Interpretation of main effects in the presence of non-significant interaction effects

Julie A. Lorah [a]

[a]Counseling and Educational Psychology, Indiana University

Abstract ■ Moderated regression models include an interaction, or product term, and can be used to assess whether the relationship between a given independent variable (IV) and a dependent variable (DV) depends on a third moderator variable (MV). If the moderation effect is significant, researchers recommend either ignoring main effects completely, or carefully interpreting them as conditional effects. However, when the moderation effect is not significant, this implies that the typical interpretation of main effects as average effects is appropriate. The present study challenges this claim since lack of significance may be due to lack of power rather than to no true population effect. To explore this idea, a simulation study is conducted and analytic illustration provided. Results indicate that when a true moderation effect exists, it may not be detected, implying the potential for misleading interpretation of main effects. To guard against this, applied researchers are encouraged to conduct power analyses prior to a moderation study; to mean-center predictors; to consider exploring the main-effects-only model by omitting the interaction effect; and to consider information criteria approaches to testing effects.

Keywords ■ interactions; interpretation; main effects; moderation.

Jlorah@iu.edu

10.20982/tqmp.16.1.p033

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

Interaction effects, or moderation effects, can be estimated to assess whether a given relationship differs as a function of a third, moderator, variable. Researchers argue that estimation of moderation effects within a discipline provide an important marker of the progress of the field (Aguinis, 1995; Frazier, Tix, & Barron, 2004; Lorah & Miksza, 2019) as they provide a more nuanced understanding of the phenomena under investigation. Researchers cite the importance of examining hypotheses of moderation in various fields, such as applied psychology (Aguinis, Beaty, Boik, & Pierce, 2005), organizational research (Champoux & Peters, 1987), biological, psychological, and social sciences (Aguinis, 1995) and examining aptitude-treatment interactions in instructional psychology (Cronbach, 1987) in both theoretical and applied research (Bodner, 2016). In addition to the increased potential for valuable insights, the estimation of moderation effects is associated with additional complexity in interpretation of results both for the interaction effects themselves, as well as the main effects estimated in the model.

An example of a test of moderation can be found in Lorah and Wong (2018) which examines whether the relationship between perceived burdensomeness and suicide ideation is moderated by thwarted belongingness after controlling for depressive symptoms among Asian American college students. They find a significant interaction effect and plotting results indicates that students who score high on both perceived burdensomeness to others and unfulfilled need to belong to others have uniquely high risk of suicide ideation. Note that in this case, substantive interpretation proceeds naturally based on plotted results and there was no need for the authors to interpret main effects separately. However, if the given interaction effect had not been significant, it seems likely that the authors may have wanted to interpret the relationship between perceived burdensomeness and suicide ideation and between thwarted belongingness and suicide ideation (the main effects).

The moderated regression model can be estimated as follows:

$$Y = b_0 + b_1 \times X + b_2 \times M + b_3 \times XM + \varepsilon \quad (1)$$

where Y is the dependent variable; X is the independent variable; M is the moderator variable; XM is the interaction or product term; b_0 is the intercept; all other b represent slope coefficients; and ϵ is the random error term (Aiken & West, 1991). Note that equation 1 specifies a moderator that can be expressed with a single variable, M, but that in the case of an ordinal or nominal variable with more than two categories, M will naturally be expressed as more than one variable. Therefore, the present formulation (equation 1) represents a special case where the moderator variable is binary or continuous. Analogously, the same situation applies to the independent variable (X). Evaluation of significance of the moderation effect can proceed with a significance test for b_3 or any model comparison procedure used to compare a model with and without the product term (Lorah, 2018).

To compute effect size for the moderation effect, an appropriate measure is f^2 (Aiken & West, 1991). This represents the variance accounted for by the interaction effect relative to total unexplained outcome variance and can be computed as follows:

$$f^2 = \frac{R_2^2 - R_1^2}{1 - R_2^2} \quad (2)$$

where R_1^2 represents variance explained for the main effects only model (equation 1 with no product term) and R_2^2 represents variance explained for the full model (equation 1). In addition, f^2 can be interpreted as a small effect at values around 0.02; medium at values around 0.15 and large at values around 0.35 (Aiken & West, 1991).

Despite their importance, tests for interaction effects have been shown to generally have low power (Aiken & West, 1991; Frazier et al., 2004) and often fail to manifest (Jaccard, Turrisi, & Wan, 1990), perhaps due to the fact that the effect size for these effects is generally quite low in applied work ((Aiken & West, 1991). Additionally, low power may be particularly exacerbated in the presence of measurement error which is multiplied to create the product term (Aiken & West, 1991). Because of this, researchers have recommended conducting power analyses as a first step in moderation studies (Lorah & Miksza, 2019) and various software options are available to do so (see Lorah & Wong, 2018).

One particularly tricky aspect of moderation analysis is interpretation of results. Introductory texts tend to emphasize the difficulty of interpreting significant interaction effects and recommend and provide procedures to plot these effects (Aiken & West, 1991; Darlington & Hayes, 2017; Jaccard et al., 1990; Jose, 2013) and specific extensions of these plots (Bodner, 2016). Much methodological literature also specifically explores the idea of mean-centering and how it can be helpful for interpretation of main effects in the presence of signification interaction effects (Dalal & Zickar, 2012; Darlington & Hayes, 2017; McClelland, Irwin, Disatnik, & Sivan, 2017; Shieh, 2011). Further, the methodological literature attempts to address the misconception that main effects in the presence of significant interaction are average effects, when in fact they are conditional effects (Aiken & West, 1991; Darlington & Hayes, 2017; Frazier et al., 2004; Lorah & Wong, 2018).

The presence of a significant interaction effect additionally complicates the interpretation of main effects. A common misconception among applied researchers is that main effects may be interpreted in the same way as in a linear regression model (Darlington & Hayes, 2017; Frazier et al., 2004; Lorah & Wong, 2018). However, this is not the case; instead of average effects, these main effects are conditional effects that apply only to the case when the other of the two predictor variables is zero. Some argue that the appropriate analysis plan involves mean-centering the predictors and interpreting these conditional effects (Aiken & West, 1991) while it could also be argued that simply ignoring the main effects in the presence of a significant interaction effect is appropriate, since plotting and interpreting the interaction itself essentially illuminates all the relationships of interest.

