Online Appendix to "Yes, But What's the Mechanism? (Don't Expect an Easy Answer)"

John G. Bullock, Donald P. Green, and Shang E. Ha

January 18, 2010

This appendix has six parts:

1. Proof that $ab = c - d$ in Equations 1 through 3.
2. Proof that OLS estimators of b and d in Equation 3 are prone to bias in infinitely large samples.
3. Proof that experimental mediation analyses are prone to bias if there are fewer experimental interventions than mediators affected by the interventions.
4. Proof that experimental estimates of indirect effects apply only to subjects who (a) are affected by the experimental interventions or (b) would be affected by the experimental interventions if they were exposed to those interventions.
5. Proof that experimental mediation analyses cannot identify average indirect effects in the presence of causal heterogeneity.
6. SPSS code to demonstrate that experimental mediation analyses cannot identify average indirect effects in the presence of causal heterogeneity.

The proofs that we present here do not break new ground. All of them may be found in other works, albeit in contexts that may make their relevance to mediation analysis unclear. To guide readers who want more detailed discussion of the issues presented here, we refer to these other works throughout this appendix.

1. Equivalence of ab and c − d in Equations 1 through 3

Let $X = X_1, \ldots, X_n$, $Y = Y_1, \ldots, Y_n$, $e_1 = e_{11}, \ldots, e_{n1}$, $e_2 = e_{12}, \ldots, e_{n2}$, and $e_3 = e_{13}, \ldots, e_{n3}$. We assume that X has been randomized such that $X \perp\!\!\!\perp e_1, e_2, e_3$. Proof by direct calculation: substituting Equation 1 into Equation 3 yields

$$Y = (b\alpha_1 + \alpha_3) + (ab + d)X + (be_1 + e_3)\,,$$

from which we see that

$$\mathrm{cov}(X, Y) = (ab + d)\mathrm{var}\,(X)$$

$$\Rightarrow ab = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}\,(X)} - d.$$

Inspection of Equation 2 shows that $c = \mathrm{cov}(X, Y)\,/\mathrm{var}\,(X)$, completing the proof. See MacKinnon, Warsi, and Dwyer (1995, 45-46) for a similar treatment.

2. OLS Estimators of b and d in Equation 3 Are Inconsistent

Consider the equation

$$Y = d\tilde{X} + b\tilde{M} + e_3, \tag{A1}$$

where $\tilde{X} = X - \overline{X}$, $\tilde{M} = M - \overline{M}$, $X = X_1, \ldots, X_n$, $M = M_1, \ldots, M_n$, and $e_3 = e_{13}, \ldots, e_{n3}$. Let $e_1 = e_{11}, \ldots, e_{n1}$ and $e_2 = e_{12}, \ldots, e_{n2}$. We assume that X has been randomized such that $X \perp\!\!\!\perp e_1, e_2, e_3$, and by extension, $\tilde{X} \perp\!\!\!\perp e_1, e_2, e_3$. Biases in the OLS estimators of d and b in Equation A1 are the same as the biases for the OLS estimators of d and b in Equation 3. Following Muller et al. (2005), we use the mean-centered predictors of Equation A1, which make for easier interpretation and simplify the calculations.

Let $\tilde{\mathbf{X}}$ be the design matrix $[\tilde{X}, \tilde{M}]$. The OLS estimators are

$$\begin{bmatrix} \hat{d} \\ \hat{b} \end{bmatrix} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(d\tilde{X} + b\tilde{M} + e_3)$$

$$= \begin{bmatrix} d \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ b \end{bmatrix} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'e_3, \text{ where}$$

$$\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'e_3 = \begin{bmatrix} \frac{\sum {e_1}_i^2 \sum \tilde{X}_i e_{3i} + a \sum \tilde{X}_i e_{1i} \sum \tilde{X}_i e_{3i} - \sum e_{1i}\tilde{X}_i \sum e_{1i}e_{3i} - a \sum \tilde{X}_i^2 \sum e_{1i}e_{3i}}{\sum {e_1}_i^2 \sum \tilde{X}_i^2 - \sum^2 e_{1i}\tilde{X}_i} \\ \\ \frac{\sum e_{1i}e_{3i} \sum \tilde{X}_i^2 - \sum e_{1i}\tilde{X}_i \sum \tilde{X}_i e_{3i}}{\sum {e_1}_i^2 \sum \tilde{X}_i^2 - \sum^2 e_{1i}\tilde{X}_i} \end{bmatrix}.$$

