Behav Res (2016) 48:1308–1317
DOI 10.3758/s13428-015-0624-x

Mean centering helps alleviate "micro" but not "macro" multicollinearity

Dawn Iacobucci[4] · **Matthew J. Schneider**[1] · **Deidre L. Popovich**[2] · **Georgios A. Bakamitsos**[3]

Published online: 7 July 2015

Abstract There seems to be confusion among researchers regarding whether it is good practice to center variables at their means prior to calculating a product term to estimate an interaction in a multiple regression model. Many researchers use mean centered variables because they believe it's the thing to do or because reviewers ask them to, without quite understanding why. Adding to the confusion is the fact that there is also a perspective in the literature that mean centering does not reduce multicollinearity. In this article, we clarify the issues and reconcile the discrepancy. We distinguish between "micro" and "macro" definitions of multicollinearity and show how both sides of such a debate can be correct. To do so, we use proofs, an illustrative dataset, and a Monte Carlo simulation to show the precise effects of mean centering on both individual correlation coefficients as well as overall model indices. We hope to contribute to the literature by clarifying the issues, reconciling the two perspectives, and quelling the current confusion regarding whether and how mean centering can be a useful practice.

Keywords Mean centering · Multicollinearity · Moderated multiple regressions · Interactions in regression

✉ Dawn Iacobucci
Dawn.Iacobucci@owen.vanderbilt.edu

[1] Northwestern University, Evanston, IL, USA

[2] Texas Tech University, Lubbock, TX, USA

[3] Stetson University, DeLand, FL, USA

[4] Vanderbilt University, 401 21st Avenue South, Nashville, TN 37203, USA

Mean centering is the act of subtracting a variable's mean from all observations on that variable in the dataset such that the variable's new mean is zero. Some researchers say that it is a good idea to mean center variables prior to computing a product term (to serve as a moderator term) because doing so will help reduce multicollinearity in a regression model. Other researchers say that mean centering has no effect on multicollinearity. The debate has left many researchers perplexed: is it good practice to mean center or not, and if it is, why?

In this paper, we wish to help researchers by clarifying the issues. We distinguish between micro and macro forms of multicollinearity, and show that whether multicollinearity is lessened by mean centering depends upon this level of analysis and interpretation.

Specifically, the conflict between the two viewpoints arises due to the ambiguity as to what constitutes multicollinearity. Multicollinearity is defined to be the presence of correlations among predictor variables that are sufficiently high to cause subsequent analytic difficulties, from inflated standard errors (with their accompanying deflated power in significance tests), to bias and indeterminancy among the parameter estimates (with the accompanying confusion regarding the interpretation and contributions of individual predictors).

Multicollinearity, or excessive correlations among predictor variables, may be detected sometimes by examining a correlation matrix, for example, the correlation between variable X_1 or X_2 with their product score X_1X_2 is likely to be large. Other times multicollinearity is more subtle, being a nonobvious linear combination of two or more of the independent variables, detectable only by the examination of the determinant of the covariance or cross-products matrix.

We define a "micro" focus on multicollinearity to be that which is based on looking at the correlation or regression coefficients. The "macro" focus on multicollinearity is based on looking at the properties of the whole matrix such as its

determinant, or the overall fit of a model as in an R^2. We first briefly review the literature and then we consider both perspectives in greater detail to show how they may be reconciled.

Brief literature review

A number of scholars have considered issues related to mean centering with regard to the inclusion of product terms in a multiple regression model to test for moderators. Let us begin with three points on which these scholars agree.

First, when multiple regression models are expanded from a supposition of two main effects, $Y=b_0+b_1X_1+b_2X_2+\epsilon$, to a model in which there exists a multiplicative term to capture the interaction, $Y=b_0+b_1X_1+b_2X_2+b_3X_1X_2+\epsilon$, the main effect variables (both X_1 and X_2) are often highly correlated with their composite product term (X_1X_2). Some researchers see this as a problem and others do not, and we will show that these different positions are largely a function of which results are under consideration; however, researchers do not disagree that, empirically, the high correlations are likely.

Second, researchers who believe that mean centering will help clarify the regression results will obviously recommend that the variables X_1 and X_2 be mean centered before the product term is computed. Researchers who do not believe the mean centering helps have no argument against mean centering per se; for example, if researchers are working with variables whose measurements include arbitrary zeros, then it may be fruitful to mean center a variable such that results are interpretable with respect to the variable's mean rather than to an arbitrary point of zero. Researchers of both camps mention the variables' measurement properties as a plausible and defensible reason for mean centering (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Irwin & McClelland, 2001; Jaccard, Wan, & Turrisi, 1990; Kromrey & Foster-Johnson, 1998).

