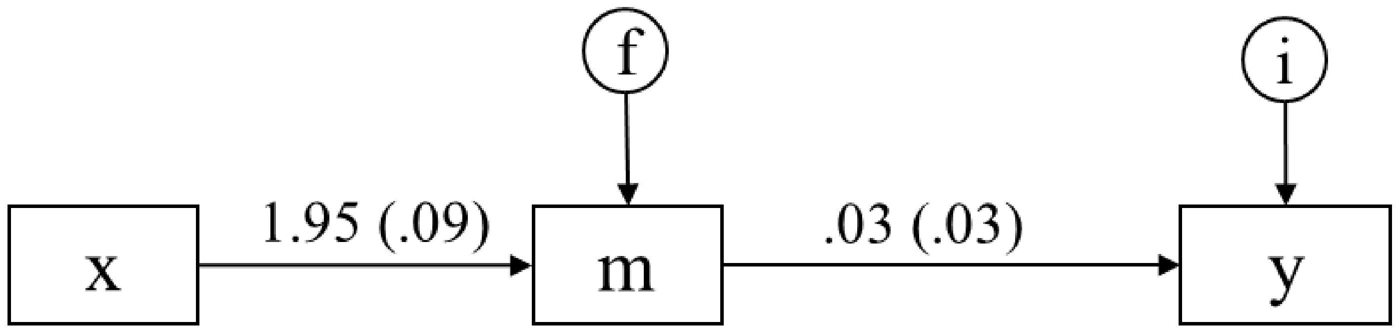
This estimate is good and equals the indirect effect. Let's see how we can recover this indirect effect, which should be 1.



Estimated (correct) parameters using ML

Estimated model 1 (instrumental variable model):

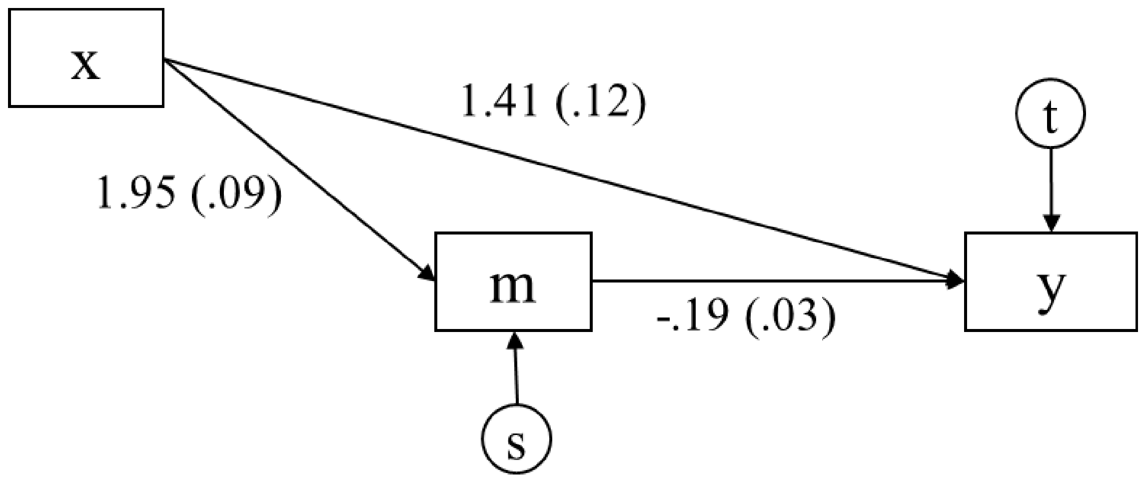
The estimate of *m* on *y* should be: $(Cov(y,x))/(Cov(m,x)) = 0.261/0.488 = .53$. ML gives an indirect effect of 1.95*.53 = 1.04, *SE* =.10, $z = 10.47$, $p < .001$. The significant covariance between the disturbances is the Hausman (1978) test, indicating the *m* is endogenous and needs to be instrumented (thus this is the test for endogeneity in *m*). Note, 2SLS gives the same result.



