Journal of Behavioral Data Science, 2023, *3* (2), 1–14.
DOI:https://doi.org/10.35566/jbds/v3n2/orourke

Considering the Distributional Form of Zeroes When Calculating Mediation Effects with Zero-Inflated Count Outcomes

Holly P. O'Rourke[1][0000−0002−2927−0333] and Da Eun Han[2][0000−0001−8699−439X]

[1] Arizona State University
holly.orourke@asu.edu
[2] University of Illinois at Urbana-Champaign
duoen10@gmail.com

Abstract. Recent work has demonstrated how to calculate conditional mediated effects for mediation models with zero-inflated count outcomes in a non-causal framework (O'Rourke & Vazquez, 2019); however, those formulas do not distinguish between logistic and count portions of the data distribution when calculating mediated effects separately for zeroes and counts. When calculating conditional mediated effects for the counts in a zero-inflated count outcome Y, the *b* path should use the partial derivative of the log-linear regression equation for X and M predicting Y. When calculating conditional mediated effects for the zeroes, the *b* path should use the partial derivative of the logistic regression equation for X and M predicting Y instead of the log-linear equation. This paper presents adjustments to the analytical formulas of conditional mediated effects for mediation with zero-inflated count outcomes when zeroes and counts are differentially predicted. Using a Monte Carlo simulation, we also empirically show that these adjustments produce different results than when the distributional form of zeroes is ignored.

Keywords: Mediation analysis · Count outcomes · Zero-inflation · ZIP · ZINB · Hurdle models

1 Introduction

Many theories in the social and behavioral sciences specify indirect mechanisms by which predictors influence outcomes. These mechanisms, also known as mediators, are incorporated into such theories through the use of mediation models. Mediation models are widely applied to theories of human behavior and test the indirect influence of a predictor variable (X) on an outcome (Y) via a mediator (M). Much methodological research on the mediation model has focused on models where the endogenous variables M and Y are continuously distributed and assume linear associations, and several extensions have been proposed as well for models where M and Y are categorical (i.e., binary or count) variables that are modeled with logistic or other exponential family distributions (Coxe & MacKinnon, 2010; Gilula, 2012; Iacobucci, 2012; Imai, Keele, & Tingley, 2010; Mackinnon, 2008; MacKinnon & Cox, 2012; Mackinnon & Dwyer, 1993; Preacher, 2015; Valeri & VanderWeele, 2013; VanderWeele, Zhang, & Lim, 2016). However, these methods are not appropriate for use where categorical endogenous variables contain zero-inflation.

Zero-inflation occurs when the proportion of observations with a value of zero on a particular variable is larger than what is expected from the variable's typical zero-uninflated distribution (for example, Poisson or negative binomial if a variable is a measure of counts). Zero-inflated (ZI) count variables are common in the social sciences. For example, consider a study of externalizing behaviors in middle school; for a given count variable measuring bullying as "number of times child was a bully in the past month", many students would have a score of zero because most children do not engage in bullying behaviors. Another example from health intervention research would be measuring drinking outcomes in a study designed to help adults with alcohol use disorder quit drinking. For a drinking count variable measured as "number of drinks consumed in the past week", many participants would have a score of zero because they are actively trying to refrain from drinking.

The traditional methods cited above for categorical mediation analysis are not equipped to handle excess zeroes in the outcome, and using these models to fit data with zero-inflated distributions may result in biased estimates. A technical body of literature does exist for causal inference methods to assess mediation with categorical variables that contain zero-inflation (Cheng et al., 2018; Wang & Albert, 2012) but this literature is not as accessible to applied researchers due to the complexity of its application. The causal literature differs from the general linear model (GLM)-based mediation literature in that it requires a working knowledge of causal inference frameworks that involve formulas for probability, and causal methods often require additional sensitivity analyses for a formal test of mediation. Furthermore, the effects involved in mediation from the causal inference literature are defined differently from GLM mediation effects, requiring the calculation and interpretation of multiple effects to determine mediation even in simple cases.

