Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Consumer Psychology 22 (2012) 595 – 598

Methods Dialogue

# Mediation analysis and categorical variables: Some further frontiers☆

Fred M. Feinberg *

*Stephen M. Ross School of Business, University of Michigan, USA*

Received 25 January 2012; accepted 30 March 2012
Available online 12 April 2012

**Abstract**

Iacobucci (2012) provides a conceptually appealing, readily implemented measure to assess mediation for a far wider range of data type combinations than traditional OLS-based analyses permit. Here, we consider potential applications and extensions along several lines, particularly in terms of random utility models, simulation-based estimation, and potential nonlinearities, as well as some methodological and cultural impediments.


*Keywords:* Mediation analysis; Bayesian statistics; Categorical data analysis; Regression; Consumer behavior; Marketing

Mediation analysis is a deservedly celebrated method in social science generally (MacKinnon, Fairchild, & Fritz, 2007), and marketing particularly (Iacobucci, 2008). For example, do older consumers buy costlier cars owing to the accumulation of experience, of capital, or merely of birthdays? Despite warnings that statistical models are correlational, and the mantra that "correlation does not imply causation", what researchers hope to fashion is a plausible *causal* story. That is, not just an indicator, but a "because". Mediation helps supply this satisfying link.

Iacobucci (2012) zeroes in on a weakness in that linkage, one that owes to history and pedagogy as much as to methodology proper. Students in the social sciences and consumer behavior spend years steeped in the logic of classical statistics, which undergirds experimental design and associated regression-based techniques. Critical hypothesis tests typically need to be shoe-horned into (asymptotic) $z$, $t$, $\chi^2$, or $F$ distributions, though that picture is slowly changing. In a mediation analysis, this works fine so long as one sticks to OLS-based relationships among the $X$, $M$, and $Y$ variables; but plain vanilla OLS isn't appropriate for categorical data. This is the problem Iacobucci (2012) addresses, in an intuitive, easily applied manner, for binary mediators and outcomes. Although the paper does not dwell on derivations, it does provide something equally illuminating: extensive simulations, indicating when the method gives bankable answers.

The core of the matter at hand is conjuring up a well-behaved test statistic. If we consider the usual schematic (e.g., Iacobucci's Fig. 1), pre-standardize all variables to eliminate regression intercepts, and ignore errors for the time being, we get something like so:

$$Y = cX$$

$$M = aX$$

$$Y = c'X + bM.$$

Basic algebra suggests that $cX = c'X + bM = c'X + abX$, so that $c = c' + ab$. Simple!

We are interested in what happens to $c$ when the mediator, $M$, intervenes, so we need a before–after statement; that is, one about $c - c'$, which equals $a \cdot b$ in the OLS framework. And this is where the trouble begins: even when both $a$ and $b$ have "nice" (limiting) $t$ distributions, their product does not (having instead a "product normal distribution", more or less). Various papers have addressed this, adding to the collective confusion. If we pre-divide $a$ and $b$ above by their standard errors, there

☆ Many thanks to both C. W. Park and Dawn Iacobucci, for allowing this opportunity to reply, and for helpful suggestions from Rich Gonzalez, Brent McFerran, Scott Rick, and Carolyn Yoon.
* Fax: +1 734 936 8716.
*E-mail address:* feinf@umich.edu.

1057-7408/$ -see front matter © 2012 Society for Consumer Psychology. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jcps.2012.03.007

are *three* common forms of the main test for mediation; chronologically:

Aroian (1947):

$$t_a t_b \Big/ \sqrt{t_a^2 + t_b^2 + 1}.$$

Goodman (1960):

$$t_a t_b \Big/ \sqrt{t_a^2 + t_b^2 - 1}.$$

Sobel (1982):

$$t_a t_b \Big/ \sqrt{t_a^2 + t_b^2}\,.$$

Complicating matters is that they are each often referred to as "The Sobel Test" (see MacKinnon, Warsi, & Dwyer, 1995, 2002; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002, for additional detail). It is well-known that these are all *large-sample* tests, where "large" is left to the discretion of the researcher, but usually means at least moderate double-digits. Their main point of disagreement is whether a "1" belongs in the denominator, and what sign it should have. Clearly, the Aroian statistic is the most "conservative", the Goodman the most inclusive; Iacobucci's (2012) novel extension hews sensibly to the former.

Once again, all this at the service of obtaining a friendly asymptotic distribution, in this case, a $z$ or $t$. This made sense three decades ago, when the most recent of these forms was published. But that's fully twenty of those fabled 18-month "Moore's Law" doublings in computational power: roughly a million-fold, the same as nudging the moon to within a few blocks. Why should we still be stranded in that primordial computational universe?

