Forthcoming, Psychological Methods (2014) with discussions of "Interpretation and Identification of Causal Mediation," R-389.

TECHNICAL REPORT
R-421
June 2014

Reply to Commentary by Imai, Keele, Tingley, and Yamamoto (2014) Concerning Causal Mediation Analysis

Judea Pearl*
Computer Science Department
University of California, Los Angeles
Los Angeles, CA, 90095-1596
judea@cs.ucla.edu
(310) 825-3243 Tel / (310) 794-5057 Fax

June 5, 2014

Abstract

This comment clarifies how Structural Causal Models (SCM) unify the graphical and potential outcome approaches to mediation, and why the resulting mediation formulas are identical in both frameworks. It further explains under what conditions ignorability based assumptions are over-restrictive and why such assumptions require graphical interpretations before they can be judged for plausibility. Finally, the comment explains the key difference between traditional and modern methods of causal mediation, and demonstrates why the notion of mediation requires counterfactual rather than Bayes conditionals to be properly defined.

Key words: Mediation formula, seeing vs. doing, sequential ignorability, graphical methods, Structural causal models, counterfactuals.

I am happy to join Imai, Keele, Tingley, and Yammamoto (henceforth Imai-et al.) in celebrating the full convergence of our respective analyses towards a unified understanding of causal mediation. I am referring to the analysis presented in (Pearl, 2001) (reproduced in (Pearl, 2014a)) on the one hand, and the analyses and implementations of (Imai et al., 2010a,b,c), on the other. In fact, when I first read (Imai et al., 2010c), I had no doubt that,

*This commentary has benefited from discussions with Kosuke Imai, David Kenny, and Bengt Muthén. I am grateful to Associate Editor, Patrick Shrout for giving me the opportunity to reply to this commentary. This research was supported in parts by grants from National Institutes of Health #1R01 LM009961-01, National Science Foundation #IIS-0914211 and #IIS-1018922, and Office of Naval Research #N000-14-09-1-0665 and #N00014-10-1-0933.

despite some dissimilarities in the presentation of the assumptions, the two works would coincide on all fronts: Definitions, basic assumptions, identification and estimation algorithms. The reasons for my confidence was that, in 2001, I approached the mediation problem from the symbiotic mathematical framework of Structural Causal Models (SCM) (Pearl, 2000, Chapter 7; Pearl, 2009a) which unifies the graphical, potential outcome and structural equation frameworks, and permits researchers to combine the merits of each representation; structural equations and graphical models best represent what a researcher believes, while potential outcomes represent what a researcher seeks to estimate.

A logical analysis of SCM theory further revealed that structural equations and potential outcomes are logically equivalent; a theorem in one is a theorem in the other. They differ only in the language in which assumptions are cast; structural equations cast assumptions in the language in which scientific knowledge is stored, while potential outcomes cast those same assumptions in terms of quantities that one wishes to estimate (e.g., counterfactuals). This means that any researcher who accepts the potential outcome framework can use the power of graphs and structural equations for advantage and be assured the validity of the result. This also means that the power of graphs lies not merely in their clarity of visualizing assumptions, but also in "computing" complex implications of those assumptions. Typical implications are: conditional independencies among variables and counterfactuals, what covariates need be controlled to remove confounding or selection bias, whether effects can be identified, and more. (Praising their transparency while ignoring their inferential power misses the main role that graphs play in modern causal analysis.)

Armed with these symbiotic tools, I derived identification conditions in the algebra of counterfactuals and presented them in two languages, potential outcomes and graphical. Not surprisingly, the mediation formulas derived in Imai et al. (2010c) coincide precisely with those derived in Pearl (2001, Eqs. (8), (17), (26), (27)). This is to be expected, since the two are but variants of the same mathematical umbrella, differing merely in the type of assumptions one is willing to posit and defend, and the language one chooses to communicate the assumptions.

The assumptions posited in Imai et al. (2010c) added two restrictions to those articulated in (Pearl, 2001):

1. Commence the analysis with two *ignorability* assumptions (B-1 and B-2 in the main paper). (The latter is automatically satisfied in randomized studies.)