<u>Incorrect parameters (OLS/SEM) "full mediation"</u>

<u>Estimated model 2:</u>

Full mediation model using ML. The estimate of *m* on *y* is quite off as is the indirect effect: .06 (ns). In addition, the chi-square test of fit is significant, $\chi^2(1) = 134.12$, $p < .001$, indicating that the constrained made, (i.e. ,*cov*(f,i) = 0) is not consistent with the data. Unfortunately, applied researchers usually release the constraint that *x* does not affect *y* directly, leading to Model 3, which is also wrong.



Incorrect parameters (OLS/SEM): "partial mediation"

Estimated model 3:

Partial mediation model using ML. Estimates are rather off. The indirect effect of x on y is -.3627, SE = .0667, z = 5.44, p < .001. The Baron and Kenny specification, with OLS, gives the same point estimate for the indirect effect (-.3627) with an SE = .0668 and z = 5.43 (Sobel, 1982). Total effect is correct and the indirect can be recovered here (i.e., direct effect + indirect effect); the standard error of the difference requires use of the delta method (Oehlert, 1992).

```
. reg3 (y = x m) (m = x), ols

Multivariate regression
--------------------------------------------------------------------------
Equation             Obs   Parms        RMSE    "R-sq"     F-Stat        P
--------------------------------------------------------------------------
y                  1,000       2    1.552205    0.1265      72.18   0.0000
m                  1,000       1      1.4891    0.3009     429.57   0.0000
--------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
           x |   1.405913   .1175257    11.96   0.000     1.175427    1.636399
           m |  -.1856263   .0329959    -5.63   0.000    -.2503363   -.1209163
       _cons |   6.821087    .337578    20.21   0.000     6.159044    7.483129
-------------+----------------------------------------------------------------
m            |
           x |   1.953847   .0942703    20.73   0.000     1.768968    2.138725
       _cons |   10.00209   .0681099   146.85   0.000     9.868518    10.13567
------------------------------------------------------------------------------

. nlcom  _b[m:x]* _b[y:m]

       _nl_1:  _b[m:x]* _b[y:m]

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |  -.3626854   .0668016    -5.43   0.000    -.4936142   -.2317566
------------------------------------------------------------------------------

.  sgmediation y , mv(m) iv(x)
```

Model with dv regressed on iv (path c)

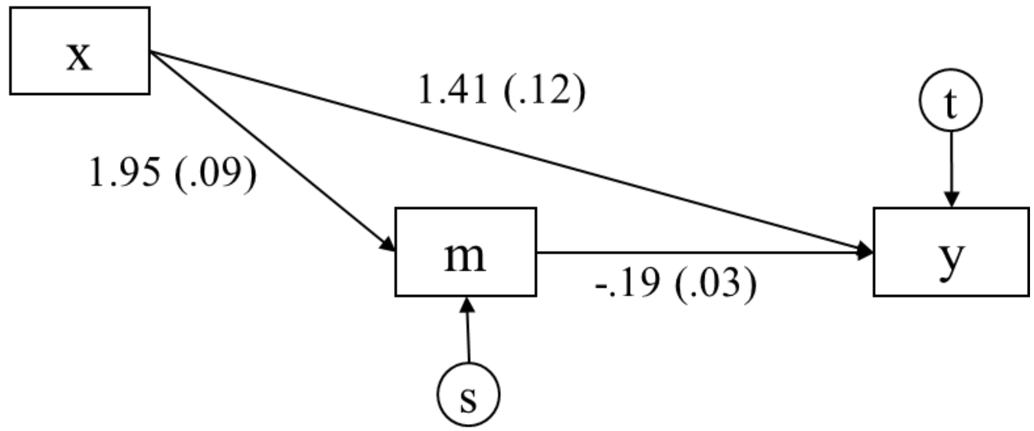
```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    109.35
       Model |  271.554254         1  271.554254   Prob > F        =    0.0000
    Residual |  2478.36475       998  2.48333141   R-squared       =    0.0987
-------------+----------------------------------   Adj R-squared   =    0.0978
       Total |    2749.919       999  2.75267167   Root MSE        =    1.5759

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.043228   .0997627    10.46   0.000      .847459    1.238996
       _cons |   4.964435   .0720781    68.88   0.000     4.822993    5.105877
------------------------------------------------------------------------------
```