Recent work on mediation for ZI counts has applied Geldhof, Anthony, Selig, and Mendez-Luck (2018)'s method of calculating mediation effects for count data that are conditional upon values of X to mediation models with zero-inflated count outcomes using a modeling framework that does not come from causal inference (O'Rourke & Vazquez, 2019). In this method, mediation effects are calculated separately for zeroes and counts when Y is zero-inflated. However, that method does not account for the unique distributional nature of the zeroes, as zeroes are predicted using the binomial logistic model while counts are fitted using the log-linear model. This article illustrates a revised formula for calculating the mediation effect for the zeroes when zeroes and counts are differentially predicted. We also demonstrate via Monte Carlo simulation that the revised formula produces different results than the original formula that does not distinguish between zeroes and counts in a zero-inflated outcome.

1.1 Mediation

The simplest single-mediator model is described by two OLS regression equations using notation from Mackinnon (2008).

$$Y = i_1 + bM + c'X + e_1 \tag{1}$$

$$M = i_2 + aX + e_2 \tag{2}$$

In these equations, X is the predictor, M is the mediator, and Y is the outcome. From Equation 1, the influence of M on Y is known as the b parameter, and the influence of X on Y controlling for M is known as the c' parameter (also known as the "direct effect"). From Equation 2, the influence of X on M is known as the a parameter. The parameters i_1 and i_2 are model intercepts and e_1 and e_2 are model errors. The mediated effect that is the focus of this article is specified as the product of the a and b parameters (ab), commonly referred to in the mediation literature (and hereafter referred to) as the "mediated effect". Other specifications of the mediated effect and their equalities with respect to count outcomes are described elsewhere (Coxe & MacKinnon, 2010; MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007; Mackinnon, 2008; O'Rourke & Vazquez, 2019).

Two common approaches to significance testing in mediation are the causal steps (Baron & Kenny, 1986; Judd & Kenny, 1981; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002) and product of coefficients (Sobel, 1982) approaches. The recommended test from the causal steps approach is the Joint Significance test, which has the best balance of power and Type I error (MacKinnon et al., 2002). The Joint Significance test uses individual $z-$ or $t-$tests of estimates of the respective a and b parameters to assess significance: if both tests are statistically significant, we can conclude that mediation is present. The product of coefficients approach was developed using a derived asymptotic standard error (Sobel, 1982) to compute a $z-$test for the mediated effect ab. More recently, it has become common to assess significance of the mediated effect by using bootstrapping to create asymmetric confidence intervals for ab (MacKinnon, Lockwood, & Williams, 2004). Bootstrapping is used because ab is a product of two variables and so it is not normally distributed (Aroian, 1947; Craig, 1936), meaning traditional formulas that assume a normal distribution of z produce biased confidence intervals for ab.

Mediation analysis is conducted with count outcomes in a similar manner to the approach described above. The difference is that instead of a normal distribution of continuous Y (the assumption under linear regression), the count outcome Y is assumed to have a Poisson distribution, negative binomial (NB) distribution if overdispersed, or beta-binomial distribution if overdispersed with a restricted upper bound. If Y is a count outcome, Poisson, NB, or beta-binomial regression can be used to fit the model specified in Equation 1, and (assuming continuous M) Equation 2 can be assessed as usual with linear regression.

1.2 Zero-Inflated Counts

Each of the models described for count outcomes has a ZI counterpart: the ZI Poisson (ZIP), ZI negative binominal (ZINB), or ZI beta-binomial (ZIBB) models. These are known as zero-inflated generalized linear models (ZI-GZLMs). Hurdle models, similar to ZI-GZLMs, can also be used to model ZI count outcomes. However, zeroes are treated differently in hurdle models compared with ZI models. ZI-GZLMs assume that there are two kinds of zeroes: "structural" (i.e., excess) zeroes that will never take on another value, and "sampling" zeroes that have some potential to be non-zero. (For the remainder of this paper, when we refer to the zeroes in Y, we are referring to the structural zeroes that are modeled separately from the counts and sampling zeroes.) Hurdle models do not make the structural vs. sampling zero distinction, but instead assume that all zeroes are generated from the same process. In other words, hurdle models treat all zeroes as structural excess zeroes that are the only source of overdispersion in the data, and these models do not include an additional probability mass which distinguishes structural zeroes from counts and sampling zeroes in other ZI-GZLMs.

