



Published in final edited form as:

J Consum Psychol. 2012 October 1; 22(4): 600–602. doi:10.1016/j.jcps.2012.03.009.

Commentary on “Mediation Analysis and Categorical Variables: The Final Frontier” by Dawn Iacobucci

David P. MacKinnon¹ and Matthew C. Cox²

¹Department of Psychology, Arizona State University

²Virginia Tech, Journal of Consumer Psychology

Dr. Iacobucci addresses the important topic of the most accurate mediation analysis when the mediator and outcome variable are categorical (Iacobucci 2012). There are many categorical mediating and outcome variables in marketing and other research that are most accurately modeled using logistic regression, Poisson regression, and survival analysis methods. A categorical independent variable is less critical because the independent variable is a predictor in all equations and coding procedures for X as binary or categorical are easily implemented. A general mediation analysis strategy for mediation models with any type of measurement scale for the mediator and the outcome variable would simplify analysis for researchers.

Mediation is fundamentally composed of two parts, one part representing the relation of the independent variable to the mediator and another part representing the relation of the mediator to the dependent variable; the *a* and *b* coefficients in Equations 2 and 3 in the article. In ordinary least squares regression or maximum likelihood estimation for continuous measures, both paths are in the same metric so the coefficient measures the change in the dependent variable for a one unit change in the independent variable. In this model both estimators of the mediated effect, the product of *a* and *b*, *ab*, and the difference of *c* and *c'*, *c - c'*, are algebraically equivalent (MacKinnon, Warsi, and Dwyer, 1995). The equivalency of *ab* and *c - c'* does not hold for categorical models such as logistic regression. It does not hold in logistic regression for example, because the error variance in logistic regression is fixed at $\pi^2/3$ for any logistic regression equation. Because the error term is fixed, the regression coefficients across equations, such as *c* and *c'*, cannot be directly compared because the *c'* coefficient, for example, measures the change in the relation of X to Y after adjustment for X plus an adjustment to keep the error variance fixed to $\pi^2/3$ (MacKinnon and Dwyer 1993).

One solution to this problem is to standardize regression coefficients prior to estimating mediation so that they are the value they would be if the error term was allowed to vary across equations as it does in continuous variable regression analysis (MacKinnon & Dwyer, 1993; Winship & Mare, 1983). When this is done, simulation studies have shown that the discrepancy between *ab* and *c - c'* is no longer substantial (MacKinnon and Dwyer 1993; MacKinnon, Lockwood, Brown, and Wang 2007). Methods based on the distribution of the product and resampling methods such as the bootstrap yield the most accurate confidence

© 2012 Society for Consumer Psychology. Published by Elsevier Inc. All rights reserved.

Do not cite or copy without first author's permission.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

intervals for the mediated effect and tests of hypothesis regarding the mediated effect for categorical data analysis (MacKinnon, Taylor, Yoon, Lockwood, and Thoemmes 2008). More on mediation in logistic and probit regression along with examples can be found in MacKinnon (2008, Chapter 11).

Another approach to the scaling problem that is now widely used in structural equation models is to model any categorical variable as an indicator of a latent continuous variable, such as Y^* . In the sample of the data, the sample measure of the categorical variable is modeled by dichotomization of values at a threshold on the Y^* latent variable. In this case, there is a model relating the observed categorical measure to the latent measure and the structural relations among these latent measures is estimated using a program such as Mplus (Muthen and Muthen 2010). Mplus is a general program that allows for accurate estimation of models with combinations of logistic, Poisson, continuous and other distributions. The use of Mplus is an excellent choice for the estimation of structural equation models for variables that differ in measurement scale. Methods also exist that do not assume a latent underlying continuous measure for observed categorical variables as described in MacKinnon et al. (2007). This application of structural equation modeling in Mplus for mixtures of continuous and categorical variables is the current method of choice.