Now,

$$\text{plim}[\hat{b}] = b + \text{plim}\left[\frac{\sum e_{1i}e_{3i} \sum \tilde{X}_i^2 - \sum e_{1i}\tilde{X}_i \sum \tilde{X}_i e_{3i}}{\sum {e_1}_i^2 \sum \tilde{X}_i^2 - \sum^2 e_{1i}\tilde{X}_i}\right]$$

$$= b + \frac{\text{cov}(e_1, e_3)}{\text{var}\,(e_1)}.$$

And by the same logic,

$$\text{plim}[\hat{d}] = d - \text{plim}\left[\frac{a \sum e_{1i}e_{3i}}{\sum {e_1}_i^2}\right] = d - a \cdot \frac{\text{cov}(e_1, e_3)}{\text{var}\,(e_1)}.$$

Thus, whenever $\text{cov}(e_1, e_3) \neq 0$, ordinary-least-squares estimators of *d* and *b* in Equation 3 are biased even in infinitely large samples.

Rosenbaum (1984) offers a wide-ranging discussion of the problem. Gelman and Hill (2007, 188-94) provide an intuitive treatment.

3. *Experimental Mediation Analyses Are Prone to Bias If Interventions Do Not Isolate Particular Mediators*

For simplicity, consider a model in which a treatment X is mediated by exactly two variables:

$$Y = \alpha_3 + dX + b_1 M_1 + b_2 M_2 + e_3,$$

where e_3 is a mean-zero error term that represents the cumulative effect of omitted variables. We assume that X has been randomly assigned such that $X \perp\!\!\!\perp e_3$. Assume that a random intervention Z affects M_1 but that it is uncorrelated with other variables that affect Y ($Z \perp\!\!\!\perp X, M_2, e_3$). The latter assumption is the *exclusion restriction*, which is required for instrumental-variables estimation of causal effects (e.g., Wooldridge, 2008, ch. 15).

Given these assumptions,

$$\begin{aligned} \operatorname{cov}(Z, Y) &= b_1 \operatorname{cov}(Z, M_1) \\ \Rightarrow b_1 &= \frac{\operatorname{cov}(Z, Y)}{\operatorname{cov}(Z, M_1)}. \end{aligned}$$

We can use the sample covariances to calculate a consistent estimator of b_1,

$$\widehat{b_1} = \frac{\widehat{\operatorname{cov}}(Z, Y)}{\widehat{\operatorname{cov}}(Z, M_1)}. \tag{A2}$$

$\widehat{b_1}$ is the traditional instrumental-variables estimator; in instrumental-variables parlance, Z is an instrument for M_1. But if Z affects both M_1 and M_2,

$$\begin{aligned} \operatorname{cov}(Z, Y) &= b_1 \operatorname{cov}(Z, M_1) + b_2 \operatorname{cov}(Z, M_2) \\ \Rightarrow b_1 &= \frac{\operatorname{cov}(Z, Y)}{\operatorname{cov}(Z, M_1)} - \frac{b_2 \operatorname{cov}(Z, M_2)}{\operatorname{cov}(Z, M_1)}. \end{aligned} \tag{A3}$$

To estimate b_1 in this case, we must estimate the right-hand side of Equation A3 with quantities (e.g., sample covariances) that can be computed from the observed values of X, M_1, M_2, and Z. We cannot do this because b_2 is unknown and cannot be estimated with the data at hand. In particular, the traditional estimator given in Equation A2 is biased in infinite samples by $-\frac{b_2 \text{cov}(Z,M_2)}{\text{cov}(Z,M_1)}$.

The experimental approach can be extended to account for multiple hypothesized mediators. But in this case, one must have at least as many instruments as hypothesized mediators: this is part of the *rank condition* for estimating effects with instrumental variables (Koopmans, 1949; Wooldridge, 2002, 85-86). To avoid problems stemming from mixtures of local average treatment effects (see Morgan & Winship, 2007, 212), we further recommend that each experimentally created instrument be crafted to affect exactly one mediator: Z_1 should affect only M_1, Z_2 should affect only M_2, and so forth. But see Angrist and Imbens (1995; also Angrist & Pischke, 2009, 173-75) for a defense of conventional practice ("two-stage least squares"), which does not demand that each instrument affect only one mediator.

On the use of instrumental variables to estimate indirect effects, see Gennetian, Morris, Bos, and Bloom (2005). There are many general treatments of instrumental-variables estimation; we recommend Angrist et al. (1996), Morgan and Winship (2007, ch. 7), and Wooldridge (2008, ch. 15).