In ZI-GZLMs, the probability of an occurrence of an excess zero is modeled as follows.

$$z \sim \text{Bernoulli}(\pi) \tag{3}$$

Where π is the probability of observing excess zeroes. Assuming a ZI Poisson distribution (the simplest in terms of parameterization because mean and variance are assumed to be equal), and where λ is the mean count from the Poisson distribution, the probability mass function of a ZIP model is as follows.

$$P(Y = 0) = \pi + (1 - \pi)e^{-\lambda} \tag{4}$$

$$P(Y \neq 0) = (1 - \pi)(\frac{\lambda^Y}{Y!})e^{-\lambda} \tag{5}$$

1.3 Mediation for Zero-Inflated Counts

One recent non-causal method of assessing mediation for count outcomes suggested computing multiple conditional mediated effects for chosen values of the predictor X (Geldhof et al., 2018), representing the nonlinear relation between X and Y as several conditionally linear relations that differ across values of X (Stolzenberg, 1980). This method was extended to mediation models for ZI count outcomes (O'Rourke & Vazquez, 2019) by calculating two separate sets of conditional mediated effects: one for the zeroes and one for the counts. For a model with any measurement level of X, continuous M, and ZI count Y, b paths were calculated separately for structural zeroes and counts using the first partial derivative with respect to M of the loglinear mediation regression equation shown in Equation 1. This loglinear mediation regression equation is given below.

$$\hat{Y} = e^{i_1 + bM + c'X} \tag{6}$$

The first partial derivative of Equation 6 being the following.

$$b_{LL} = \frac{\partial \hat{Y}}{\partial M} = b(e^{i_1 + bM + c'X}) \tag{7}$$

The formula in Equation 7 for b_{LL} (b path from the *loglinear* equation) was used in conjunction with the a path from Equation 2 to calculate conditional mediated effects as follows.

$$a * b_{LL} \tag{8}$$

Two sets of k conditional mediated effects were calculated separately for zeroes and counts using the formula in Equation 8, with k being equal to the number of chosen values of X (this number would typically be $k = 2$ for binary X, $k = 3$ for continuous X at low, medium. and high values) and M fixed at its mean. Equation 8 was used to calculate sets of conditional mediated effects for both ZIP and ZINB models, as both Poisson and negative binomial models utilize log link functions.

1.4 Distributional Form for Zeroes

The method described above produced the desired sets of conditional mediated effects, however, using the partial derivative formula b_{LL} for both the structural zeroes and the counts disregarded the form of the assumed distribution for the structural zeroes. Specifically, the structural zeroes are modeled with a logistic distribution. Therefore, the logistic regression for predicting structural zeroes has a logit link function of

$$ln(\frac{\pi}{1-\pi}) \tag{9}$$

The mean function corresponding to this logit link function is as follows.

$$\hat{Y} = \frac{e^{i_1 + bM + c'X}}{e^{i_1 + bM + c'X} + 1} \tag{10}$$

Taking the first partial derivative of Equation 10 with respect to M gives us b_{LG} (the b path from the *logistic* mediation regression equation).

$$b_{LG} = \frac{\partial \hat{Y}}{\partial M} = b\frac{e^{i_1 + bM + c'X}}{(e^{i_1 + bM + c'X} + 1)^2} \tag{11}$$

This would result in a conditional mediated effect for the zeroes of

$$a * b_{LG} \tag{12}$$

Both of the first partial derivative formulas for b presented here are known quantities for estimating a mediation path with either a count or binary non-ZI endogenous variable (Geldhof et al., 2018; Li, Schneider, & Bennett, 2007), but they have not been used in tandem to handle two-part mediation models for ZI endogenous variables.

The current paper aims to utilize both of these formulas for the b path, and to demonstrate that using the b path from the logit link function b_{LG} when calculating mediated effects for the zeroes results in different estimates of the conditional mediated effects than using the b path from the log link function b_{LL} for both zeroes and counts. In the next section, we describe a Monte Carlo simulation study that demonstrates that the formulas produce different estimates of the conditional mediated effects for the zeroes (hereafter referred to as "conditional mediated effects").