The Distribution of the Product, $ZaZb$, test

Iacobucci (2012) proposes a solution to the problem of estimates from different analysis methods such as logistic and ordinary least squares regression. The idea is to obtain the values of $Za=a/s_a$ and $Zb=b/s_b$ for each estimate and divide this by the variance of the product of $ZaZb$. The formula in Iacobucci (2012) is for the second order solution for the standard error of ab . The formula in Equation 1 for the first order multivariate delta standard error (Sobel 1982) is given below in Equation 1 and does not include 1 in the denominator.

$$Z_{ZaZb} = \frac{ZaZb}{\sqrt{za^2 + zb^2}} \quad (1)$$

As mentioned in the article, the ratio in Equation 1 is algebraically equivalent to the ratio of the mediated effect divided by the multivariate delta standard error. As a result, the Z is tested for significance by comparing it with tabled values of the normal distribution or confidence limits may be calculated using the normal distribution. However, the Z_{ZaZb} does not always have a normal distribution and more accurate statistical tests and confidence intervals may be obtained by using the distribution of the product (MacKinnon et al. 2002; MacKinnon et al. 2007; Tofighi and MacKinnon 2011). The mediated effect is the product of regression coefficients and $ZaZb$ is the product of Z scores so the correct distribution is the distribution of the product. The distribution of the product provides more accurate confidence limits and statistical tests as demonstrated in several statistical simulation studies (MacKinnon, Lockwood, and Williams 2004). So the test described in the article would be improved by using the distribution of the product to conduct significance tests and create confidence intervals. These tests can be conducted using the PRODCLIN program and an improved program, RMediation, available here <http://www.amp.gatech.edu/RMediation>. For both programs, the user inputs the values a , standard error of a , b , standard error of b , correlation between a and b , and Type I error rate for the confidence intervals usually .05. The distribution of the product method has been evaluated in an extensive statistical simulation study for logistic regression and the method provided the most accurate tests and confidence limits for logistic regression analysis tests for mediation (MacKinnon, Taylor, Yoon, Lockwood, and Thoemmes 2008). Bootstrapping has also been proposed for significance testing and confidence limit estimation because it also more accurately models

the distribution of the product (MacKinnon, Lockwood, and Williams 2004; MacKinnon, Chapter 12 for more on computer intensive tests).

Another important issue for the test that uses separate values for Za and Zb is the possible presence of a correlation between the two coefficients. In the continuous case for ordinary least squares or maximum likelihood, the correlation between a and b is zero (Tofighi, MacKinnon, and Yoon 2009). In many analyses there is a correlation between the coefficients so this correlation should be included in the standard error as shown in Equation 2 for the mediated effect. For example, latent variable mediation models can have nonzero correlation between the a and b coefficients as described in MacKinnon (2008, Chapter 7). Note that the distribution of the product program described above allow for a correlation between the a and b coefficients. If there is a correlation between the a and b coefficients, and the test based on Z values is preferred, then Equation 3 should be used, where r is the correlation between a and b .

$$Z_{ab} = \frac{ab+r}{\sqrt{a^2s_b^2+b^2s_a^2+2a^2s_b^2b^2s_a^2r+s_a^2s_b^2+r^2}} \quad (2)$$

It is possible that a and b paths from different forms of analysis may be correlated. If Za and Zb are estimated in separate equations the correlation will not be available (note that the correlation between the estimates is available from structural equation modeling programs such Mplus with the TECH3 option). If the two coefficients come from different analyses then it may be difficult to obtain an estimate of the covariance necessary for the standard error but there may be analytical solutions for this problem. That is, it may be possible to derive this correlation for certain types of analyses and this correlation may be inserted in the equations above.

$$Z_{zab} = \frac{ZaZb+r}{\sqrt{za^2+zb^2+2za^2zb^2r+1+r^2}} \quad (3)$$

There is another issue related to categorical variables in mediation analysis that is related to the nonlinear nature of statistical models for categorical variables. The nonlinear nature of these models indicates that the value of the mediated effect depends on the values of the variables studied. That is, the mediated effect is likely to differ at different values of the mediating variable. As a result, a more general solution is needed that would integrate the effect across all values studied. Such formulas are beginning to appear and are based on a causal perspective (Pearl 2011; VanderWeele 2010)