However, there is little guidance regarding interpretation for the situation where the interaction effect is not found to be significant. In this case, it is clear that the interaction effect should not be interpreted substantively, since no evidence is provided that it exists. Further, rather than claiming a null effect, researchers have suggested claiming inconclusive findings and conducting a post hoc power analysis particularly for cross-level interactions in multilevel models (Aguinis, Gottfredson, & Culpepper, 2013), although this may be unhelpful as post hoc power analysis has been shown to be inappropriate as a general technique (Hoenig & Heisey, 2001). Some guidance indicates that if the interaction is nonsignificant, the researcher may proceed with multiple comparison procedures from an ANOVA model (Kirk, 2013) indicating that the interaction can be ignored and that the analysis can proceed as if the interaction effect had not been included.

Sadly, the guidance suggesting that main effects may be interpreted as simple additive effects in the presence of non-significant interaction effects may be misleading since the non-significant finding may be correct or it may simply be a Type II error (failure to find a true effect). One technique that may be helpful in this case is employing a model comparison procedure, such as Bayesian information criterion (BIC; Raftery, 1995). The BIC is designed to provide evidence for or against a simpler model (Weaklim, 2004). This indicates that the BIC could be used to provide evidence for the model excluding the interaction effect which could helpfully allow researchers to proceed with interpretation by ignoring the interaction effect. The BIC is computed by adding the model deviance to the number of predictors multiplied by the natural log of the sample size (Hox, 2010). A smaller value indicates a better fit (Hox, 2010).

Since the actual existence of the interaction effect will obviously not be known, it is unclear how applied researchers should proceed with interpretation when interactions are non-significant. Further, this is likely a particularly common scenario in moderation research given the low effect size that is typical for interaction effects; in addition, when tested interaction effects are non-significant, it is likely that the main effects are of particular substantive interest in the study and so correct interpretation of these main effects is crucial for the study.

The present study explores this issue with a simulation study designed to assess the possibilities for misleading interpretation of main effects with non-significant interactions, an analytic illustration of results, and guidance for applied researchers based on the findings. By simulating data according to common scenarios in the applied literature, the interpretation of main effects in the presence of Type II error (non-significant interaction terms) can be specified and contrasted to the interpretation if the null hypothesis for the interaction effect had been correctly rejected. Specifically, the following research question is examined: When interpreting main effects in the presence of a non-significant interaction effect, under what conditions is the possibility for misinterpretation of these effects severe?

Methods

Monte Carlo simulation was used to simulate data with known properties. A total of 192 different conditions were simulated, with 10,000 datasets simulated per condition (for a total of 1,920,000 unique datasets within the experiment). These 192 conditions were created by varying total sample size of each dataset (value of 50, 100, 500, or 1000); varying the population mean value of X, referred to as centrality condition (value of 0, 1, or 2); and varying the moderator effect size value (16 conditions: $b_3 = 0$ through 1.5, in increments of .1). These conditions were all fully crossed for a total of $4 \times 3 \times 16 = 192$ conditions. All other model parameters were held constant.

These conditions were chosen in particular to be representative of data typically encountered in the behavioral sciences and to allow for a clear demonstration of interpretation of results. For minimum sample sizes, researchers have suggested 104 plus the number of parameters (Hox, 2010) or 15 subjects per predictor (Stevens, 2002). Researchers have suggested that common samples sizes in applied work include 30 (number of industrialized countries); 50 (number of U. S. states); 100 (number of U.S. standard metropolitan statistical areas); and 1000 (small survey). For the present study, these considerations were used when selecting the minimum sample size (50) and the maximum sample size (1000) and additionally two intermediate sample size conditions were added. Since variables are often mean-centered for moderation analyses (Dalal & Zickar, 2012; Darlington & Hayes, 2017; McClelland et al., 2017; Shieh, 2011), the independent variable X was simulated with a population mean value of zero. In addition, population mean values of one and two were considered in order to demonstrate the impact of coding on model parameters and the associated interpretation. Since X was generated with standard deviation of one, these centrality conditions represent the addition of one and two standard deviations, respectively.

Similarly, the condition with no moderation effect was simulated in order to provide a baseline, whereas the remaining 15 effect size conditions were simulated in order to demonstrate the potential impact on interpretation. The values for b_3 were chosen to approximate common effect size conditions (specifically f^2 ranging from approximately 0 to .6; see Table 1 for specific average simulated values of f^2). This is consistent with the benchmarks for f^2 of small = 0.02; medium = 0.15; and large = 0.35 (Aiken & West, 1991).

Data were generated according to the moderation model specified in Equation 1. The independent variable X was generated as a standard random normal variable with one of three different population mean values: 0, 1, or 2. Therefore, X is normally distributed with mean of either 0, 1, or 2 and variance of one. The moderator variable, M, was generated as a binary variable with 50% of scores at value of -0.5 and 50% at value of 0.5. The random error term was generated as a standard normal variable (mean of zero, variance of one).

Values for the dependent variable, Y, were generated based on Equation 1 with $b_0 = b_1 = b_2 = 1$ and b_3 varied by condition. These chosen slope values imply that the main effect is constant across simulated conditions.

All data were simulated using R (R Core Team, 2017) and a moderation model according to Equation 1 was estimated using `lm()` within R. In addition, a main effects only model (Equation 1 without the product term) was estimated for model comparison purposes. Simulations with significant interaction terms were identified using the Wald test by dividing the coefficient (b_3) by its standard error and comparing this to +1.96. In addition, the interac-

Table 1 ■ Average observed effect size, f^2, for each b_3 condition

b_3	f^2
0	0.009
0.1	0.012
0.2	**0.019**
0.3	0.032
0.4	0.05
0.5	0.073
0.6	0.101
0.7	**0.134**
0.8	0.172
0.9	0.216
1	0.264
1.1	0.317
1.2	**0.375**
1.3	0.438
1.4	0.508
1.5	0.582

Note. Entries in bold used to represent small, medium, and large effect sizes, respectively. Values were computed as an average across all 3 population mean values for X and all 4 sample size conditions implying that each f^2 value is the mean of $3 \times 4 \times 10,000 = 120,000$ replications.

tion terms were also evaluated using BIC by choosing the model with the lower BIC value.