Third, Aiken and West (1991) attribute to Marquardt (1980) the terminology of distinguishing "essential" and "nonessential" multicollinearity. The terms are not great, given that they are somewhat value-laden, but these terms are used in this literature (cf., Bradley & Srivastava, 1979; Dalal & Zickar, 2012; Shieh, 2010). "Essential" multicollinearity describes correlations between variables for constructs that are very likely to be correlated; Aiken and West (1991, p.36) give the example of a likely correlation "between the age of a child and his/her developmental stage." In contrast, "nonessential" multicollinearity describes correlations that arise due to issues of measurement or in the moderated multiple regression context, the fact that X_1 and X_2 are likely correlated with their product term X_1X_2 because, of course, they are contained within it. No researchers believe that mean centering affects essential multicollinearity, and they differ on whether they believe that mean centering reduces nonessential multicollinearity.

Beyond those basic points, there is less consistency among researchers' points of view. Usually there is agreement on facts, but there exist disagreements regarding the assessments of those facts.

For example, looking at the results of a moderated multiple regression of the form $Y=b_0+b_1X_1+b_2X_2+b_3X_1X_2+\epsilon$ from the "micro" perspective, that is, examining the individual predictors and regression coefficients, researchers agree that mean centering X_1 and X_2 has no effect on the product term X_1X_2, nor the power with which the moderator effect may be detected (cf. Allison, 1977; Dalal & Zickar, 2012; Kromrey & Foster-Johnson, 1998; Shieh, 2011, 2010, 2009; Smith & Sasaki, 1979). Furthermore, most researchers concur that mean centering X_1 and X_2 will reduce their correlations with the product term X_1X_2. Researchers in the "micro" camp will point to this fact as evidence that the mean centering helps reduce (micro) multicollinearity. Mean centering facilitates the likelihood of finding significance for the main effect terms, X_1 and X_2. This multicollinearity is the sort labeled "nonessential," because it is a function of data processing (i.e., taking a product), not of inherent relationships among constructs (i.e., essential multicollinearity). Some researchers of the "macro" camp take the value-laden word "nonessential" a step further when they then state that researchers "do not care about" this kind of multicollinearity or that it is a "myth" that mean centering can alter substantive conclusions (cf., Dalal & Zickar, 2012, pp. 342, 358) or that whether researchers mean center or not, their results will be "equivalent" (Kromrey & Foster-Johnson, 1998, p. 42). In fact, results on X_1 and X_2 can vary, so we suspect that those macro researchers' characterizations are based on the frequent orientation in such modeling that the moderator relationship is the key term under investigation (a philosophy with which we agree), with the main effects included to partial out their variance but not for theoretical reasons (e.g., Stone & Hollenbeck, 1984). However, it may certainly be the case that some researchers are interested in both the main effect terms as well as the interaction term (a philosophy with which we also agree), hence, they may well care about reducing the micro level of multicollinearity (e.g., Aiken & West, 1991; Friedrich, 1982; Irwin & McClelland, 2001).

Similarly, results of a moderated multiple regressions considered at a "macro" level, from the determinant of the input matrix to the overall model fit R^2, are not affected by being mean centered (Dalal & Zickar, 2012; Echambadi & Hess, 2007)—a fact on which both camps agree, yet the macro camp uses these results to support their claim that mean centering has no effect on multicollinearity. Researchers in the micro camp might well agree that mean centering has no effect on macro multicollinearity, but their interest is in micro multicollinearity, which is affected.

With that brief overview of the literature providing an introduction to the issues, we now proceed to define and demonstrate the relationships at the "micro" and "macro" levels in moderated multiple regressions. We note that the literature we just reviewed briefly also considers other related issues, including multicollinearity that arises due to power terms (e.g., X^2), not just product terms (Bradley & Srivastava, 1979), and the effects of non-normality and analogous effects on three-way interactions (Shieh, 2010). Some articles contain equations and proofs, and some offer demonstrations with very small data sets or simulations (e.g., Allison, 1977; Shieh, 2011). In our investigations that follow, we use all of these methods to offer a more comprehensive view of the micro and macro issues.