Model with mediator regressed on iv (path a)

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    429.57
       Model |  952.531502         1  952.531502   Prob > F        =    0.0000
    Residual |   2212.9845       998  2.21741934   R-squared       =    0.3009
-------------+----------------------------------   Adj R-squared   =    0.3002
       Total |    3165.516       999  3.16868468   Root MSE        =    1.4891

------------------------------------------------------------------------------
           m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.953847   .0942703    20.73   0.000     1.768856    2.138837
       _cons |   10.00209   .0681099   146.85   0.000     9.868437    10.13575
------------------------------------------------------------------------------
```

Model with dv regressed on mediator and iv (paths b and c')

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =     72.18
       Model |  347.807358         2  173.903679   Prob > F        =    0.0000
    Residual |  2402.11164       997  2.40933966   R-squared       =    0.1265
-------------+----------------------------------   Adj R-squared   =    0.1247
       Total |    2749.919       999  2.75267167   Root MSE        =    1.5522

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           m |  -.1856263   .0329959    -5.63   0.000    -.2503757    -.120877
           x |   1.405913   .1175257    11.96   0.000     1.175287    1.636539
       _cons |   6.821087    .337578    20.21   0.000     6.158642    7.483532
------------------------------------------------------------------------------

Sobel-Goodman Mediation Tests

                     Coef         Std Err     Z           P>|Z|
Sobel               -.36268538    .06680162  -5.429      5.658e-08
Goodman-1 (Aroian)  -.36268538    .066874  -5.423      5.847e-08
Goodman-2           -.36268538    .06672916  -5.435      5.474e-08

                    Coef      Std Err    Z          P>|Z|
a coefficient   =  1.95385    .09427    20.726          0
b coefficient   = -.185626   .032996  -5.62574    1.8e-08
Indirect effect = -.362685   .066802  -5.42929    5.7e-08
  Direct effect =  1.40591   .117526   11.9626          0
   Total effect =  1.04323   .099763   10.4571          0

Proportion of total effect that is mediated:  -.34765697
Ratio of indirect to direct effect:           -.25797141
Ratio of total to direct effect:               .74202859
```



Incorrect parameters: PROCESS by Hayes

Estimated model 4:

Partial mediation model estimated using SPSS PROCESS v2.13. Bootstrapped estimate (k = 50,000 replications) = -.3627, Bootstrapped *SE* = .0657, and Bootstrapped LLCI = -.4962 and ULCI = -.2383. My replication of their bootstrapping procedure using a program I wrote for Stata gives a Bootstrapped *SE* = .0654, and Bootstrapped LLCI = -.4920 and ULCI = -.2355. The indirect estimate is off and not much different from those of Model 3. The correct effect must be recovered as per above using the delta method.

Let’s try with the “causal mediation” approach (Imai, Keele, & Tingley, 2010)

```
. medeff (regress m x) (regress y m x), mediate(m) treat(x) sims(1000) seed(1)
Using 0 and 1 as treatment values

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    429.57
       Model |  952.531502         1  952.531502   Prob > F        =    0.0000
    Residual |   2212.9845       998  2.21741934   R-squared       =    0.3009
-------------+----------------------------------   Adj R-squared   =    0.3002
       Total |    3165.516       999  3.16868468   Root MSE        =    1.4891

------------------------------------------------------------------------------
           m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.953847   .0942703    20.73   0.000     1.768856    2.138837
       _cons |   10.00209   .0681099   146.85   0.000     9.868437    10.13575
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =     72.18
       Model |  347.807358         2  173.903679   Prob > F        =    0.0000
    Residual |  2402.11164       997  2.40933966   R-squared       =    0.1265
-------------+----------------------------------   Adj R-squared   =    0.1247
       Total |    2749.919       999  2.75267167   Root MSE        =    1.5522

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.405913   .1175257    11.96   0.000     1.175287    1.636539
           m |  -.1856263   .0329959    -5.63   0.000    -.2503757    -.120877
       _cons |   6.821087    .337578    20.21   0.000     6.158642    7.483532
------------------------------------------------------------------------------
```

```
---------------------------------------------------------------------------------------
        Effect                   |  Mean           [95% Conf. Interval]
-------------------------------+-------------------------------------------------------
        ACME                     | -.3627823     -.4981014     -.2201736
        Direct Effect            |  1.411134       1.18156       1.65286
        Total Effect             |  1.048351      .8531834      1.238896
        % of Tot Eff mediated    | -.3455007      -.425211     -.2928276
---------------------------------------------------------------------------------------
```