For the conditions with no true interaction effect ($b_3 = 0$), the Type I error rate was computed by assessing the proportion of samples with significant interaction effects as indicated by both the Wald test and BIC. For the conditions with non-zero interaction effect (b_3 not equal to zero), power was assessed analogously.

To assess the possibility for misinterpretation, samples with significant versus non-significant interaction effects are considered separately. The average slope coefficient for M (b_2) was computed for each condition and a correct and incorrect interpretation is offered. Subsequently, an analytic illustration is offered.

Results

The effect size value, f^2, was computed separately for each individual b_3 condition (Table 1). The values range from 0.009 when $b_3 = 0$ to .582 when $b_3 = 1.5$. Since small, medium, and large f^2 values are expected to be around 0.02, 0.15, and 0.35, respectively (Aiken & West, 1991), it is concluded that the present range of effect size conditions adequately covers typical interaction effect size conditions that might be observed in the applied literature. More specifically, a value of b_3 of 0.2 approximately corresponds with $f^2 = 0.02$; a value of b_3 of 0.7 roughly corresponds with $f^2 = 0.15$ and a value of b_3 of 1.2 roughly corresponds with $f^2 = 0.35$. Therefore, the present study will refer to these three values of b_3 as small, medium, and large effect sizes (see rows marked in bold in Table 1 and Table 2). Note that, as expected the average value for b_1 was 1 and the average value for b_3 varied by simulated b_3 condition, and was consistent with expectations.

The Type I error rate and power was assessed for each effect size and sample size condition and averaged across the three centrality conditions (see Table 2 and Figure 1). Note that before averaging results across the three centrality conditions, the results were assessed separately within these three conditions to ensure they did not differ systematically. In fact, the pattern of results assessed separately for these three conditions were virtually identical for the Wald test as well as for the BIC. There was no systematic pattern across these three conditions and the largest difference in power between the three centrality conditions within a given effect size/sample size condition was 0.018 for the Wald test and 0.017 for the BIC. These small differences are assumed due to sampling variation.

As expected, the conditions with no true interaction effect maintained a Type I error rate very close to the nominal rate of 0.05. The observed rates ranged from 0.051 to 0.057 using the Wald test (see Table 2, Panel A, first row). Although BIC represents an information criteria approach, as opposed to hypothesis testing, the rate of false positives can be assessed for the $b_3 = 0$ conditions. Since BIC explicitly considers the value of sample size in its computation, it is expected that the false positive rate varies as a function of sample size. This can be seen in the present data where the false positive rate varies from 0.01 to 0.06 (see

Table 2 ■ Proportion of samples indicating significant interaction effect (rejected null) by b_3 and sample size condition; $b_3 = 0$ represents Type I error rate; $b_3 > 0$ represents power.

	Panel A: Results based on Wald test					Panel B: Results based on BIC			
b_3	$N = 50$	$N = 100$	$N = 500$	$N = 1000$	b_3	$N = 50$	$N = 100$	$N = 500$	$N = 1000$
0	0.057	0.053	0.05	0.051	0	0.06	0.037	0.014	0.01
0.1	0.071	0.08	0.197	0.355	0.1	0.075	0.058	0.082	0.149
0.2	**0.108**	**0.168**	**0.601**	**0.884**	**0.2**	**0.114**	**0.129**	**0.393**	**0.698**
0.3	0.18	0.314	0.912	0.997	0.3	0.188	0.259	0.796	0.983
0.4	0.283	0.502	0.993	1	0.4	0.291	0.437	0.974	1
0.5	0.403	0.683	1	1	0.5	0.412	0.623	0.999	1
0.6	0.528	0.831	1	1	0.6	0.538	0.786	1	1
0.7	**0.652**	**0.918**	**1**	**1**	0.7	0.661	0.891	1	1
0.8	0.758	0.967	1	1	0.8	0.765	0.953	1	1
0.9	0.845	0.991	1	1	0.9	0.851	0.984	1	1
1	0.906	0.997	1	1	1	0.91	0.995	1	1
1.1	0.945	0.999	1	1	1.1	0.948	0.998	1	1
1.2	**0.971**	**1**	**1**	**1**	**1.2**	**0.973**	**1**	**1**	**1**
1.3	0.984	1	1	1	1.3	0.985	1	1	1
1.4	0.993	1	1	1	1.4	0.994	1	1	1
1.5	0.997	1	1	1	1.5	0.997	1	1	1

Note. First row of each sub-table ($b_3 = 0$) represents Type I error in detecting interaction effect. Subsequent rows ($b_3 > 0$) represent power for detecting interaction effect. Type II error rates are represented by 1-power. Entries in bold used to represent small, medium, and large effect sizes, respectively. Values were computed as an average across all 3 centrality conditions implying that each proportion is based on $3 \times 10,000 = 30,000$ replications. Results indicated that there were no differences between results for each of the three centrality conditions, which is why results are presented averaged across these conditions.

Table 2, Panel B, first row). This indicates, for most sample size conditions, a conservative test, which is consistent with previous simulation results for interaction effects assessed with BIC (Lorah, 2018).

Power can be observed based on the non-zero b_3 rows in Table 2. The minimum desired power of .80 (Cohen, 1992) is used as a benchmark in the present study. For a small effect size ($b_3 = 0.2$), power may be as low as 0.108 for a small sample size of N=50 and still remains below 0.8 for N=100 and N=500. For a medium effect size, power remains below 0.8 for N=50 but is higher, at 0.918 for N=100. For a large effect size, power is approaching 1, but still lower at 0.971 for N=50. In comparison to the Wald test, the power for the BIC may be slightly higher at lower sample sizes and slightly lower at higher samples (See Table 2, Panel B).

As the goal of the present study is to examine interpretation of main-effects, Table 3 and Figure 2 provide a summary of the average value of the main effect coefficient for M (b_2) under different conditions and for different subsets of the simulated datasets for all conditions with N=50. Note that although the value of b_2 is uncorrelated with sample size condition in the entire sample (r=-0.0004), due to differences in power there are systematic differences in this value among each of the data subsets considered when sample size is varied. For that reason, only simulations with N=50 are presented, but results for other sample size conditions demonstrated similar patterns.