Mean centering helps alleviate "micro" multicollinearity

The arguments in favor of mean centering and its role in reducing multicollinearity are based on the "micro" focus, looking at the effect of centering on single correlation or regression coefficients, one at a time (cf., Irwin & McClelland, 2001; Jaccard, Wan, & Turrisi, 1990; Smith & Sasaki, 1979). From this micro focus, it is correct to state that transforming X_1 or X_2 to deviation scores,[1] $(X_1-\overline{X}_1)$ and $(X_2-\overline{X}_2)$, before computing a product term to represent their interaction, $(X_1-\overline{X}_1)(X_2-\overline{X}_2)$, will typically produce a correlation, $r\left(\left((X_1-\overline{X}_1)(X_2-\overline{X}_2)\right), Y\right)$, that is much smaller than its corresponding correlation on the raw (uncentered) variables, $r((X_1X_2), Y)$. Similarly the correlation between the product score and either of its components will be smaller for the mean centered variables, specifically $r\left(\left((X_1-\overline{X}_1)(X_2-\overline{X}_2)\right), (X_1-\overline{X}_1)\right)$ will be smaller than $r((X_1X_2),X_1)$ and $r\left(\left((X_1-\overline{X}_1)(X_2-\overline{X}_2)\right), (X_2-\overline{X}_2)\right)$ will be smaller than $r((X_1X_2),X_2)$. To understand these relationships, consider the following simple proofs.

Let us begin with the familiar equation for a Pearson product-moment correlation coefficient:

$$r_{X,Y} = \frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}}. \tag{1}$$

Next, let us simply substitute the product score X_1X_2, wherever there exists the single variable X in equation (1) to obtain the equation for the correlation between a product score X_1X_2 and Y:

$$r_{(X_1X_2),Y} = \frac{cov((X_1X_2),Y)}{\sqrt{var(X_1X_2)\cdot var(Y)}}. \tag{2}$$

Similarly, let us consider the equation for the correlation between the product score, X_1X_2, and one of its components—we will take X_1 but these results will obviously also hold for X_2. Here we take Eq. (2) and substitute X_1 in everywhere there had been a Y:

$$r_{(X_1X_2),X1} = \frac{cov((X_1X_2),X_1)}{\sqrt{var(X_1X_2)var(X_1)}}. \tag{3}$$

Now, let us consider the "micro" claim that multicollinearity is reduced when the predictor variables are mean centered. The question is whether the correlation between X_1 (or X_2) and X_1X_2 is smaller when the variables have been mean centered. The equation for that correlation is derived as follows, by simply taking Eq. (3) and replacing X_1 and X_2 with their respective deviation score counterparts, $(X_1-\overline{X}_1)$ and $(X_2-\overline{X}_2)$:

$$r_{\left(\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)\right),\left(X_1-\overline{X}_1\right)} = \frac{cov\left(\left(\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)\right),\left(X_1-\overline{X}_1\right)\right)}{\sqrt{var\left(\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)\right)var\left(\left(X_1-\overline{X}_1\right)\right)}}. \tag{4}$$

In the comparison we seek, it is sufficient to compare the numerators of Eqs. (3) and (4), as will become clear shortly. Thus, let us begin by looking at the numerator in Eq. (3), which is the covariance between the product score and one of its components, X_1, when X_1 and X_2 are in their original (uncentered) form.

It has been shown that the covariance between a product score and another variable may be written as follows (under multivariate normality, Arnold & Evans, 1979; Bohrnstedt & Goldberger, 1969; Goodman, 1960):

$$cov(AB,C) = \varepsilon(A)cov(B,C) + \varepsilon(B)cov(A,C). \tag{5}$$

The notation ε(A) stands for the expected value of all A's in the population distribution; alternatively, μ_A. Thus, for our purposes, the numerator in Eq. (3) may be written as:

$$cov((X_1X_2),X_1) = \varepsilon(X_1)cov(X_1,X_2) + \varepsilon(X_2)cov(X_1,X_1). \tag{6}$$

Note that the last term in Eq. (6) is actually a variance, hence:

$$cov((X_1X_2),X_1) = \varepsilon(X_1)cov(X_1,X_2) + \varepsilon(X_2)var(X_1). \tag{7}$$

[1] Deviation scores are often denoted more simply as a lower case variable label, e.g., $x_1 = (X_1-\overline{X}_1)$ and $x_2 = (X_2-\overline{X}_2)$, and we agree with a reviewer that such notation would make this presentation more elegant. However, we chose to retain the difference notation so that the deviation was explicit, and the reader did not have to remember that X_1 was the raw variable and x_1 was its centered counterpart.