The intervening and recent literatures have suggested two alternatives, bootstrapping and Bayesian approaches, respectively. Bootstrapping (Shrout and Bolger 2002; Preacher and Hayes 2008) is appealing when one has a classical statistical program "hammer" and is willing to whack many nails: it requires no new programming, solution concepts, or anything other than a bit of (automated) persistence in the form of "resampling with replacement". It is a sensible strategy when computer horsepower is cheap, but methods haven't fully caught up. The Bayesian approach shines when methods *do* catch up. Despite being sometimes fraught with Philosophy, the Bayesian approach is conceptually simple; instead of maximizing a likelihood, take a very large sample from it (perhaps with a "prior" tacked on). This conceptual simplicity nearly instantly hits a wall in implementation: conditional densities must be derived and programmed, long "chains" (sometimes many) need to be run, convergence must be monitored, "thinning" might be needed, and notoriously finicky log-marginal-densities must be calculated.

Researchers have enough trouble understanding their data without all these headaches. So, why would they want to assess mediation using Bayesian methods? For starters, it completely solves the "product of coefficients" problem, in fact rendering it trivial. If you want the distribution of the product of two parameters, just multiply their samples together: done. In fact, any function of multiple parameters is just a bit of arithmetic away, no Hessians or Jacobians or limiting densities or new *assumptions* in sight. Small or unbalanced samples present no problems. Inference is exact. In our present scenario, that means inference for the critical quantity in mediation, $a \cdot b$, is exact, since we have its actual distribution. Programs like WinBUGS (with an afternoon of acquaintance) and MlwiN (with none at all) allow Bayesian analysis for the most common "general linear models" (GLMs), with all the derivational nastiness hidden under the hood. And even SAS in its most recent incarnation (version 9) includes Bayesian estimation in three of its most venerated modules, GENMOD, LIFEREG, and PHREG.

Iacobucci (2012) is a welcome advance because it shows behavioral researchers how to dramatically extend the range of their analyses by sticking close to what they've done since Baron and Kenny's (1986) overwhelmingly influential article.[1] But colleagues in the empirical modeling camp are already starting to work mediation into their souped-up, custom-programmed modeling frameworks. There appear to be two keys to this: (1) realizing that GLM-based models posit a linear-additive specification for mean effects; and (2) recognizing that because product of coefficients can't be assumed $z$ or $t$ in real data, adopting a simulation-based estimation technology, primarily Bayesian, pays immediate dividends.

For example, Zhang, Wedel, and Pieters (2009) present what appears to be the first mediation analysis in marketing proper estimated using Bayesian techniques. They do the needed yeoman's work of updating the classic mediation picture (their Fig. 1) with latent instrumental variables for $X$ (IVs), $M$ (mediator), $Y$ (DV), and well as $Z$ (controls). Far more important from our perspective is that their estimation showcases both bootstrapping and MCMC (Markov chain Monte Carlo, the main technique in Bayesian analyses). Although their substantive results are beyond the scope of this Comment, their statistical findings (Table 2) are of clear interest: most of the bootstrapped estimates and their standard errors appear inflated, by around 10%. Additional studies will be required to see if this generalizes, but the key point is how simple it is to perform mediation analysis in a Bayesian setting: just multiply samples for the two coefficients in question. And, again, such inferences are exact; no additional assumptions or asymptotics in sight. Researchers in other areas have also approached Iacobucci's (2012) goal via simulation-based estimation. For example, Elliott, Raghunathan and Li (2010) perform a Bayesian analysis with dichotomous mediators and outcomes (recall that

[1] In marketing, Zhao, Lynch, and Chen (2010) provide an extended critique of assumptions underlying the Baron–Kenny analysis, as well as artifacts that can arise via Sobel type tests, which are known to have low power.

dichotomous $X$ isn't a problem, even in the classical analysis, since that's just a dummy variable).

But how is one to actually DO all this, without becoming a stats geek and programmer? Fortunately, very recent work addresses that, too. Imai, Keele, Tingley, and Yamamoto (2010) report on their `mediation` package in the open-source statistical language `R`, which allows for non-, semi-, and fully parametric inference in the linear (e.g., Baron–Kenny) and GLM frameworks. It uses both nonparametric bootstrap and quasi-Bayesian estimation for nearly any combination of {continuous, ordered, binary} mediators × {continuous, binary} outcomes, "right out of the box". We would also direct readers to Woody's (2011) exceptionally detailed and practitioner-oriented summary of mediation issues, estimation strategies (Sobel, bootstrap, MCMC, and especially Structural Equations Models), heterogeneity, and associated "best practices".

As Iacobucci (2012) conveys, one remaining frontier is the bestiary of categorical data types: multinomial, ordered, rank-ordered, pick-*k*-of-*n*, "divide 100 points", etc., and nefarious combinations thereof. Multinomial data, with the sort of "random utility" representation underlying logit-type models, may lure us into believing we can "just multiply" Monte Carlo samples for coefficients, analogously to Zhang et al. (2009) in the linear case. While that may not be outwardly wrong—dedicated simulation studies would need to assess this—it most certainly would presume that the *entire pattern* of choices among available options be used to explain mediation. But what, instead, if it involved trade-offs just between one subset of alternatives and its complement? The single coefficient in the (logit) utility is not meant to detect such trade-offs, let alone more subtle ones. The same sorts of issues arise for "exploded" models commonly used for ranked data, which are likewise based on a random utility formulation.