4. *Experimental Estimates of Indirect Effects Apply Only to Subjects Who Are Affected by the Experimental Intervention Or Who Would Be Affected By It If They Were Exposed to It*

Let M_i, the mediator from Equation 3, be a dummy variable. Let Z_i be a dummy variable indicating whether i has been exposed to an intervention designed to change his value of M ($Z_i = 1$) or has not been exposed to such an intervention ($Z_i = 0$). Let $M_i(1)$ be the value of M_i when $Z_i = 1$; similarly, let $M_i(0)$ be the value of M_i when $Z_i = 0$. Let $Y_i(m, z)$ be the value of Y_i if

$M_i = m$ and $Z_i = z$. For example, $Y_i(1,1)$ is the value of Y_i if $M_i = 1$ and $Z_i = 1$. $M_i(1), M_i(0)$, and $Y_i(m,z)$ are *potential outcomes*.

Assume that:

1. Z_i is independent of the potential outcomes ($Z_i \perp\!\!\!\perp M_i(0), M_i(1), Y_i(m,0), Y_i(m,1) \; \forall \; i$).
2. Z affects M ($E\left[M_i(1) - M_i(0)\right] \neq 0$).
3. Z_i satisfies the exclusion restriction. In the context of Equation 3, this implies $Z_i \perp\!\!\!\perp X_i, e_{i3} \; \forall \; i$.

Barring randomization problems, the first two assumptions are likely to be met by any randomized intervention that is designed to affect M. As we show in Part 3 of this appendix, the third assumption is also necessary if we are to use experiments to identify indirect effects. Note that the third assumption implies that $Y_i(1,1) = Y_i(1,0)$ and $Y_i(0,1) = Y_i(0,0)$. We now simplify the notation by denoting these potential outcomes $Y_i(1)$ and $Y_i(0)$.

Now, let $M_i = \beta_0 + \beta_1 Z_i + \epsilon_i$, where $Z_i \perp\!\!\!\perp \epsilon_i$ and $E\left[\epsilon_i\right] = 0$. Combining this with Equation 3, we have $Y_i = \alpha_3 + dX + b\left(\beta_0 + \beta_1 Z_i + \epsilon_i\right) + e_{i3}$. It follows that

$$\operatorname{cov}(Z_i, Y_i) = b\beta_1 \operatorname{var}(Z_i) = \{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]\}\operatorname{var}(Z_i)\,, \text{ and}$$

$$\operatorname{cov}(Z_i, M_i) = \beta_1 \operatorname{var}(Z_i) = \{E[M_i|Z_i = 1] - E[M_i|Z_i = 0]\}\operatorname{var}(Z_i)\,.$$

From these equations, we see that

$$b = \frac{\operatorname{cov}(Z_i, Y_i)}{\operatorname{cov}(Z_i, M_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[M_i|Z_i = 1] - E[M_i|Z_i = 0]}. \tag{A4}$$

We can rewrite the numerator of Equation A4:

$$
\begin{aligned}
E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= E[Y_i(0) + (Y_i(1) - Y_i(0))M_i|Z_i = 1] - E[Y_i(0) + (Y_i(1) - Y_i(0))M_i|Z_i = 0] \\
&= E[Y_i(0) + (Y_i(1) - Y_i(0))M_i(1)] - E[Y_i(0) + (Y_i(1) - Y_i(0))M_i(0)] \\
&= E[(Y_i(1) - Y_i(0))\,(M_i(1) - M_i(0))].
\end{aligned}
$$

At this point, we need to invoke the *monotonicity assumption* (e.g., Angrist et al., 1996): the experimental intervention does not increase the value of the mediator for some subjects while decreasing it for others. Formally, $M_i(1) \geq M_i(0)$ for all subjects or $M_i(0) \leq M_i(1)$ for all subjects. Unlike some of the assumptions that we discuss in our article, we do not think that this assumption is troubling: it is likely to be met in most psychological applications.

Without loss of generality, assume $M_i(1) \geq M_i(0)$. Then $M_i(1) - M_i(0)$ is 0 or 1 for all subjects. It follows that the numerator in Equation A4 is

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[(Y_i(1) - Y_i(0))\,|M_i(1) > M_i(0)] \cdot \Pr[M_i(1) > M_i(0)].$$

Now turn to the denominator of the right-hand side of Equation A4. We have

$$
\begin{aligned}
E[M_i|Z_i = 1] - E[M_i|Z_i = 0] &= E[M_i(1) - M_i(0)] \\
&= \Pr[M_i(1) > M_i(0)].
\end{aligned}
$$

It follows that

$$b = \frac{\operatorname{cov}(Z_i, Y_i)}{\operatorname{cov}(Z_i, M_i)} = E[(Y_i(1) - Y_i(0))\,|M_i(1) > M_i(0)].$$

The coefficient *b* is therefore the average effect of *M* on *Y* for “compliers” alone—i.e., for subjects whose value of *M* would be changed by exposure to *Z*. Estimators of *b* estimate this "local average treatment effect," not an average treatment effect for all subjects. It follows that estimates of the indirect effect, *ab*, will apply to compliers if and only if *a* is the effect of *X* on *M* for this subgroup.