2. Satisfy these two assumptions with the *same* set (W) of observed covariates.

Clearly, all identification results produced under these restrictions will be valid in the symbiotic system of SCM (Pearl, 2001), in which these restrictions were not imposed.

In (Pearl, 2014a) I identify the set of circumstances where these two added restrictions lead to missed opportunities, and the current commentary by Imai-et al. identify conditions under which the added restrictions will cause no practical loss of opportunities. The two studies complement each other and provide valuable information; they tell us when the inference systems of (Imai et al., 2010a,b,c) operate in perfect harmony with the methodology presented in (Pearl, 2001).

Specifically, Imai-et al. show that the restrictions imposed by sequential ignorability play a role only in observational studies, but not in studies where treatment is randomized.

Additionally, the extra-restriction of conditioning on the *same* set of covariates may not be too severe in certain observational studies. I concur with most of these observations, and commend Imai-et al. for bringing them to readers attention.

I cannot accept, however, their conclusions that: "Including irrelevant covariates may complicate the modeling but does not compromise the identification of causal mediation effects under the as-if randomization assumption" (Imai et al., 2014, this issue) Whether covariates are relevant or irrelevant depends on whether the "as-if randomization assumption" holds after their inclusion, which makes the sentence above circular, if not contradictory. The "as if randomized" assumption can easily be violated by including what may appear to be "irrelevant" pre-treatment covariates.[1] Moreover, the validity of the "as-if randomization assumption" may depend on many other assumptions encoded in the model, hence no mortal can judge its plausibility without the aid of graphs.[2] Fortunately, the graphical procedure presented in my paper (Pearl, 2014a) allow us to mechanize the choice of the "relevant covariates," and I hope Imai-et al. can implement this procedure in their flexible software. A prerequisite for accomplishing this function is to let users articulate assumptions in the language of scientific understanding, namely graphs, and let estimation procedures and covariate selection be derived (mechanically) from those assumptions, rather than chosen a priori.

In the remaining of this note, I concentrate on an issue that is common to all players in the causal mediation analysis. It concerns ways of improving the understanding of causal mediation among the uninitiated.

Impediments to such understanding come from several research communities.

1. Potential outcomes enthusiasts reject mediation when the mediator is non-manipulable.

2. Traditional statisticians fear that, without extensive reading of Aristotle, Kant and Hume, they are not well equipped to tackle the subject of causation, especially when it involves claims based on untested assumptions.

3. Traditional mediation analysts do not understand the sudden intrusion of counterfactuals into their field, which thus far has been dominated by regression analysis.

4. Economists, who adore counterfactuals (though find difficulties defining them (Pearl, 2009b, p. 379)) are not convinced that mediation analysis could help policy makers.

I will address the third group, namely, the traditional mediation analysts usually connected with the school of Baron and Kenny (BK) (1986), since the difficulties faced by this school are endemic of other groups as well, and constitute the key impediment to a wider acceptance of causal mediation. As traditionalists examine modern definitions of direct and

[1]For a lively discussion concerning the harm of including seemingly "irrelevant covariates" see (Pearl, 2009c; Rubin, 2009; Shrier, 2009; Sjölander, 2009). The collider X in Figure 9 of the main paper (Pearl, 2014a) is an example of a covariate that would compromise identificationd if included in the analysis (assuming a randomized treatment).

[2]Skeptics are invited to guess whether $M_t \perp\!\!\!\perp T|Y$ holds in the model of Figure 1A, namely, whether the effect of T on M is ignorable conditional on Y. Graphs replace such formidable mental tasks with transparent scientific judgements on whether the graph structure is plausible, followed by a simple test for the backdoor criterion (see Pearl 2014a, Appendix A).

indirect effects, even those based on structural equations (e.g., Eqs. (7)–(10) in the main paper), the thing that strikes them odd is the absence of a conditioning operator in any of these definitions. Whereas in the linear SEM tradition "effects" are associated with conditional expectations or regression slopes conditioned on holding some variables constant, here, we plug the value of the variables we wish to keep constant (or "control for") directly into the equation (or into the subscript of a counterfactual), but we never place that variable behind a conditioning bar. In other words, we write $E\{f_Y[1, M = m]\}$ or $E[Y_{1,m}]$ but not $E(Y|T = 1, M = m)$.