For comparison, let's next consider the numerator of Eq. (4), which is the covariance of the product score and one of its components when the predictor variables have been mean centered. According to Eq. (5), the covariance between $\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)$ and $\left(X_1-\overline{X}_1\right)$ may be written as:

$$\begin{aligned} cov\Big(\Big(\Big(X_1-\overline{X}_1\Big)\Big(X_2-\overline{X}_2\Big)\Big),\Big(X_1-\overline{X}_1\Big)\Big) \\ = \varepsilon\Big(\Big(X_1-\overline{X}_1\Big)\Big)cov\Big(\Big(X_1-\overline{X}_1\Big),\Big(X_2-\overline{X}_2\Big)\Big) \\ + \varepsilon\Big(\Big(X_2-\overline{X}_2\Big)\Big)var\Big(\Big(X_1-\overline{X}_1\Big)\Big). \end{aligned} \quad (8)$$

At this point, the easy observation is made that the covariance in (the left side of) Eq. (8) is expected to be zero because the expected values of deviations for X_1 is zero, $\varepsilon\left(\left(X_1-\overline{X}_1\right)\right)=0$, as is that for X_2, $\varepsilon\left(\left(X_2-\overline{X}_2\right)\right)=0$ for normally distributed variables, (see Appendix 1 for the proof). Of course, the fact that the expected (or average) values of these mean centered variables are zero does not imply that in any given data set the covariance (and hence the correlation) will be precisely zero. However, on average, the covariances (and correlations) will be zero.

Thus, if the expected value of the covariance between $\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)$ and $\left(X_1-\overline{X}_1\right)$ is zero, then in all likelihood, the correlation in Eq. (4) among centered predictors will be reduced compared to the correlation in Eq. (3) among the original (uncentered) predictors. Therefore we may conclude that, yes, from this micro perspective, mean centering helps to reduce the micro form of multicollinearity.

It is also important to note that mean centering variables does not change the nature of the relationships between any variable in the set that does not include the product term. That is, any correlation among $\{X_1,X_2,Y\}$ will be the same as the correlation between the corresponding variables $\left\{\left(X_1-\overline{X}_1\right),\left(X_2-\overline{X}_2\right),Y\right\}$ (see Appendix 2 for the proof). As Dunlap and Kemery (1987, p. 420) say, "the intercorrelations among [the] original variables are unchanged, whereas the correlations involving cross-product terms are reduced dramatically." Specifically, $r_{X_1,X_2}=r_{\left(X_1-\overline{X}_1\right),\left(X_2-\overline{X}_2\right)}$, $r_{X_1,Y}=r_{\left(X_1-\overline{X}_1\right),Y}$, and $r_{X_2,Y}=r_{\left(X_2-\overline{X}_2\right),Y}$.

At this point, these relationships have been proven, but it might also help to see an illustration. Table 1 shows a random sample of 30 observations for three variables drawn from a multivariate normal population with correlation parameters $\rho_{X1,Y}=0.6$, $\rho_{X2,Y}=0.6$, and $\rho_{X1,X2}=0.3$ (though the relationships and equalities we are about to illustrate are replicable on any of the reader's own available datasets). Table 2 shows the correlations among these two predictors, X_1 and X_2, their product score, and the dependent variable. The correlation matrices are presented in both their original form and after the two main effect predictors had been mean centered (prior to the computation of the product score). Note that the equalities stated previously hold: $r_{X_1,X_2}=r_{\left(X_1-\overline{X}_1\right),\left(X_2-\overline{X}_2\right)}=0.278$, $r_{X_1,Y}=r_{\left(X_1-\overline{X}_1\right),Y}=0.587$, and $r_{X_2,Y}=r_{\left(X_2-\overline{X}_2\right),Y}=0.665$.

Table 1 Illustration data

Y	X1	X2	Y	X1	X2
3	3	4	3	3	4
4	5	3	4	4	4
4	4	2	4	5	4
4	4	4	2	3	3
5	4	6	3	5	4
3	1	3	2	4	3
2	4	1	4	3	4
3	5	3	3	3	2
5	4	5	4	4	3
4	5	3	1	2	1
5	6	4	3	3	4
3	2	3	5	5	6
2	3	3	3	2	3
4	4	3	4	5	3
4	5	3	2	2	3

Raw data, N = 30

In these results, we see that researchers are correct in asserting that micro multicollinearity is lessened by mean centering (e.g., Irwin & McClelland, 2001; Jaccard, Wan, & Turrisi, 1990; Smith & Sasaki, 1979). Specifically, correlations involving a product score based on centered variables like those in Eq. (4) in the right-hand side of Table 2 tend to be smaller than their uncentered counterparts like the correlations in Eq. (3) in the left-hand side of Table 2; we see that $r_{\left(\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)\right),\left(X_1-\overline{X}_1\right)} < r_{(X_1X_2),X1}$, as -0.167 < 0.759 (the magnitude is lessened, regardless of the sign change, and the correlation is not significantly different from zero). Similarly, $r_{\left(\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)\right),\left(X_2-\overline{X}_2\right)} < r_{(X_1X_2),X2}$, as 0.082 < 0.818.