On local average treatment effects, see Imbens and Angrist (1994) and Sobel (2008, pp. 244-47). (Angrist et al., 1996, 451) show that if the monotonicity assumption is violated, the estimand (e.g., *b* in Equation 3) can be viewed as a weighted average of the average treatment effect for compliers and the average treatment effect for "defiers."

A particular limitation of LATE estimation is that one cannot know exactly who the compliers are. But it is possible to use summary statistics (e.g., percentage women, percentage white) to characterize the population of compliers: see Angrist and Pischke (2009, 166-72).

5. *Randomization Cannot Identify Average Indirect Effects in the Presence of Causal Heterogeneity*

Let $Y_i = c_i X_i + e_{i1}$. Y_i is the outcome of interest for subject i, $X_i \in \{0, 1\}$ is a treatment, c_i is the effect of X_i on Y_i, and e_{i1} represents the cumulative effect of other variables. (Intercepts are redundant in this model: $e_{i1} = \alpha_i + e^*_{i1}$, where α_i is an intercept.) The effect of X_i may vary from subject to subject, in which case $c_i \neq c_j$ (for $i \neq j$).

When $X_i = 1$, we denote the value of Y_i as $Y_i(1)$. And when $X_i = 0$, we denote the value of Y_i as $Y_i(0)$. The effect of X_i on Y_i is $c_i = Y_i(1) - Y_i(0)$. We cannot observe both $Y_i(1)$ and $Y_i(0)$ for any subject—this is the "fundamental problem of causal inference" (Holland, 1986, 947)—and we therefore cannot observe c_i for any subject. But if we randomly assign values of X, we *can* estimate the average effect of X, $\bar{c} = E\left[Y_i(1) - Y_i(0)\right]$, without bias. We do this by observing the average $Y_i(1)$ for the treatment group, $\overline{Y_i(1)|X = 1}$, and the average $Y_i(0)$ for the control group, $\overline{Y_i(0)|X = 0}$. This lets us calculate

$$\overline{Y_i(1)|X = 1} - \overline{Y_i(0)|X = 0}.$$

And if X is independent of $Y_i(1)$ and $Y_i(0)$ (as is usually the case when X is randomized),

$$\begin{aligned} E\left[\overline{Y_i(1)|X=1} - \overline{Y_i(0)|X=0}\right] &= E\left[\overline{Y_i(1)}\right] - E\left[\overline{Y_i(0)}\right] \\ &= E\left[\overline{Y_i(1) - Y_i(0)}\right] \\ &= \bar{c}. \end{aligned}$$

However, we cannot use experiments to identify average indirect effects when those effects vary. Let $M_i = a_i X_i + e_{i2}$ and $Y_i = d_i X_i + b_i M_i + e_{i3}$. $M_i(1)$ and $M_i(0)$ are the values that M_i assumes when $X_i = 1$ and $X_i = 0$, respectively. The value of Y_i when $X_i = 1$ and $M_i = M_i(1)$ is

$$\begin{aligned} Y_i\left(1, M_i(1)\right) &= d_i + b_i M_i(1) + e_{i3} \\ &= d_i + b_i\left(a_i + e_{i2}\right) + e_{i3}. \end{aligned}$$

The value of Y_i when $X_i = 1$ and $M_i = M_i(0)$ is

$$\begin{aligned} Y_i\left(1, M_i(0)\right) &= d_i + b_i M_i(0) + e_{i3} \\ &= d_i + b_i\left(e_{i2}\right) + e_{i3}. \end{aligned}$$