2 Simulation Study

2.1 Simulation Conditions

We conducted a Monte Carlo simulation in R 4.2.3 in conjunction with Mplus version 8.10 (Muthén & Muthén, 2017). In this simulation study, we manipulated two factors: Sample size ($N = 100$, 250, 500, 750, and 1500) and population distribution of the counts in the outcome (Poisson vs. negative binomial). Simulation manipulations resulted in 2 x 5 = 10 conditions. Sample sizes were chosen to represent a range from small to large samples based on sample sizes commonly observed in the behavioral sciences. Manipulation of sample size allowed for us to examine possible effects of sample size on results by examining whether the difference in estimates of the conditional mediated effect grew smaller as sample size increased. The Poisson and negative binomial distributions for counts were chosen as the two distributions that are most commonly observed and practically applied with ZI-GzLMs. Differences in estimates of the conditional mediated effects were expected to be stable across the two levels of distribution of the count outcomes.

Population parameter values were not varied over conditions, and the parameter values used in each simulation model are given in Table 1.

Table 1. Simulation Study Parameter Values

Parameter	Population Value
a	0.59
b (Zeroes)	-0.14
b (Counts)	0.14
c' (Zeroes)	-0.01
c' (Counts)	0.01
μ_M	0
σ_M	1
μ_Y (Zeroes)	0
μ_Y (Counts)	3
ϕ*	1

*for ZINB models only

Parameter value magnitudes were chosen to reflect large (0.59) effect size for a and small (0.14) effect sizes for b as established in prior simulation research on single mediator models (Fritz & MacKinnon, 2007; MacKinnon et al., 2002; O’Rourke & MacKinnon, 2015), and parameter value signs were chosen such that conditional mediated effects would differ in sign for the zeroes and counts, as discussed below. The *c’* path was assigned a very small effect size in accordance with a model that would approach full mediation, a condition where *c’* = 0 (Mackinnon, 2008), but still would factor into calculations of conditional mediated effects. Table 2 shows population calculations of the conditional mediated effects across values of X based on the parameter values given in Table 1, using both the log link and logit link b paths.

Table 2. Calculation of Population Conditional Mediated Effects for Log Link and Logit Link Formulas

	Log Link Function b path	
	X = 0	X = 1
General Formula	$a*b(e^{i_1+bM+c'X})$	$a*b(e^{i_1+bM+c'X})$
$e^{i_1+bM+c'X}$	$e^{0+(-0.14)(0)+(-0.01)(0)}=1$	$e^{0+(-0.14)(0)+(-0.01)(1)}=1.01$
$b(e^{i_1+bM+c'X})$	$-0.14*1=-0.14$	$-0.14*1.01=-0.141$
$a*b(e^{i_1+bM+c'X})$	$0.59*-0.14=-0.0826$	$0.59*-0.141=-0.0834$
Conditional Mediated Effect	**-0.0826**	**-0.0834**
	Logit Link Function b path	
	X = 0	X = 1
General Formula	$a*b\frac{e^{i_1+bM+c'X}}{(e^{i_1+bM+c'X}+1)^2}$	$a*b\frac{e^{i_1+bM+c'X}}{(e^{i_1+bM+c'X}+1)^2}$
$e^{i_1+bM+c'X}$	$e^{0+(-0.14)(0)+(-0.01)(0)}=1$	$e^{0+(-0.14)(0)+(-0.01)(1)}=1.01$
$\frac{e^{i_1+bM+c'X}}{(e^{i_1+bM+c'X}+1)^2}$	$\frac{1}{(1+1)^2}=0.25$	$\frac{1.01}{(1.01+1.01)^2}=0.2499$
$b\frac{e^{i_1+bM+c'X}}{(e^{i_1+bM+c'X}+1)^2}$	$-0.14*0.25=-0.035$	$-0.14*0.2499=-0.0349$
$a*b\frac{e^{i_1+bM+c'X}}{(e^{i_1+bM+c'X}+1)^2}$	$0.59*-0.035=-0.0207$	$0.59*-0.0349=-0.0206$
Conditional Mediated Effect	**-0.0207**	**-0.0206**