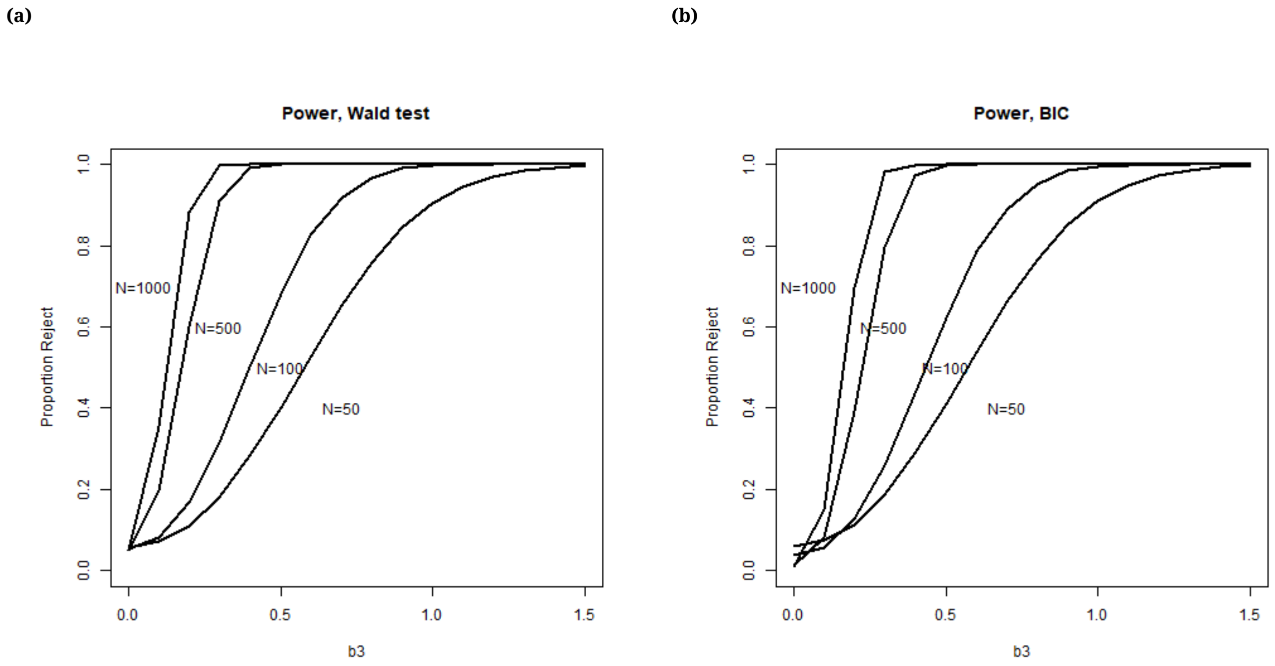
To assess interpretation of main effects, the datasets were divided into those where a significant interaction was found via the Wald test and those where a significant interaction was not found via the Wald test; then the datasets were divided again by those with significant versus non-significant interactions via the BIC. Results for both the Wald and BIC tests were similar and so BIC results are presented in the appendix for reference. Presumably, if an applied researcher finds the interaction effect to be significant, they will proceed by correctly interpreting the main effects as conditional effects. However, in the datasets where the interaction effect test indicates that the interaction is not significant, the researcher would likely proceed by interpreting the main effect as an average effect. This is correct for non-significant interaction ($b_3 = 0$) but incorrect for a true significant interaction ($b_3 > 0$). Thus, the issue being presently investigated is particularly when a true interaction exists, but it is found non-significant (i.e. Type II error, bold in Table 3), what is the appropriate interpretation for main effects?

Table 3 ■ Mean value of b_2 (slope of M) and its standard error as a function of b_3 and population mean value for X condition; results for $N = 50$ conditions, full interaction model; subset of samples with non-significant Wald test for interaction effect (NS Wald); significant Wald test for interaction effect (Sig Wald); and total set of samples (Total) considered separately.