Mean centering does not help alleviate "macro" multicollinearity

On the other side of the debate is the position that mean centering is not effective in reducing multicollinearity (cf., Dalal & Zickar, 2012; Echambadi & Hess, 2007). These arguments are based on the "macro" focus, looking at the determinant of the predictor matrix or global indices such as R^2. These researchers are not against mean centering, per se. For example, mean centering may provide a more parsimonious interpretation of data analysis, namely that the effects of one variable would be interpreted as a function of being above or below the mean on another variable (vs. above or below some arbitrary

Table 2 Comparative correlations

Correlations among original, raw variables					Correlations among centered variables				
	X_1	X_2	X_1X_2	Y		$(X_1-\overline{X}_1)$	$(X_2-\overline{X}_2)$	$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$	Y
X_1	1.000				$(X_1-\overline{X}_1)$	1.000			
X_2	0.278	1.000			$(X_2-\overline{X}_2)$	0.278	1.000		
X_1X_2	0.759	0.818	1.000		$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$	−0.167	0.082	1.000	
Y	0.587	0.665	0.765	1.000	Y	0.587	0.665	−0.091	1.000

intercept term). However, these researchers would argue that the decision to mean center is an altogether different issue, and one that does not alleviate multicollinearity (cf., Dunlap & Kemery, 1987; Echambadi & Hess, 2007).

A macro focus on multicollinearity considers all the interrelationships among the entire set of predictor variables, X_1, X_2, and X_1X_2. From this vantage, multicollinearity is not reduced because while mean centering reduces the off-diagonal elements (such as the covariance of X_1 with X_1X_2), it also reduces the elements on the main diagonal (such as X_1X_2 with itself, that is, its variance). That is, while the correlations might be reduced among X_1, X_2, and X_1X_2, the correlation between X_1X_2 and Y is also reduced. To better understand these relationships, consider the following demonstrations.

Table 3 presents the regression results for the data previously analyzed in Table 2. Note that the "micro" level results vary for X_1 and X_2 (but not for X_1X_2). These include the parameter estimates, raw regression coefficients b or standardized coefficients β, their standard errors, and of course, by implication, their t-statistics, p-values, and conclusions regarding which effects in the model are significant. (The b-weight and standard error for the highest order interaction term, X_1X_2, are not affected, although the corresponding β is modified.)

At the same time, note that the "macro" level results are constant, that the total amount of variance explained in Y is the same $R^2 = 0.622$, whether that variance is explained via $\{X_1, X_2,$ and $X_1X_2\}$ or $\{(X_1-\overline{X}_1), (X_2-\overline{X}_2),$ and $(X_1-\overline{X}_1)(X_2-\overline{X}_2)\}$. The total variance explained is the same, it is simply redistributed.[2] Recall from Table 2 that the correlations between the product scores and their components are lessened, but it is also true that the correlation between the product score and the dependent variable is lessened as well, $r_{((X_1-\overline{X}_1)(X_2-\overline{X}_2)),Y} < r_{(X_1X_2),Y}$, as $-0.091 < 0.765$. Thus we may conclude that indeed, from a macro perspective, researchers are correct to assert that mean centering does not reduce or affect the macro form of multicollinearity (cf., Dalal & Zickar, 2012; Echambadi and Hess (2007).

It can be stated that mean centering has no effect on this macro form of multicollinearity when one examines the determinant of the matrix. To understand the concept of a determinant better, consider Fig. 1. This figure plots the determinant of a 2 × 2 matrix in which a correlation between two variables varies from 0.0 to 1.0. The figure shows that the determinant of a correlation matrix ranges from 0 to 1. The determinant will equal one only if the correlation is zero, otherwise the determinant will be less than one. The determinant becomes zero or near zero when variables are very highly correlated, that is, in the presence of micro multicollinearity.

Multicollinearity is a problem because determinants that are zero or close to zero can cause computational difficulties. For example, the estimation of regression weights, $b=(X'X)^{-1}X'Y$ requires the inverse of $X'X$. The calculation of the inverse, in turn, requires the term $1/\det(X'X)$, and obviously the computation will blow up if the denominator is zero. Hence, the determinant of $X'X$ must not be zero.