2.2 Data Generation and Data Analysis

The R MplusAutomation package (Hallquist & Wiley, 2018) was used to simulate data. For each of the conditions, 500 replications with complete data were simulated. The paths related to mediation (b and *c’*) were specified to be equal in magnitude but opposite in sign for the zeroes and counts in Y in accordance with commonly observed patterns of results in applied ZI-GZLMs. Binary X was simulated with a Bernoulli distribution with $X \in 0, 1$, and M was simulated with a continuous Gaussian distribution $M \sim N(0, 1)$. The counts in Y were simulated to have a mean of 3 and the zeroes in Y, a mean of 0. Replications for Y with a ZIP distribution did not include a dispersion parameter, and when Y was simulated to have a ZINB distribution, the dispersion parameter was $\phi = 1$ as specified by the variance of Y in the Mplus MODEL command.

After all datasets were generated, the R MplusAutomation package was then used to create and run Mplus scripts analyzing all replications within each condition and then import results into R. This process was repeated for each of the 10 conditions. For each replication, a ZI-GZLM was fitted to the data using Maximum Likelihood estimation. Conditional mediated effects were calculated at X = 0 and X = 1 using the Mplus "Model Constraint" command. For the zeroes in Y, sets of conditional mediated effects were calculated using both the original method with b_{LL} for the b path and the revised method with b_{LG} for the b path. This resulted in four conditional mediated effects for comparison in further analyses. Bootstrapped confidence intervals were also generated to assess significance of each conditional mediated effect. Sample Mplus and R scripts can be found on the GitHub project at https://github.com/horourke/MZI2.

2.3 Simulation Study Outcomes and Outcome Analyses

We assessed differences in results for the conditional mediated effect estimates by examining relative parameter difference (i.e., relative bias for the ab_{LG} estimate). The relative difference was calculated as the difference between the population value of ab_{LG} and the respective estimates $\widehat{ab_{LL}}$ and $\widehat{ab_{LG}}$, over the population value of ab_{LG}.

$$\frac{ab_{LG} - \widehat{ab_{LL}}}{ab_{LG}} \tag{13}$$

$$\frac{ab_{LG} - \widehat{ab_{LG}}}{ab_{LG}} \tag{14}$$

Efficiency was calculated using the standard deviations of the raw estimates of the conditional mediated effects averaged across each condition. We also examined statistical power for each condition, calculated as the proportion of replications for which the p value associated with each conditional mediated effect was less than .05 and the bootstrapped confidence intervals of each conditional mediated effect did not include zero.

In preparation for analysis of the relative difference outcome, data were restructured to long format such that use of the b formula (b_{LL} vs. b_{LG}) could be coded as an additional binary predictor of a given outcome. Analyses conducted in R examined the impact of the condition on the dependent variable of interest at the replication level, with one replication considered as one observation. Analysis of variance (ANOVA) was used to investigate the differences in study conditions for relative parameter differences. Analyses were conducted separately for each outcome at X = 0 and X = 1. Factors representing study conditions in each ANOVA were sample size, population distribution of the counts in the outcome, and method of calculating the b path. In addition to main effects, all possible two- and three-way interactions were included as predictors in each ANOVA. Only ANOVA estimates that were significant at $p < .05$ with corresponding partial η^2 values of .02 (small amount of variance explained) or higher were considered meaningfully significant for the interpretation of results.

3 Results

3.1 Relative Difference

The average relative difference over replications for each condition is shown in Table 3. Results from the ANOVAs for both X = 0 and X = 1 indicated that only the method of calculating the b path (b_{LG} vs. b_{LL}) was a meaningfully significant predictor of relative difference. Method of calculating the b path explained 20.6% and 15.1% of the variability in the outcome respectively, which were large effect sizes (X = 0: $p < .001$, partial η^2 = .206; X = 1: $p < .001$, partial η^2 = .151).