		NS Wald		Sig Wald		Total		
b_3	Mean of X	b_2	SE of b_2	b_2	SE of b_2	b_2	SE of b_2	Prop. Reject
0	0	1	0.29	0.98	0.3	1	0.29	0.06
0.1	0	**1**	0.29	0.99	0.31	1	0.29	0.07
0.2	0	**1**	0.29	1	0.31	1	0.29	0.12
0.3	0	**1**	0.29	1.01	0.3	1	0.29	0.19
0.4	0	**1**	0.29	1	0.29	1	0.29	0.28
0.5	0	**1**	0.29	1	0.3	1	0.29	0.4
0.6	0	**1**	0.29	0.99	0.29	1	0.29	0.53
0.7	0	**1**	0.29	1	0.29	1	0.29	0.65
0.8	0	**1.01**	0.29	1	0.29	1	0.29	0.76
0.9	0	**0.99**	0.3	1	0.29	1	0.29	0.84
1	0	**1**	0.28	1	0.29	1	0.29	0.91
1.1	0	**1**	0.3	0.99	0.29	0.99	0.29	0.95
1.2	0	**1**	0.31	1	0.29	1	0.29	0.97
1.3	0	**0.97**	0.31	1.01	0.29	1	0.29	0.98
1.4	0	**1**	0.34	1	0.29	1	0.29	0.99
1.5	0	**0.97**	0.41	1	0.29	1	0.29	1
0	1	1	0.39	1.03	0.74	1	0.42	0.06
0.1	1	**0.93**	0.39	0.6	0.63	0.91	0.42	0.07
0.2	1	**0.85**	0.38	0.34	0.42	0.8	0.42	0.1
0.3	1	**0.79**	0.37	0.3	0.36	0.7	0.42	0.18
0.4	1	**0.74**	0.37	0.27	0.34	0.6	0.42	0.28
0.5	1	**0.68**	0.35	0.22	0.34	0.49	0.41	0.41
0.6	1	**0.64**	0.35	0.18	0.36	0.4	0.42	0.52
0.7	1	**0.61**	0.34	0.14	0.36	0.3	0.42	0.65
0.8	1	**0.57**	0.33	0.09	0.36	0.2	0.41	0.76
0.9	1	**0.55**	0.35	0.02	0.38	0.1	0.42	0.84
1	1	**0.52**	0.33	-0.06	0.39	0	0.42	0.91
1.1	1	**0.50**	0.34	-0.13	0.4	-0.1	0.42	0.95
1.2	1	**0.43**	0.34	-0.22	0.4	-0.2	0.42	0.97
1.3	1	**0.45**	0.35	-0.31	0.41	-0.3	0.42	0.99
1.4	1	**0.46**	0.33	-0.41	0.41	-0.4	0.42	0.99
1.5	1	**0.53**	0.36	-0.51	0.42	-0.51	0.43	1
0	2	1	0.6	0.87	1.4	1	0.67	0.06
0.1	2	**0.84**	0.59	0.18	1.15	0.79	0.67	0.07
0.2	2	**0.71**	0.57	-0.29	0.74	0.6	0.66	0.11
0.3	2	**0.58**	0.55	-0.44	0.52	0.41	0.67	0.17
0.4	2	**0.46**	0.52	-0.49	0.49	0.19	0.67	0.29
0.5	2	**0.36**	0.5	-0.57	0.48	-0.01	0.67	0.4
0.6	2	**0.29**	0.48	-0.64	0.5	-0.2	0.68	0.53
0.7	2	**0.21**	0.47	-0.74	0.51	-0.41	0.67	0.65
0.8	2	**0.16**	0.45	-0.84	0.54	-0.59	0.68	0.75
0.9	2	**0.10**	0.44	-0.96	0.57	-0.79	0.67	0.85
1	2	**0.05**	0.43	-1.1	0.59	-0.99	0.67	0.91
1.1	2	**-0.04**	0.41	-1.26	0.62	-1.19	0.67	0.94
1.2	2	**-0.06**	0.4	-1.44	0.63	-1.4	0.67	0.97
1.3	2	**-0.15**	0.41	-1.62	0.65	-1.59	0.67	0.98
1.4	2	**-0.09**	0.45	-1.81	0.66	-1.8	0.68	0.99
1.5	2	**-0.12**	0.41	-2	0.66	-2	0.66	1

Note. NS Wald = samples where the Wald test indicated no significant interaction effect; Sig Wald = samples where the Wald test indicated a significant interaction effect; Total = all samples; Prop. Reject = proportion of samples where the null is rejected based on the Wald test. Values in these columns represent the mean and the standard error (SE), which is computed as the standard deviation of the b_2 values for the given condition. Values where b_3 is not zero represent those cases where a Type II error is made and where the potential for misinterpretation is explored. Proportion of samples rejected represents Type I error when $b_3 = 0$ and power when $b_3 > 0$. Type II error rates can be computed as 1-power.

Figure 1 ■ Power curves for interaction effect varying b3 and sample size, as evaluated by the Wald test (left panel) and the BIC test (right panel)



Based on Table 3, the coefficient for M (b_2) is roughly 1 whenever the independent variable, X, is simulated with population mean of zero. This coefficient of 1, which is technically a conditional effect, is the same as the average effect of M (i.e. when $b_3 = 0$, $b_2 = 1$). Thus, with a population mean value of zero which corresponds to mean-centering of the independent variable, the average interpretation rather than conditional interpretation of the moderator effect is only slightly misleading. For example, consider the following two interpretation for $b_2 = 1$. The average effect interpretation is that for one unit increase in M, Y is expected to increase by one unit when controlling for associated covariates. In contrast, the conditional effect interpretation is that for one unit increase in M, Y is expected to increase by one unit, but only for samples where $X = 0$, controlling for associated covariates. Switching one interpretation for the other is not ideal, but only slightly misleading.

When X is simulated with a population mean of one, the results look different. Again the average effect for M (visible for conditions with $b_3 = 0$) is one. However, depending on the effect size, the value for b_2 varies quite a bit from one (values range from about $.43$ to 1 for non-significant Wald conditions). When X is simulated with a population mean of two, the deviations are even larger. For example, the value for b_2 ranges from about $-.12$ to $.86$. Note again that these values for b_2 are found in the conditions displaying Type II error. In other words, the researcher is likely to ignore the interaction since it did not reach significance, and then interpret these main effects as average effects. For example, consider the condition with $b_3 = 1.5$ and X is simulated with a population mean of two. Interpreting the effect as an average effect would result in the following statement: For a one-unit increase in M, Y is expected to decrease by .12 units, controlling for associated covariates. Because we know this is the result of a Type II error and that actually the statement holds only when $X = 2$, the statement appears to be fairly misleading.

The conditions where the Wald test was significant follow a similar pattern, but the coefficients are slightly different, as these are the datasets where the Wald test for interaction effect reached significance. These are also the conditions where the researcher is likely to correctly interpret main effects as conditional effects. Note that patterns for the BIC are overall fairly similar (see specific results in the Appendix).

The results from the main-effects-only model (i.e. Equation 1 without the product term) are displayed in Table 4. Although this model is incorrectly specified (i.e. the model itself omits the interaction effect), the results for the value of b_2 are typically very close to 1. So in general, when no evidence is provided for a significant interaction, interpreting the moderator effect as an average effect is

Table 4 ■ Mean value of b_2 (slope of M) and its standard error for each population mean value for X condition, averaged across all other conditions, main-effects-only model

Centrality	Mean b2	SE
0	1.0002	0.2005
1	0.9999	0.2004
2	0.9997	0.1999

Note. The mean and standard error (SE) are computed based on all samples implying the values are averaged across effect size and sample size conditions, which was done after confirming that there were no systematic differences in results across these conditions. SE is computed as the standard deviation of the b_2 values for the given centrality condition.

only particularly misleading when the model includes the product term and the independent variable is not mean-centered. However, if no significant interaction is found, and if the independent variable is mean-centered, or the product term is omitted from the model, the interpretation of the main effect as an average effect is fairly reasonable.