To understand $X'X$, first think of one's data as a matrix X, in which respondents form the rows, and the columns comprise X_1, X_2, and X_1X_2 to be used to predict the dependent variable Y (Y is in a separate vector). The product of the matrix transpose (X') and itself, $X'X$, is referred to as the uncorrected sums of squares and cross products matrix, or the uncorrected SSCP matrix. When the matrix X is corrected for or centered around its means, we call the result a deviation matrix, X_d, and $X_d'X_d$ is the corrected SSCP matrix. The covariance matrix among the predictor variables, C, is equal to the corrected SSCP divided by (n-1). The diagonal of the covariance matrix contains the variables' variances. Next, these variances are extracted and put into a vector, the square root taken to obtain the standard deviations, and their reciprocals are taken, and the result is called $S^{-1/2}$. The familiar correlation matrix, R, is equal to the covariance matrix rescaled by the standard deviations, $R=S^{-1/2}CS^{-1/2}$.

Table 4 contains the raw uncorrected sums of squares and cross products matrices for the original variables from Table 2, X_1, X_2, and X_1X_2 at the left, and $(X_1-\overline{X}_1)$, $(X_2-\overline{X}_2)$, and $(X_1-\overline{X}_1)(X_2-\overline{X}_2)$ at the right. The elements in the two

[2] As a reviewer kindly pointed out, the betas do not capture variance per se; the squared part correlations, or effect-size indices like η^2, which are a function of the regression weights, capture variance more directly.

Table 3 Comparative regressions

Regression results for the original variables (R^2 = 0.622)						Regression results on the centered variables (R^2 = 0.622)					
	b-weight	Standard error	t-statistic	p-value	β-weight		b-weight	Standard error	t-statistic	p-value	β-weight
intercept	−0.440	1.500	−0.29	0.772		intercept	0.023	0.130	0.18	0.862	
X_1	0.576	0.392	1.47	0.154	0.667	$(X_1-\overline{X}_1)$	0.364	0.111	3.29	0.003	0.422
X_2	0.744	0.472	1.57	0.127	0.809	$(X_2-\overline{X}_2)$	0.509	0.116	4.37	0.0002	0.554
X_1X_2	−0.063	0.118	−0.53	0.600	−0.403	$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$	−0.063	0.118	−0.53	0.600	−0.066

matrices are quite different, yet the determinants of the full *X'X* matrices (from an X matrix that includes a column of ones for the intercept) are equal.

Similarly, Table 5 presents the covariance matrices. The pattern of identity—the set of elements that are constant in the left and right matrices—is the same as in the correlation matrices in Table 2 (e.g., cov(X_1,Y) = cov $\left((X_1-\overline{X}_1), Y\right)$ = 0.731). Note the variance of the product term is greatly reduced by mean centering, whereas the other variances remain constant. It is also of note that these determinants are identical to each other. Thus, the claim is also correct that the macro form of multicollinearity is not reduced as a function of mean centering.

Figure 1 illustrates how multicollinearity affects the determinant in a problematic manner, yet mean centering does not alter that functional form—the relationship holds whether the variables are mean centered or not. To show the effect of mean centering on other results of a regression, we ran a Monte Carlo study. We set $\rho_{X1,Y}=\rho_{X2,Y}=0.5$ to represent modest-sized effects of two predictors on the dependent variable, and varied the extent of multicollinearity, $\rho_{X1,X2}$ from 0.0 to 0.9. For each level of $\rho_{X1,X2}$, we generated a random sample of size N = 100, with three variables, X_1, X_2, and Y, with population means, μ_{X1} = 3 (to approximate a 5-point rating scale), μ_{X2} = 4 (to approximate a 7-point rating scale), and μ_Y = 0 (given that its value is immaterial to the effects of mean centering of predictor variables). We computed a product score, X_1X_2. We ran a regression using the main effects and interaction, X_1, X_2, and X_1X_2 to predict Y. We obtained the β estimates, their standard errors, and the p-values representing their significance tests. We also took the same generated variables and mean centered X_1 and X_2. We computed the new product score, ran another regression, obtained the new βs, standard errors, and p-values. We repeated this procedure for 1,000 replications.