Running the sensitivity analysis gives:

```
-----------------------------------------------------------------
Sensitivity results
-----------------------------------------------------------------
        Rho at which ACME = 0             |    -.1754
        R^2_M*R^2_Y* at which ACME = 0:  |     .0308
        R^2_M~R^2_Y~ at which ACME = 0:  |     .0188
-----------------------------------------------------------------
95% Confidence interval
```

"The results show that for the point estimate of the ACME to be zero, the correlation between *s* and *t* must be approximately -.18." Wait a minute: see slide 34. The correlations between *q* with *m* and *q* with *y* are 0.566 and -0.623 (average absolute correlation is .59).

Let's estimate the SEM instrumental variable model and get standardized estimate:

```
. sem (y<-m) (m<-x), cov(e.y*e.m) stand

[snip]

------------------------------------------------------------------------------
             |                 OIM
Standardized |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  y          |
           m |   .5728632   .0578491     9.90   0.000     .4594811    .6862453
       _cons |  -.2267609   .4053222    -0.56   0.576    -1.021178     .567656
  -----------+----------------------------------------------------------------
  m          |
           x |   .5485515   .0203764    26.92   0.000     .5086145    .5884885
       _cons |   5.621712   .1426161    39.42   0.000      5.34219    5.901234
-------------+----------------------------------------------------------------
     var(e.y)|   1.290193   .0593674                      1.178928    1.411959
     var(e.m)|   .6990912    .022355                       .656621    .7443084
-------------+----------------------------------------------------------------
 cov(e.y,e.m)|  -.5682898   .0390302   -14.56   0.000    -.6447877    -.491792
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

Questions

If you wish to be gently introduced to the issue of endogeneity in mediation see in order:

Jacquart, P., Cole, M. S., Gabriel, A. S., Koopman, J., & Rosen, C. C. (2017). Studying leadership: Research design and methods. In J. Antonakis & D. V. Day (Eds.), The Nature of Leadership (3 ed., pp. 411-437). Thousand Oaks: Sage.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), The Oxford Handbook of Leadership and Organizations (pp. 93-117). New York: Oxford University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. The Leadership Quarterly, 21, 1086-1120.

Antonakis, J., Brulhart, M., & Lalive, R. (2020). The endogeneity problem in mediation analysis. Working paper. Department of Organizational Behavior, University of Lausanne.

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*, 1086-1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93-117). New York: Oxford University Press.

Antonakis, J., Brulhart, M., & Lalive, R. (2017). The endogeneity problem in mediation analysis. *Working paper*. *Department of Organizational Behavior, University of Lausanne.*

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

Bollen, K. A. (2012). Instrumental Variables in Sociology and the Social Sciences. *Annual Review of Sociology, 38*(1), 37-72.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1-22.

Fischer, T., Dietz, J., & Antonakis, J. (2017). Leadership process model: A review and synthesis. *Journal of Management, 43*(6), 1726-1753.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251-1271.

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

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

Kline, R. B. (2015). The Mediation Myth. *Basic and Applied Social Psychology, 37*(4), 202-213.

Oehlert, G. W. (1992). A note on the Delta Method. *The American Statistician, 46*, 27-29.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717-731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. [Article]. *Behavior Research Methods, 40*(3), 879-891.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shaver, J. M. (2005). Testing for mediating variables in management research: Concerns, implications, and alternative strategies. *Journal of Management, 31*(3), 330-353.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods, 7*(4), 422.

Smith, E. R. (2012). Editorial. *Journal of Personality and Social Psychology*, *102*(1), 1-3.

Sobel, M. E. (1982). Asymptotic Intervals for Indirect Effects in Structural Equations Models. In S. Leinhart (Ed.), *Sociological Methodology* (pp. 290−312). San Francisco: Jossey-Bass.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research, 37*(2), 197-206.