Note that these potential interpretations of model results are offered to demonstrate the possibilities for misleading interpretation, not to suggest that the conditional main effects interpretation is always appropriate. In the preceding discussion, the "correct" interpretation was offered only based on knowledge of a true moderation effect (i.e. true parameters are known in a simulation study). In an applied setting, the true parameters would be unknown and so the preceding discussion can simply be taken as a warning to proceed with caution when interpreting these models. Based on these results, further guidance and analysis options for applied researchers are offered in the discussion section.

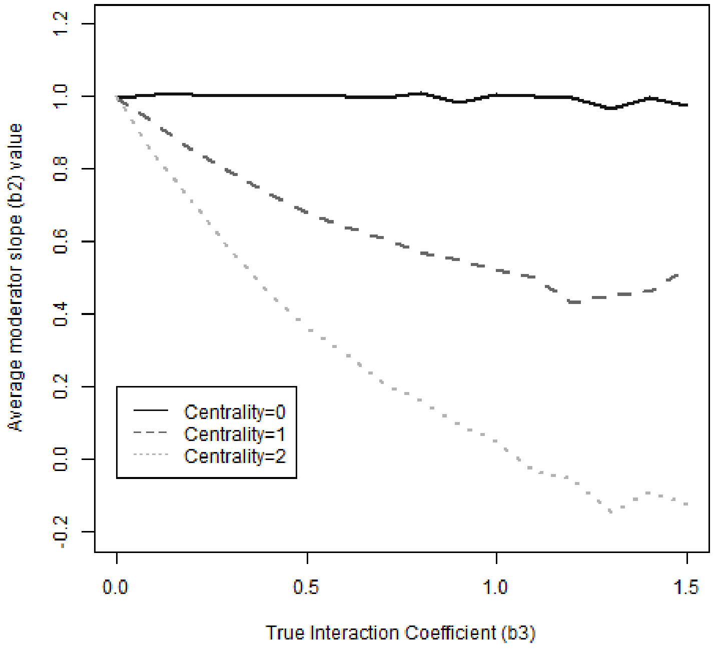
Analytic Illustration

The expectation for the main effect for M (b_2) can be derived based on the concept of simple slopes (Aiken & West, 1991). This is now demonstrated using the $b_3 = 1.5$ effect size condition with X simulated with a population mean of zero versus one. Based on the parameters specified for the present investigation, the regression equation (Equation 1) to predict Y is $Y' = 1 + X + M + 1.5 \times XM$ which simplifies to $Y' = 1.5 + 1.75 \times X$ when $M = 0.5$ and to $Y' = .5 + .25 \times X$ when $M = -0.5$. Recall the interpretation of b_2 (coefficient for M) for a moderation model is conditional on the case when $X = 0$. Specifically, it is interpreted as the expected increase in Y for one unit increase in M when $X = 0$. Therefore, we can compute the predicted value of Y in the case when $M = .5$ and the case when $M = -.5$ by substituting $X = 0$ into the simple slopes equations just provided and then take the difference. Doing this results in a predicted Y value of 1.5 and 0.5, respectively. Taking the difference results in $1.5 - 0.5 = 1$ which is consistent with the simulated average value for b_2 in conditions where X is simulated with a population mean of zero.

However, when X has a different mean, for example a population mean of one, the intercept is moved one unit to the left. This can be clarified, again, by examination of simple slopes. In this case, X becomes $X - 1$ and the predicted Y can be specified as $Y' = 1 + 1 \times (X - 1) + 1 \times M + 1.5 \times (X - 1) \times M$. The line for observations where $M = 0.5$ can be specified by substituting this value for M resulting in $Y' = -.25 + 1.75 \times X$ and for cases where $M = -0.5$ as the result is $Y' = .25 + .25 \times X$. By substituting $X = 0$ into these two equations we can compute the difference in expected Y for one unit increase in M. In this case, the expectation of Y when $M = 0.5$ is $-.25$ and the expectation of Y when $M = -0.5$ is .25. Therefore, we would expect that at the point where $X = 0$, a one-unit increase in M is related to a 0.5 unit decrease in Y. This value of -0.5 roughly corresponds with the average simulated value provided in Table 3. In general, this solution holds for any set of parameters and centrality condition by simply applying the simple slopes concept with the appropriate values.

These results can be understood visually by plotting the results obtained from the preceding simple slopes analysis, as has been done in Figure 3, left panel. Note that the vertical line at $X = 0$ corresponds to the intercept for the analysis where X is simulated with a population mean of zero and the vertical line at $X = -1$ corresponds to the intercept for the analysis where X is simulated with a population mean of one. In contrast, Figure 3, right panel, provides an example of an analogous plot where no interaction effect exists, for demonstration purposes. This corresponds to the model where the $b_3 = 0$ effect size condition is used. Note that since the relationship between X and Y does not depend on the value of M, the slope of X is identical for any value of M. The plot shows this for two example values of M.

Figure 2 ■ Average moderator slope (b_2) value for simulations with no significant interaction for $N = 50$ conditions. Where $b_3 > 0$, these values are associated with Type II error conditions.



Discussion

The results from this study explored the likelihood that tests for interaction effect may display low power. Further, the results demonstrated the possibility for misinterpretation of main effects in a moderation model. Although it is common for researchers to misinterpret the conditional main effects as average main effects in the presence of significant interaction effects (Darlington & Hayes, 2017; Frazier et al., 2004; Lorah & Wong, 2018), it is expected that researchers will interpret main effects as average main effects when the interaction effects are nonsignificant. However, particularly due to the possibility of a Type II error, these interpretations have the possibility to be misleading.