Readers versed in the distinction between "seeing" vs. "doing" (Lindley, 2002; Pearl, 1993; Pearl, 2009b, pp. 421–428; Spirtes et al., 1993) or "controlling for" vs. "setting" will recognize immediately that, in mediation, the proper operator is "doing," not "seeing" and that it is this difference that gives causal mediation analysis a claim to the title "causal." Most traditionalists, however, are not attuned to this distinction and, when presented with the modern definitions of direct and indirect effect tend to voice skepticism: "Do we really need those counterfactuals?" or "Do we really need to treat a structural equation in this manner? Why not condition on $M = m$?"

The urge to condition on variables held constant is in fact so intense that I hold it accountable for a century of blunders and confusions; from "probabilistic causality" (Suppes, 1970; [Pearl, 2011b]) to "evidential decision theory (Jeffrey, 1965; [Pearl, 2009b, pp. 108–109]) and Simpson's paradox (Simpson, 1951; [Pearl, 2009b, pp. 173–180; Pearl, 2014b]); from Fisher's error in handling mediation (Fisher, 1935; [Rubin, 2005]) to "Principal Stratification" mishandling of mediation (Rubin, 2004; [Pearl, 2011a]) from misinterpretations of structural equations (Freedman, 1987; Hendry, 1995; Holland, 1995; Sobel, 2008; Wermuth, 1992; [Bollen and Pearl, 2013; Pearl, 2009b, pp. 135–138]) to the structural-regressional confusion in econometric textbooks today ([Chen and Pearl, 2013]).[3]

What caused this confusion, and how did it enter the world of mediation? The urge to condition stems from the absence of probabilistic notation for the notion of "holding M constant," which has forced generations of statisticians to use a surrogate in the form of "conditioning on M''; the only surrogate licensed to them by probability theory.

The history of mediation analysis offers a compelling narrative on why the conditioning habit took roots, and why it should be uprooted.

Examine the basic mediation model (Figure 1A) with M (partially) mediating between T and Y. Why are we tempted to "control" for M when we wish to estimate the *direct* effect of T on Y? The reason is that, if we succeed in preventing M from changing then whatever changes we measure in Y would be attributable solely to variations in T and we would be justified then in proclaiming the response observed as "direct effect of T on Y." Unfortunately, the language of probability theory does not possess the notation to express the idea of "preventing M from changing" or "physically holding M constant." The only operator probability allows us to use is "conditioning" which is what we do when we "control for M" in the conventional way. In other words, instead of physically holding M constant (say at $M = m$) and comparing Y for units under $T = 1$ to those under $T = 0$, we allow M to vary but ignore all units except those in which M achieves the value $M = m$. Students of

[3]In this paragraph, the unbracketed citations refer to articles where confusions are present, while bracketed citations refer to articles where confusions are unveiled or resolved.

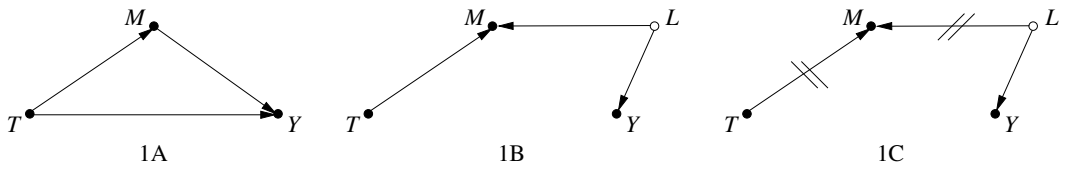

Figure 1: Demonstrating the difference between "controlling for M" and "fixing M." 1A: the classical mediation model. 1B: A model where the direct effect of T on Y is zero and, yet, "controlling for" M would yield a non-zero difference between units under $T = 0$ and those under $T = 1$. 1C: "Fixing" M amounts to overruling the influences of T and L on M, leading to correct estimate of the direct effect ($= 0$).

causality know that these two operations are profoundly different, and give totally different results, except in the case of no omitted variables. Yet to most traditionalists, this would come as a total surprise, and would elicit requests for explicit demonstration. Stunned by the cultural divide between the two camps, and having not found a convincing demonstration in the literature,[4] I believe it is appropriate to provide one at this commentary; it is absolutely pivotal to the understanding of causal mediation.