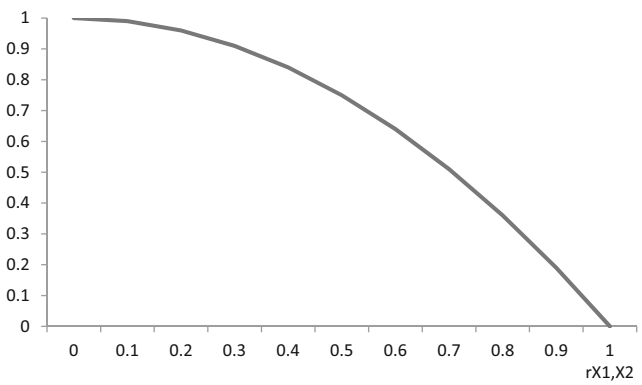


Fig. 1 Determinant as a function of multicollinearity $r_{X1,X2}$

Figure 2 shows the average over the 1,000 pairs of regressions for the β_1 (or β_2) estimates (given that the strength of the relationship between X_1 and Y was the same as that between X_2 and Y, both main effect βs were equal, for simplicity). As we mentioned previously, it is known that a linear transformation like mean centering will not affect the highest order term in a model, in this case, the X_1X_2 interaction term (Allison, 1977; Dalal & Zickar, 2012; Kromrey & Foster-Johnson, 1998; Shieh, 2011, 2010, 2009; Smith & Sasaki, 1979), thus we focus on the results for the main effect predictors. Much like the plot in Fig. 1 for the determinant, Fig. 2 shows that greater multicollinearity dampens the estimate of β. As was also true of Fig. 1, the plot in Fig. 2 is identical whether the predictor variable had been mean centered or not. That is, mean centering does not change the dampening effect of multicollinearity on estimates of regression coefficients.

By comparison, mean centering reduces standard errors and thus benefits p-values and the likelihood of finding β_1 or β_2 significant. Figure 3 shows that as multicollinearity increases, the standard error of the β estimate increases slightly, which in turn decreases the likelihood of finding predictor variables to be significant. That effect is exacerbated when taken together with the observation from Fig. 2 that multicollinearity lessens the β estimate, thus multicollinearity further decreases the likelihood that the β will be significant, and this effect is shown in the rising p-values in Fig. 4. Figures 3 and 4 depict the nature of how mean centering enhances a regression analysis—the effects of the correlations between predictors on the standard errors and p-values for β_1 or β_2 are greatly reduced. Thus, mean centering is beneficial in reducing effects of micro multicollinearity.

Note also in Figs. 3 and 4 that even the improvement offered by mean centering has its limits. When correlations among predictors approach or exceed 0.7, results begin to be affected even if the variables have been mean centered.

Table 4 Comparative SSCP (sums of squares and cross products) and determinants

X'X matrix, original variables (det = 1386656)				X'X matrix, mean centered predictors (det = 1386656)			
	X_1	X_2	X_1X_2		$(X_1-\overline{X}_1)$	$(X_2-\overline{X}_2)$	$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$
X_1	460			$(X_1-\overline{X}_1)$	41.867		
X_2	388	377		$(X_2-\overline{X}_2)$	10.933	36.967	
X_1X_2	1624	1484	6296	$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$	−6.302	2.916	37.908

Parameter estimates will be lower, standard errors higher, p-values higher, and significant findings fewer. When one or more bivariate correlations exceed 0.7, indicating 50 % or more redundancy between the variables, the researcher would be well served by creating a composite variable, based on factor analysis, so as to minimize any possible subsequent detrimental effects of multicollinearity.

Discussion

Multicollinearity arises in numerous contexts. In this paper, we considered whether mean centering variables help alleviate multicollinearity in multiple regression models. There are two perspectives regarding whether mean centering variables prior to the construction of their product term to serve as an interaction term helps to alleviate multicollinearity. In this paper, we reconciled the perspectives by distinguishing a *micro* analysis of correlations from a *macro* analysis of the determinants of matrices.

As we have shown, the debate and confusion about whether mean centering helps to alleviate multicollinearity is a function of discussing multicollinearity in slightly different contexts. In particular, there is a micro and macro view of multicollinearity and both camps are somewhat correct: mean centering reduces multicollinearity if that is meant to characterize individual correlation or regression coefficients—the micro form of multicollinearity, and yet mean centering does not reduce multicollinearity if that is meant to characterize the fit of the regression model as a whole, i.e., R^2, or the determinant of the covariance matrix—the macro form of multicollinearity. The micro focus reacts to high correlations, and the macro focus reacts to any linear combination of correlations that indicate that the predictors show some linear redundancies.