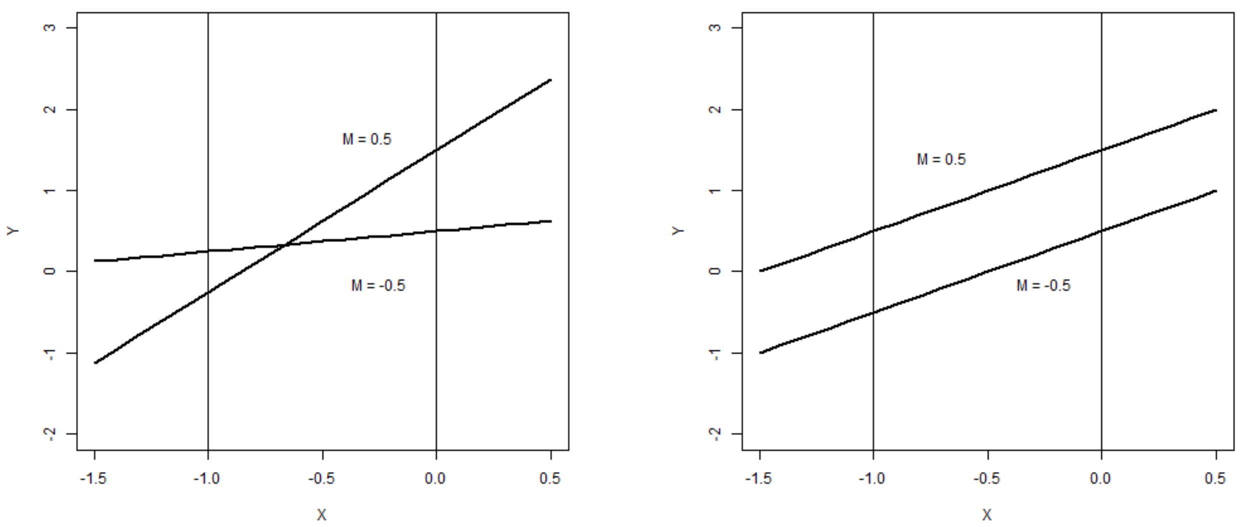
Conditional main effects provide interpretation specifically when the value of the independent variable is zero. It is possible for zero to be an implausible value for a given variable (i.e. age for an elderly population) in which case interpreting conditional effects as if they are average effects could be potentially very misleading. Clearly, regardless of whether the interaction is significant, researchers should use caution when interpreting main effects in any moderation model.

In addition to diligence in interpretation, there are steps researchers can routinely take to guard against the possibility of misinterpretation. The first step is to ensure adequate power to detect a moderation effect before the start of the study. This involves conducting power analysis specifically for the moderation effect, rather than just for main effects, which is a recommendation consistent with the literature (Lorah & Miksza, 2019; Lorah & Wong, 2018). In addition, routinely mean-centering predictors used in any product term can help guard against substantially incorrect interpretation. As seen in Table 3, and generally expected, when an interpretation is conditional on a mean value for another variable, rather than an uncommon value for that variable, it is more useful.

Another way to find evidence for the null model would be to use an information criteria approach, such as BIC which, unlike hypothesis testing, can provide evidence for a null model (Raftery, 1995). This would allow the researcher to feel more confident in ignoring the interaction effect and proceeding with typical interpretation of main effects. This was demonstrated in the present study, although results indicated some amount of consistency between results from the Wald test and BIC, implying that use of BIC alone is insufficient to prevent misinterpretation of main effects.

Lastly, the researcher should consider estimating a

Figure 3 ■ Left: Plot demonstrating interaction based on example simulated model with $b_3 = 1.5$ effect size (described in Analytic Illustration section); right: Plot demonstrating no interaction effect; model corresponds to a model analogous to the interaction model, but with a coefficient of zero for the product term implying the $b_3 = 0$ effect size condition (described in Analytic Illustration section)



(a) Demonstrating interactions

(b) Demonstrating no interaction

model without the interaction effect. This model will provide average, rather than conditional main effects, and in particular, the researcher may want to compare the regression coefficients for the main effects from the full interaction model to those from the main effects only model (the model with no interaction effect). If the regression coefficients are similar, either model could be reported; however, if they vary substantially, further investigation may be helpful.

There are a few limitations associated with this study. First, the results obtained do not necessarily generalize beyond the simulated conditions examined. Specifically, the present study examined a binary moderator variable and a continuous independent variable. Future research should further examine categorical moderators with more than two categories and continuous moderators as well as categorical independent variables. In addition, future research should examine additional models which may be estimated with moderation effects, such as structural equation models, hierarchical linear models, and models with categorical outcome variables.

In summary, the following recommendations are provided for researchers:

- Conduct power analyses specifically for the moderation effect before beginning the study.
- Routinely mean-center predictors used in product terms. Since it can never be certain whether the true effect exists or not, this step guards against substantial misinterpretation of main effects.
- Estimate the main effects only model, in addition to the interaction model. If parameter estimates for main effects seem substantially different between the models, further investigation is warranted. If there is no sound theoretical reason to include the interaction term in the final model and it is not found to be significant, consider removing it and reporting results for the main effects only model instead.
- Consider assessing interactions effects with information criteria approach, such as BIC, rather than only relying on hypothesis testing procedures, in order to have the opportunity to provide evidence for a null effect. This step alone may be insufficient to ensure main effect interpretation is not misleading. Use of BIC should be combined with either mean-centering and/or estimation of the main effects only model.

Regardless of whether interactions effects are significant or not, when estimating moderation models, researchers should exercise caution when interpreting results from these models, and particularly when interpreting main effects. It is clear that the possibility for misleading interpretation of main effects in moderation models is large, and specific steps researchers may take to avoid misleading interpretation have been provided.

Authors' note

This research was presented at the VIII European Congress of Methodology [EAM 2018] in July 2018.

References

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management Research*, *21*, 1141–1158.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*(1), 94–107.

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6), 1490–1528.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Bodner, T. E. (2016). Tumble graphs: Avoiding misleading end point extrapolation when graphing interactions from a moderated multiple regression analysis. *Journal of Educational and Behavioral Statistics*, *41*(6), 593–604.

Champoux, J. E., & Peters, W. S. (1987). Form, effect size, and power in moderated regression analysis. *Journal of Occupational Psychology*, *60*, 243–255.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin*, *102*, 414–417.

Dalal, D. K., & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods*, *15*(3), 339–362.

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. New York, NY: Guilford.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*(1), 115–134.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power. *The American Statistician*, *55*(1), 19–24. doi:10.1198/000313001300339897

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park: Sage Publications.

Jose, P. E. (2013). *Doing statistical mediation & moderation*. New York: The Guilford Press.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences*. Los Angeles: Sage.