Table 3. Relative Difference of Conditional Mediated Effects Collapsed Across Conditions

ZINB				
	X = 0		X = 1	
n	ab_{LL}	ab_{LG}	ab_{LL}	ab_{LG}
100	2.917	-0.052	4.620	-0.101
250	3.226	0.017	3.857	-0.005
500	3.089	0.011	3.330	0.000
750	3.062	0.007	3.176	0.000
1500	3.065	0.009	3.074	0.006
ZIP				
	X = 0		X = 1	
n	ab_{LL}	ab_{LG}	ab_{LL}	ab_{LG}
100	3.105	-0.001	4.599	-0.049
250	3.243	0.029	3.765	0.011
500	3.089	0.012	3.282	0.003
750	3.076	0.012	3.169	0.007
1500	3.066	0.012	3.073	0.010

Examining Table 3 for X = 0, the average relative difference was around 3 or above for all conditions for ab_{LL} estimates. When using the ab_{LG} formula to calculate conditional mediated effects, the average relative difference only reached an absolute value above .05 for the ZINB model at the smallest sample size, and the average relative difference was otherwise extremely small for all conditions using the ab_{LG} formula.

For X = 1, results from Table 3 indicate that the average relative difference of the estimates of ab_{LL} ranged from $[3.073, 4.620]$ for all conditions, with average relative difference decreasing as sample size increased[1]. As with calculations for

[1] Sample size was a statistically significant predictor of relative difference for the ANOVA where X = 1, however the partial $\eta^2 < .02$.

X = 0, the average relative difference of the conditional mediated effects using the ab_{LG} formula only reached an absolute value above .05 for the condition fitting the ZINB model at the smallest sample size, and the average relative difference was otherwise not problematic (i.e., below an absolute value of .05) for all conditions using the ab_{LG} formula.

3.2 Efficiency

For all values of X and regardless of sample size and distribution of outcomes, estimates of ab_{LG} had smaller variability (i.e., were more efficient) than for estimates of ab_{LL}, as shown by the averaged standard deviations of the estimates in Table 4. The difference in efficiency between the two sets of conditional mediated effect estimates decreased monotonically as sample size increased such that the conditional mediated effect estimates calculated with ab_{LL} were least efficient at the smallest sample sizes.

Table 4. Efficiency of Conditional Mediated Effects Collapsed Across Conditions

	ZINB			
	X = 0		X = 1	
n	ab_{LL}	ab_{LG}	ab_{LL}	ab_{LG}
100	0.152	0.036	0.242	0.033
250	0.090	0.021	0.119	0.021
500	0.061	0.015	0.073	0.015
750	0.047	0.012	0.053	0.011
1500	0.034	0.008	0.037	0.008
	ZIP			
	X = 0		X = 1	
n	ab_{LL}	ab_{LG}	ab_{LL}	ab_{LG}
100	0.147	0.034	0.221	0.032
250	0.086	0.021	0.109	0.020
500	0.059	0.014	0.069	0.014
750	0.045	0.011	0.050	0.011
1500	0.033	0.008	0.035	0.008

3.3 Power

Power values by condition can be found in Table 5. For conditional mediated effects where X = 0, there were negligible differences in power between the methods of calculating the b path. For conditional mediated effects where X = 1, power was slightly larger for estimates of ab_{LG} than for estimates of ab_{LL}. Power increased as the sample size increased for all conditions, and there were negligible differences in power between the ZIP and ZINB conditions. Power never reached a level of .8 (Cohen, 1988) in any of the conditions, likely due to the small magnitude of the b paths.

Table 5. Power of Conditional Mediated Effects Collapsed Across Conditions

ZINB				
	X = 0		X = 1	
n	ab_{LL}	ab_{LG}	ab_{LL}	ab_{LG}
100	0.000	0.008	0.000	0.018
250	0.068	0.106	0.000	0.118
500	0.214	0.250	0.050	0.258
750	0.400	0.418	0.222	0.420
1500	0.692	0.700	0.612	0.702
ZIP				
	X = 0		X = 1	
n	ab_{LL}	ab_{LG}	ab_{LL}	ab_{LG}
100	0.000	0.020	0.000	0.032
250	0.090	0.120	0.000	0.134
500	0.274	0.298	0.084	0.308
750	0.432	0.442	0.286	0.446
1500	0.728	0.730	0.670	0.732