Assume that there is a latent variable L causing both M and Y as shown in Figure 1B. To simplify the discussion, assume further that the structural equations are $Y = 0 \cdot T + 0 \cdot M + L$ and $M = T + L$. Obviously, the direct effect of T on Y in this case is zero, but this is not what we would get if we "control for M" and compare subjects under $T = 1$ to those under $T = 0$ at the same level of $M = 0$. In the former group we would find $Y = L = M - T = 0 - 1 = -1$ whereas in the latter group we would find $Y = L = M - T = 0 - 0 = 0$. In other words, in order to keep the same score of $M = 0$ for the two groups, L had to change from $L = -1$ to $L = 0$. Thus, we are unwittingly comparing apples and oranges (i.e., subjects for which $L = -1$ to those with $L = 0$) and, not surprisingly, we obtain an erroneous estimate of (-1) for a direct effect that, in reality is zero.

Now let us examine what we obtain from the counterfactual expression

$$CDE(M) = E[Y(1, M)] - E[Y(0, M)]$$

for $M = 0$ (same for $M = 1$). Substituting the structural equation for the counterfactuals, we get

$$\begin{aligned} CDE(M = 0) &= E[Y(1, 0)] - E[Y(0, 0)] \\ &= E[0 \cdot 1 + 0 \cdot 0 + L] - E[0 \cdot 0 + 0 \cdot 0 + L] \\ &= E[L - L] = 0 \end{aligned}$$

as expected. The reason we obtained the correct result is that we simulated correctly what we set out to do, namely, to physically hold M constant, rather than condition on M. In

[4]The inappropriateness of conditioning on a mediator is demonstrated in (Pearl, 1998; Robins and Greenland, 1992) and by many authors since. The demonstration provided below, however, is algebraic and may be more convincing to researchers new to graphical modeling.

the former case L remains unchanged, because the physical operation of holding M constant and changing T does not affect L. In the latter, when we "condition" on a constant M, L must compensate for varying T to satisfy the equation $M = T + L$. In short, counterfactual conditioning reflects a physical intervention while statistical conditioning reflects filtered observation. To avoid confusion between the two, I used the notation $E[Y|do(T = t)]$ as distinguished from ordinary conditional expectation, $E[Y|T = t]$ (Pearl, 2009b, Chapter 3).

The habit of translating "hold M constant" into "condition on M" became deeply entrenched in the statistical culture (see Lindley, 2002; Pearl, 1993; Spirtes et al., 1993), not by deliberate negligence but due to the coarseness of its language (probability theory) which fails to provide an appropriate operator for "holding M constant." Absent such operator, statisticians (including Fisher (1935)) were pressed to use the only operator available to them: conditioning, and a century of confusion came into being.

Traditional mediation analysts of the BK school were not unaware of the dangers lurking from conditioning (Judd and Kenny, 1981, 2010). However, lacking an appropriate operator for "fixing M," they settled on a compromise; they defined the direct effect as

$$c' = E[Y|T = 1, M = 0)] - E[Y|T = 0, M = 0)]$$

and accompanied this definition with a warning that it is valid only under the assumption of "no omitted variables."

Causal analysis circumvents this compromise upon realizing that the operator needed for "fixing M," while undefinable in probability theory, is well defined in SEM, both parametric and nonparametric, through the $do(M = m)$ operator. It calls for modifying the model by replacing the equation that determines M with a constant $M = m$, and keeping all other equations unaltered (Balke and Pearl, 1995; Pearl, 1993). This "surgical" operator permits researchers to state their intent using expressions such as $E(Y|do(M = m))$ or $Y(1, M)$, yielding $CDE(M) = E[Y(1, M)] - E[Y(0, M)]$. Modern treatment of direct and indirect effects owes its development to this notational provision and to the SEM semantics of interventions (Haavelmo, 1943; Spirtes et al., 1993) and counterfactuals (Balke and Pearl, 1995).