Lorah, J. A. (2018). Estimating individual-level interaction effects in multilevel models: A monte carlo simulation study with application. *Journal of Applied Statistics*, *45*(12), 2238–2255. doi:10.1080/02664763.2017.1414163

Lorah, J. A., & Miksza, P. (2019). Applications of moderation analysis for music education research. *Bulletin of the Council for Research in Music Education*, *220*, 21–41.

Lorah, J. A., & Wong, Y. J. (2018). *Contemporary applications of moderation analysis in counseling psychology*. doi:10.1037/cou0000290

McClelland, G. H., Irwin, J. R., Disatnik, D., & Sivan, L. (2017). Multicollinearity is a red herring in the search for moderator variables: A guide to interpreting moderated multiple regression models and a critique of iacobucci, schneider, popvich, and bakamitsos (2016). *Behavior Research Methods*, *49*, 394–402.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Shieh, G. (2011). Clarifying the role of mean centering in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology*, *64*, 462–477.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Weaklim, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods and Research*, *33*, 167–187.

Appendix A: Mean value of b_2 (slope of M) and its standard error by b_3 and population mean value for X condition; results for $N = 50$ conditions, full interaction model; subset of samples with non-significant BIC test for interaction effect (NS BIC); significant BIC test for interaction effect (Sig BIC); and total set of samples (Total) considered separately.

		NS Wald		Sig Wald		Total		
b_3	Mean of X	b_2	SE of b_2	b_2	SE of b_2	b_2	SE of b_2	Prop. Reject
0	0	1	0.29	0.98	0.3	1	0.29	0.06
0.1	0	1.01	0.29	0.99	0.31	1	0.29	0.07
0.2	0	1	0.29	1	0.3	1	0.29	0.12
0.3	0	1	0.29	1	0.29	1	0.29	0.19
0.4	0	1	0.29	1	0.29	1	0.29	0.28
0.5	0	1	0.29	1	0.3	1	0.29	0.4
0.6	0	1	0.29	0.99	0.29	1	0.29	0.53
0.7	0	1	0.29	1	0.29	1	0.29	0.65
0.8	0	1.01	0.3	1	0.29	1	0.29	0.76
0.9	0	0.99	0.3	1	0.29	1	0.29	0.84
1	0	1	0.28	1	0.29	1	0.29	0.91
1.1	0	1	0.3	0.99	0.29	0.99	0.29	0.95
1.2	0	1	0.31	1	0.29	1	0.29	0.97
1.3	0	0.97	0.32	1.01	0.29	1	0.29	0.98
1.4	0	0.97	0.33	1	0.29	1	0.29	0.99
1.5	0	0.97	0.41	1	0.29	1	0.29	1
0	1	1	0.39	1.02	0.74	1	0.42	0.06
0.1	1	0.93	0.38	0.6	0.62	0.91	0.42	0.07
0.2	1	0.86	0.38	0.35	0.42	0.8	0.42	0.1
0.3	1	0.79	0.37	0.31	0.37	0.7	0.42	0.18
0.4	1	0.74	0.37	0.27	0.35	0.6	0.42	0.28
0.5	1	0.68	0.35	0.23	0.34	0.49	0.41	0.41
0.6	1	0.65	0.35	0.19	0.36	0.4	0.42	0.52
0.7	1	0.62	0.34	0.14	0.36	0.3	0.42	0.65
0.8	1	0.58	0.33	0.09	0.36	0.2	0.41	0.76
0.9	1	0.56	0.34	0.03	0.38	0.1	0.42	0.84
1	1	0.53	0.33	-0.06	0.39	0	0.42	0.91
1.1	1	0.51	0.34	-0.13	0.4	-0.1	0.42	0.95
1.2	1	0.44	0.35	-0.22	0.4	-0.2	0.42	0.97
1.3	1	0.46	0.36	-0.31	0.41	-0.3	0.42	0.99
1.4	1	0.47	0.33	-0.4	0.41	-0.4	0.42	0.99
1.5	1	0.51	0.36	-0.51	0.42	-0.51	0.43	1
0	2	1	0.59	0.89	1.39	1	0.67	0.06
0.1	2	0.84	0.59	0.19	1.14	0.79	0.67	0.07
0.2	2	0.71	0.57	-0.27	0.75	0.6	0.66	0.11
0.3	2	0.59	0.55	-0.42	0.52	0.41	0.67	0.17
0.4	2	0.47	0.52	-0.48	0.49	0.19	0.67	0.29
0.5	2	0.37	0.5	-0.56	0.48	-0.01	0.67	0.4
0.6	2	0.3	0.48	-0.63	0.5	-0.2	0.68	0.53
0.7	2	0.22	0.46	-0.73	0.52	-0.41	0.67	0.65
0.8	2	0.17	0.45	-0.83	0.54	-0.59	0.68	0.75
0.9	2	0.11	0.43	-0.95	0.57	-0.79	0.67	0.85
1	2	0.06	0.43	-1.1	0.59	-0.99	0.67	0.91
1.1	2	-0.02	0.4	-1.26	0.62	-1.19	0.67	0.94
1.2	2	-0.04	0.4	-1.44	0.63	-1.4	0.67	0.97
1.3	2	-0.14	0.41	-1.62	0.65	-1.59	0.67	0.98
1.4	2	-0.06	0.42	-1.81	0.66	-1.8	0.68	0.99
1.5	2	-0.12	0.41	-2	0.66	-2	0.66	1

Note: NS BIC = samples where the BIC indicated no significant interaction effect; Sig BIC = samples where the BIC indicated a significant interaction effect; Total = all samples; Prop. Reject = proportion of samples where the null is rejected based on the BIC test. Values in these columns represent the mean and the standard error (SE), which is computed as the standard deviation of the b_2 values for the given condition. Proportion of samples rejected represents Type I error when $b_3 = 0$ and power when $b_3 > 0$. Type II error rates can be computed as $1-$power.

Citation

Lorah, J. A. (2020). Interpretation of main effects in the presence of non-significant interaction effects. *The Quantitative Methods for Psychology*, *16*(1), 33–45. doi:10.20982/tqmp.16.1.p033

Received: 01/08/2019 ~ Accepted: 05/12/2019