4 Discussion

We used a Monte Carlo simulation to demonstrate the differences in results using log-linear vs. logistic regression equations for zeroes when calculating conditional mediated effects in a mediation model where Y is a ZI count. The conditional mediated effects for the zeroes in Y were calculated using two different b path formulas, b_{LL} and b_{LG}, where b_{LG} used the distributional form of the zeroes. Comparing estimates of ab_{LL} and ab_{LG}, we found that the conditional mediated effects differed significantly in magnitude at both values of X, for both ZINB and ZIP models and across all sample sizes examined. Specifically, results for relative difference showed that estimates of ab_{LL} were significantly different from estimates of ab_{LG}, and that this difference held across sample sizes and outcome distributions. Conditional mediated effects for zeroes calculated using the b_{LG} formula were also more efficient, and when X was non-zero, had slightly higher power. These results indicate that the choice of b path formula has a meaningful impact on the interpretation of results for the conditional mediated effects and should be considered when using this method to conduct mediation analysis with ZI count variables.

For conditional mediated effects calculated at non-zero X, power was slightly higher for ab_{LG} than for ab_{LL}. This means that when using the different formulas for b, we may make different conclusions about the significance of the conditional mediated effects when X is non-zero (for example, we could observe that the conditional mediated effect $ab_{LL[X=1]}$ was not significant and $ab_{LG[X=1]}$ was significant), which further highlights the importance of considering the distributional form of the zeroes when conducting mediation analysis where ZI counts are present in the data.

In this paper, we focused specifically on the mediation model that contained Y as a ZI count, meaning we discussed the issue of separate distributions of zeroes and counts with respect to only the *b* path (from M to Y) in mediation. However, this issue is applicable as well to models with a ZI count mediator, in which case we would calculate an *a* path using a log-linear regression equation for counts in M and an *a* path using a logistic regression equation for zeroes in M. Furthermore, it would be possible to use this method in a model where both M and Y are ZI counts. Under such circumstances, the *a* path for the counts in M could be calculated using the first partial derivative with respect to X of the log-linear transformation of Equation 2, and the *b* path for the counts in Y could be calculated using Equation 7. The *a* and *b* paths for the zeroes in M and Y would then be calculated using the first partial derivatives of the logit transformations of their respective regression equations (Equation 11 for the *b* path).

The utilization of these formulas can also be applied to future methodological work on mediation analysis with ZI count variables. The method described in this paper that is an extension of O'Rourke and Vazquez (2019) can be extended to more complex models that are frequently used by applied researchers, such as models that incorporate time (i.e., longitudinal models). This process of mediation can be expanded to ZI-GZLMs for repeated measures nested within individuals that are fitted in the multilevel modeling framework.

It is important for researchers to have accessible methods of assessing mediation in complex nonlinear models. This paper advances accessible methodology in the pursuit of best practices for investigating mediators when data are nonnormal. The simulation results presented here highlight the complexity of calculating mediated effects in models where ZI counts are present.

References

Aroian, L. A. (1947, June). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, *18*(2), 265–271. doi: https://doi.org/10.1214/aoms/1177730442

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. doi: https://doi.org/10.1037/0022-3514.51.6.1173

Cheng, J., Cheng, N. F., Guo, Z., Gregorich, S., Ismail, A. I., & Gansky, S. A. (2018, September). Mediation analysis for count and zero-inflated count data. *Statistical Methods in Medical Research*, *27*(9), 2756–2774. doi: https://doi.org/10.1177/0962280216686131

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed ed.). Hillsdale, N.J: L. Erlbaum Associates.