I believe that, with this narrative in mind, traditional SEM analysts should not have any difficulties accepting the premises of causal mediation. First, these analysts already accept structural equations as the basis for modeling (most statisticians do not). Second, counterfactuals in this narrative emerge naturally, as abbreviated structural equations (see Eq. (4) in the main paper). Third, traditional SEM analysts can easily appreciate the benefits of causal mediation analysis, since it endows them with two new capabilities: 1. Extending mediation analysis to nonlinear functions and highly interactive variables, continuous as well as discrete. 2. Distinguishing between the necessary and sufficient notions of mediation.

I hope this exchange helps clarify the logic and scope of causal mediation analysis as well as the unifying power of the SCM methodology. I thank Imai-et al. for commenting on my paper and contributing to this clarification.

References

Balke, A. and Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 11–18.

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

Bollen, K. and Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research* (S. Morgan, ed.). Springer, Dordrecht, Netherlands, 301–328.

Chen, B. and Pearl, J. (2013). Regression and causation: A critical examination of econometrics textbooks. *Real-World Economics Review* **65** 2–20.

Fisher, R. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Freedman, D. (1987). As others see us: A case study in path analysis (with discussion). *Journal of Educational Statistics* **12** 101–223.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press, New York.

Holland, P. (1995). Some reflections on Freedman's critiques. *Foundations of Science* **1** 50–57.

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

Imai, K., Keele, L., Tingley, D. and Yamamoto, T. (2010b). Causal mediation analysis using R. In *Advances in Social Science Research Using R* (H. Vinod, ed.). Springer (Lecture Notes in Statistics), New York, 129 − −154, <http://imai.princeton.edu/research/mediationR.html>.

Imai, K., Keele, L. and Yamamoto, T. (2010c). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.

Imai, K., Tingley, D. and Yamamoto, T. (2014). Commentary: Practical implications of theoretical results for causal mediation analysis. *Psychological Methods* This issue.

Jeffrey, R. (1965). *The Logic of Decisions*. McGraw-Hill, New York.

Judd, C. and Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* **5** 602–619.

Judd, C. and Kenny, D. (2010). Data analysis in social psychology: Recent and recurring issues. In *he handbook of social psychology* (D. Gilbert, S. T. Fiske and G. Lindzey, eds.), 5th ed. McGraw-Hill, Boston, MA, 115–139.

Lindley, D. (2002). Seeing and doing: The concept of causation. *International Statistical Review* **70** 191–214.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research* **27** 226–284.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2009c). Myth, confusion, and science in causal analysis. Tech. Rep. R-348, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.

Pearl, J. (2011a). Principal stratification – a goal or a tool? *The International Journal of Biostatistics* **7**. Article 20, DOI: 10.2202/1557-4679.1322. Available at: http://www.bepress.com/ijb/vol7/iss1/20.

Pearl, J. (2011b). The structural theory of causation. In *Causality in the sciences* (P. M. Illari, F. Russo and J. Williamson, eds.), chap. 33. Clarendon Press, Oxford, 697–727.

Pearl, J. (2014a). Interpretation and identification of causal mediation. *Psychological Methods* This volume.

Pearl, J. (2014b). Understanding Simpson's paradox. *The American Statistician* **68** 8–13.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31** 161–170.

Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.

Rubin, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine* **28** 1420–1423.

Shrier, I. (2009). Letter to the editor: Propensity scores. *Statistics in Medicine* **28** 1317–1318.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **13** 238–241.

Sjölander, A. (2009). Letter to the editor: Propensity scores and M-structures. *Statistics in Medicine* **28** 1416–1423.

Sobel, M. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* **33** 230–231.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag, New York.

Suppes, P. (1970). *A Probabilistic Theory of Causality.* North-Holland Publishing Co., Amsterdam.

Wermuth, N. (1992). On block-recursive regression equations. *Brazilian Journal of Probability and Statistics* (with discussion) **6** 1–56.