$Y_i\,(1, M_i(0))$ is the value that Y_i would assume if $X_i = 1$ but M_i took on the value that it would have if $X_i = 0$. We cannot observe this quantity for any individual because we cannot assign X_i to simultaneously equal 1 and 0. Moreover, experiments cannot produce unbiased averages of this quantity: the quantity is counterfactual, unknowable even in principle. This is problematic because the indirect effect of X_i involves this quantity. Specifically, the indirect effect is the change in Y_i that we would observe if we held X_i constant at 1 but changed the mediator from $M_i(0)$ to $M_i(1)$:

$$\begin{aligned} Y_i\,(1, M_i(1)) - Y\,(1, M_i(0)) &= d_i + b_i\,(a_i + e_{i2}) + e_{i3} - [d_i + b_i\,(e_{i2}) + e_{i3}] \\ &= a_i b_i.^{1} \end{aligned}$$

We cannot observe a_i or b_i for any individual. If we conduct an experiment in which only X is manipulated, we can estimate $\bar{a} = \sum a_i/n$, the average effect of X on M. And if we conduct an experiment in which M and X are manipulated, we can estimate $\bar{b} = \sum b_i/n$, the average effect of M on Y while holding X constant. But by the laws of covariance, the product of these averages does not generally equal the average indirect effect. Instead,

$$E[a_i]E[b_i] = E[a_i b_i] - \text{cov}(a_i, b_i)\,.$$

In words, the product method will produce biased estimates of average indirect effects, and the bias will equal the covariance of a and b. See Glynn (2009, p. 10-13) for a demonstration that the

[1] When $X_i = 0$, the indirect effect of X_i is $Y_i(0, M_i(1)) - Y_i(0, M_i(0))$. Almost all mediation analyses implicitly assume that this quantity is equal to the indirect effect when $X_i = 1$, but there is typically no empirical reason to make this “no-interaction” assumption: see Robins, 2003, 76-77; Sobel, 2008. For simplicity, we focus here on direct and indirect effects when $X_i = 1$, but the same problem obtains when $X_i = 0$.

same result holds when the “difference method” is used to compute the average indirect effect in the presence of causal heterogeneity.

6. SPSS Code for Simulations Demonstrating Bias in Mediation Analysis Caused by Heterogeneous Treatment Effects

```
* Create ID numbers from 1 to 10000.
INPUT PROGRAM.
LOOP id=1 TO 10000.
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.

* Generate a moderator variable called q.
* Change 5000 to some other number in order to change the distribution of q.
RECODE id (0 thru 5000=0)(else=1) into q.

* Generate normal disturbances for the two equations.
COMPUTE e = RV.NORMAL(0,1) .
COMPUTE u = RV.NORMAL(0,1) .

* Generate uniformly distributed independent variable x.
COMPUTE x = RV.UNIFORM(0,1) .

* Generate a mediator variable m that is a function of x, q, and an interaction.
COMPUTE m = q*(x+e) + (1-q)*(-x+e) .

* Generate a dependent variable y that is a function of m, q, and an interaction.
* Note that x has no direct effect on y.
COMPUTE y = q*(m+u) + (1-q)*(-m+u) .

* This regression correctly estimates the average total effect of x on y.
```

```
REGRESSION /DEPENDENT y /METHOD=ENTER x .

* This regression correctly estimates the average direct effect of x on m.
REGRESSION /DEPENDENT m /METHOD=ENTER x .

* This regression incorrectly estimates the direct effect of x on y and the direct
* effect of m on y.  Recall that x in fact has no direct effect on y, but the
* regression says otherwise.  Moreover, this regression misestimates the direct
* effect of m on y, declaring that m has no direct effect.
REGRESSION /DEPENDENT y /METHOD=ENTER m x .

* Show that one obtains unbiased results when one partitions the sample such
* that there are no heterogeneous effects within subgroups:

    * Rerun the last regression for the subsample where q=1.
    REGRESSION /SELECT= q EQ 1 /DEPENDENT y /METHOD=ENTER m x .

    * Rerun the last regression for the subsample where q=0.
    REGRESSION /SELECT= q EQ 0 /DEPENDENT y /METHOD=ENTER m x .
```

References

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association, 90,* 431-42.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444-55.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gennetian, L. A., Morris, P. A., Bos, J. M., & Bloom, H. S. (2005). Constructing instrumental variables from experimental data to explore how treatments produce effects. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 75-114). New York: Russell Sage.

Glynn, A. N. (2009). "The product and difference fallacies for indirect effects." Harvard University. Manuscript.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-60.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62,* 467-75.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica, 17,* 125-44.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30,* 41-62.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference*. New York: Cambridge University Press.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89,* 852-863.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70-81). New York: Oxford University Press.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, 147,* 656-66.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics, 33,* 230-51.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2008). *Introductory econometrics: A modern approach*. 4th ed. Mason, OH: South-Western Cengage Learning.