Coxe, S., & MacKinnon, D. P. (2010, November). Abstract: Mediation analysis of Poisson distributed count outcomes. *Multivariate Behavioral Research*, *45*(6), 1022–1022. doi: https://doi.org/10.1080/00273171.2010.534375

Craig, C. C. (1936, March). On the frequency function of \$xy\$. *The Annals of Mathematical Statistics*, *7*(1), 1–15. doi: https://doi.org/10.1214/aoms/1177732541

Fritz, M. S., & MacKinnon, D. P. (2007, March). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233–239. doi: https://doi.org/10.1111/j.1467-9280.2007.01882.x

Geldhof, G. J., Anthony, K. P., Selig, J. P., & Mendez-Luck, C. A. (2018, March). Accommodating binary and count variables in mediation: A case for conditional indirect effects. *International Journal of Behavioral Development*, *42*(2), 300–308. doi: https://doi.org/10.1177/0165025417727876

Gilula, Z. (2012, October). Mediation with categorical variables: Consider ordinal models, empirical Bayes, and alternatives to R2. *Journal of Consumer Psychology*, *22*(4), 599. doi: https://doi.org/10.1016/j.jcps.2012.03.008

Hallquist, M. N., & Wiley, J. F. (2018, July). *MplusAutomation* : An R package for facilitating large-scale latent variable analyses in M *plus*. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. doi: https://doi.org/10.1080/10705511.2017.1402334

Iacobucci, D. (2012, October). Mediation analysis and categorical variables: The final frontier. *Journal of Consumer Psychology*, *22*(4), 582–594. doi: https://doi.org/10.1016/j.jcps.2012.03.006

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. doi: https://doi.org/10.1037/a0020761

Judd, C. M., & Kenny, D. A. (1981, October). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, *5*(5), 602–619. doi: https://doi.org/10.1177/0193841X8100500502

Li, Y., Schneider, J. A., & Bennett, D. A. (2007). Estimation of the mediation effect with a binary mediator. *Statistics in Medicine*, *26*(18), 3398–3414. doi: https://doi.org/10.1002/sim.2730

MacKinnon, D., Lockwood, C., Brown, C., Wang, W., & Hoffman, J. (2007, October). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, *4*(5), 499–513. doi: https://doi.org/10.1177/1740774507083434

Mackinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

MacKinnon, D. P., & Cox, M. G. (2012, October). Commentary on “Mediation analysis and categorical variables: The final frontier” by Dawn Iacobucci. *Journal of Consumer Psychology*, *22*(4), 600–602. doi: https://doi.org/10.1016/j.jcps.2012.03.009

Mackinnon, D. P., & Dwyer, J. H. (1993, April). Estimating Mediated Effects in Prevention Studies. *Evaluation Review*, *17*(2), 144–158. doi: https://doi.org/10.1177/0193841X9301700202

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104. doi:

https://doi.org/10.1037/1082-989X.7.1.83

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004, January). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus.* Los Angeles, CA: Muthén & Muthén.

O'Rourke, H. P., & Vazquez, E. (2019, July). Mediation analysis with zero-inflated substance use outcomes: Challenges and recommendations. *Addictive Behaviors*, *94*, 16–25. doi: https://doi.org/10.1016/j.addbeh.2019.01.034

O'Rourke, H. P., & MacKinnon, D. P. (2015, June). When the test of mediation is more powerful than the test of the total effect. *Behavior Research Methods*, *47*(2), 424–442. doi: https://doi.org/10.3758/s13428-014-0481-z

Preacher, K. J. (2015, January). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, *66*(1), 825–852. doi: https://doi.org/10.1146/annurev-psych-010814-015258

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290. doi: https://doi.org/10.2307/270723

Stolzenberg, R. M. (1980). The measurement and decomposition of causal effects in nonlinear and nonadditive models. *Sociological Methodology*, *11*, 459. doi: https://doi.org/10.2307/270872

Valeri, L., & VanderWeele, T. J. (2013, June). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, *18*(2), 137–150. doi: https://doi.org/10.1037/a0031034

VanderWeele, T. J., Zhang, Y., & Lim, P. (2016, September). Brief report: Mediation analysis with an ordinal outcome. *Epidemiology*, *27*(5), 651–655. doi: https://doi.org/10.1097/EDE.0000000000000510

Wang, W., & Albert, J. M. (2012, November). Estimation of mediation effects for zero-inflated regression models. *Statistics in Medicine*, *31*(26), 3118–3132. doi: https://doi.org/10.1002/sim.5380