Though the statistical issues described in Dr. Iacobucci's article and discussed above are certainly important, they assume that the true underlying mediation model is correct. There are many additional assumptions of the mediation models that are critical including no omitted variables affect the relation of X to M and M to Y and correct functional form for relations among variables (MacKinnon 2008). In many respects, the hardest parts of the mediation analysis do not depend on statistics but on the substantive background to justify the mediation theory. In marketing and social sciences this background has typically been based on strong theory which provides the rationale for the model tested. The specification of these models and the comprehensive and sustained investigation of their implication remains the most challenging aspect of the scientific study of mediating variables.

Acknowledgments

We thank Professor Dawn Iacobucci for the invitation to comment on her paper. This work was supported in part by National Institute on Drug Abuse Grant DA09757.

References

- Iacobucci, Dawn. Mediating and Categorical Variables: The Final Frontier. *Journal of Consumer Psychology*. 2012 to appear.
- MacKinnon, David P. Introduction to Statistical Mediation Analysis. Erlbaum; New York: 2008.
- MacKinnon, David P.; Dwyer, James H. Estimating Mediated Effects in Prevention Studies. *Evaluation Review*. 1993; 17(2):144–158.
- MacKinnon, David P.; Fritz, Matthew S.; Williams, Jason; Lockwood, Chondra M. Distribution of the Product Confidence Limits for the Indirect Effect: Program PRODCLIN. *Behavior Research Methods*. 2007; 39(3):384–389. [PubMed: 17958149]
- MacKinnon, David P.; Lockwood, Chondra M.; Williams, Jason. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research*. 2004; 39(1):99–128. [PubMed: 20157642]
- MacKinnon, David P.; Lockwood, Chondra M.; Hendricks Brown, C.; Wang, Wei; Hoffman, Jeanne M. The Intermediate Endpoint Effect in Logistic and Probit Regression. *Clinical Trials*. 2007; 4(5): 499–513. [PubMed: 17942466]
- MacKinnon, David P.; Lockwood, Chondra M.; Hoffman, Jeanne M.; West, Stephen G.; Sheets, Virgil. A Comparison of Methods to Test Mediation and Other Intervening Variable Effects. *Psychological Methods*. 2002; 7(1):83–104. [PubMed: 11928892]
- MacKinnon, David P.; Taylor, Aaron B.; Yoon, Myeonsung; Lockwood, Chondra M.; Thoemmes, Felix. A Comparison of Methods to Test Mediation and Other Intervening Variable Effects in Logistic Regression. 2008. manuscript submitted for publication
- MacKinnon, David P.; Warsi, Ghulam; Dwyer, James H. A Simulation Study of Mediated Effect Measures. *Multivariate Behavioral Research*. 1995; 30(1):41–62. [PubMed: 20157641]
- Muthén, Linda K.; Muthén, Bengt O. Mplus (Version 6.0) Computer Software. Los Angeles, CA: Muthén and Muthén; 2010.
- Pearl, Judea. The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms. *Prevention Science*. (in press).
- Sobel, Michael E. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*. 1982; 13:290–312.
- Tofighi, Davood; MacKinnon, David P. RMediation: An R Package for Mediation Analysis Confidence Intervals. *Behavior Research Methods*. 2011; 43(3):1–9. [PubMed: 21287104]
- Tofighi, Davood; MacKinnon, David P.; Yoon, Myeonsung. Covariances Between Regression Coefficient Estimates in a Single Mediator Model. *British Journal of Mathematical and Statistical Psychology*. 2009; 62(3):457–484. [PubMed: 18652738]
- VanderWeele, Tyler J. Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects. *Epidemiology*. 2010; 21(4):540–551. [PubMed: 20479643]
- Winship, Christopher; Mare, Robert D. Structural Equations and Path Analysis for Discrete Data. *American Journal of Sociology*. 1983:54–110.