Another frontier involves the use of Structural Equation Models, which have been well-covered elsewhere as they apply to mediation (e.g., Cole and Maxwell 2003). SEMs allow such a wide vista of relationships that they should be informed by prior theory, not thoughtlessly and exhaustively run to converge upon a best-fitting model. SEMs have been traditionally estimated using classical methods, but Bayesian approaches are slowly gaining traction (Lee 2007); notably, one major SEM program, AMOS, has started to implement Bayesian estimation, a process sure to accelerate. Very recent work (Wang and Zhang 2011) applies Bayesian estimation in SEMs with censored data, e.g., web site visit logs covering a specific time period, thereby omitting (censoring) anything before or after. Practically speaking, difficulties arising from multiplying parameters may soon become a relatively minor issue even in SEM-based mediation analyses incorporating a variety of variable types as both mediators ($M$) and outcomes ($Y$).

One last frontier that can only be covered very briefly here concerns nonlinearity, specifically in terms of potential interaction effects. In a series of papers devoted to the topic, Pearl develops the "Mediation Formula," which is claimed to assess mediation for many data types in even highly nonlinear models. In his own description (Pearl 2012), "The Mediation Formula represents the average increase in the outcome $Y$ that the transition from $X=x$ to $X=x'$ is expected to produce absent any direct effect of $X$ on $Y$." Pearl (2001, 2012, and in many others) illustrates the formula, which is estimable by ordinary regression, for linear models, both with and without interaction, for logit and probit models, and for when any or all of $\{X, Y, M\}$ are binary, the situation also addressed in Iacobucci (2012). This presents a fertile avenue for future research, since essentially all prior approaches have known problems when the true model for $Y$ is a linear function not only of $X$ and $M$, but of an interaction term $XM$ as well.

In short, we live in interesting times with respect to mediation. Simulation-based methods—in particular, those relying on Bayesian estimation via data augmentation (Tanner and Wong 1987; Edwards and Allenby 2003), which "fills in" many data types to allow an underlying OLS-based representation—may soon allow researchers to assess mediation for essentially any sort of variable, including censored, complex categorical forms, interactions, and even to allow coefficient heterogeneity (e.g., random effects or hierarchical models). Until that day arrives, behavioral researchers need an intuitive, reliable, implementable method that works with the main types of data they encounter, and this is precisely what Iacobucci (2012) provides.

## References


Aroian, Leo A. (1947). The probability function of a product of two normally distributed variables. *Annals of Mathematical Statistics*, *18*, 265–271.

Baron, Reuben M., & Kenny, David 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.

Cole, David A., & Maxwell, Scott E. (2003). Testing mediational models with longitudinal data: Myths and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*, 558–577.

Edwards, Yancy, & Allenby, Greg (2003). Multivariate analysis of multiple response data. *Journal of Marketing Research*, *40*, 321–334.

Elliott, Michael R., Raghunathan, Trivellore E., & Li, Yun (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, *11*, 353–372.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*, 708–713.

Iacobucci, Dawn (2008). *Mediation analysis.* Thousand Oaks, CA: Sage.

Iacobucci, Dawn (2012). Mediation analysis and categorical variables: The final frontier. *Journal of Consumer Psychology* [Fill In Issue Information].

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010). Causal mediation analysis using R. In H. Vinod (Ed.), *Lecture notes in statistics: Advances in social science research using R* (pp. 129–154). New York: Springer.

Lee, Sik-Yum (2007). *Structural equation modeling: A Bayesian approach.* Chichester, UK: John Wiley and Sons.

MacKinnon, David P., Warsi, Ghulam, & Dwyer, James H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41–62.

MacKinnon, David P., Lockwood, Chondra M., Hoffman, Jeanne M., West, Stephen G., & Sheets, Virgil (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614.

Pearl, Judea (2001). Direct and indirect effects. *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann Publishers.

Pearl, Judea (2012). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, & L. Bernardinelli (Eds.), *Causality: Statistical perspectives and applications, chapter 12*. : J. Wiley.

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

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Sobel, Michael E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In Samuel Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). San Francisco: Jossey-Bass.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–550.

Wang, Lijuan, & Zhang, Zhiyong (2011). Estimating and testing mediation effects with censored data. *Structural Equation Modeling*, *18*(1), 18–34.

Woody, Erik (2011). An SEM perspective on evaluating mediation: What every clinical researcher needs to know. *Journal of Experimental Psychopathology*, *2*(2), 210–251.

Zhang, Jie, Wedel, Michel, & Pieters, Rik (2009). Sales effects of attention to feature advertisements: A Bayesian mediation analysis. *Journal of Marketing Research*, *46*, 669–681.

Zhao, Xinshu, Lynch, John G., Jr., & Chen, Qimei (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*, 197–206.