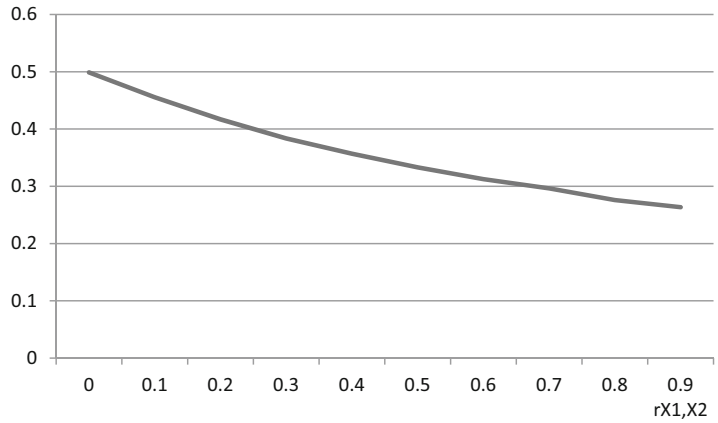
It is quite likely that researchers naturally gravitate toward a micro or macro perspective on multicollinearity as a function of their paradigmatic contexts and needs. For example, for many research questions in the social sciences, research questions revolve around testing the significance and contributions of particular independent variables, and their individual influences on dependent variables. From this perspective, multicollinearity puts the likelihood that any given predictor will be found significant at risk. Mean centering helps these researchers and is good practice when testing and reporting on effects of individual predictors.

For other research questions, a macro view might dominate, such as when a modeler wishes to improve a model's overall fit or predictions. These researchers may not care as much about the particular contributions of separate independent variables, instead caring about adding predictors so as to maximize macro results including the model's overall R^2. For example, increasingly, practitioners are analyzing real world, "big data" and are facing the macro issues that plague computer scientists and artificial intelligence researchers regarding identifying conditions under which estimation will be problematic, as estimates approach machine error boundaries. For these researchers, mean centering will not affect (harm or help) their analyses.

Finally, along the continuum between micro and macro may be placed the indices that many social scientists find

Table 5 Comparative covariances

Covariances among raw variables (det = 0.771)					Covariances among centered variables (det = 0.771)				
	X_1	X_2	X_1X_2	Y		$(X_1-\overline{X}_1)$	$(X_2-\overline{X}_2)$	$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$	Y
X_1	1.444				$(X_1-\overline{X}_1)$	1.444			
X_2	0.377	1.275			$(X_2-\overline{X}_2)$	0.377	1.275		
X_1X_2	6.051	6.129	44.064		$(X_1-\overline{X}_1)(X_2-\overline{X}_2)$	−0.217	0.101	1.170	
Y	0.731	0.779	5.269	1.076	Y	0.731	0.779	-0.102	1.076



*Note: the plot is the same for β_1 or β_2 computed on mean centered or raw variables.

Fig. 2 β_1 or β_2 as a function of multicollinearity $r_{X1,X2}$*

useful, namely variance inflation factors (VIFs). The VIF for a predictor variable X_i is $VIF_i = 1/(1-R_i^2)$, where R_i^2 is obtained from the multiple regression in which X_i is predicted from all the other independent variables. The square root of the VIF for X_i indicates how much larger the standard error of the regression coefficient for X_i is compared to what the standard error for X_i would have been if it were uncorrelated with the other independent variables. If there is no multicollinearity involving variable X_i, its VIF will be 1.0, that is to say the standard error of the regression coefficient has not been inflated, there is no biasing problem, and the likelihood that the predictor variable will be found significant is as good as it can be. In contrast, if X_i has a VIF of, say, 9, then the standard error for the regression coefficient representing variable X_i will have been inflated by a factor of $\sqrt{9} = 3$, and multicollinearity might therefore be causing problems in the data analyses. It has been suggested that a VIF for any given predictor variable should not exceed 10.0 (Marquardt 1970). Consider, for example, the uncentered data VIFs for X_1, X_2, and X_1X_2 are 14.175, 18.164, and 39.483, compared to the VIF indices for their centered counterparts, 1.128, 1.104, and 1.048, confirming from another perspective that mean centering helps reduce potentially bad effects of interrelated variables. While VIF indices take into account the effects of all the predictor variables and thus may be considered to be somewhat macro, their emphasis is on one predictor variable at a time, and they were also improved by the mean centering, so they might be considered somewhat more micro in character than macro.

Given the distinction between micro and macro perspectives, different researchers will use this information as they deem most appropriate for their scholarly interests. Researchers who focus on the macro level of analyses, seeking to use several predictor variables to enhance model fits can mean center their variables if they prefer to interpret their results relative to means rather than arbitrary intercepts, and they can proceed confidently knowing that doing so will not affect or improve multicollinearity at a macro level, including determinants and model fits such as R^2s. Alternatively, researchers who focus on the micro level of analyses, seeking to test hypotheses about the significance of the contributions of individual predictors should mean center their variables because micro multicollinearity will be reduced, and the patterns of significant predictors will be clarified.

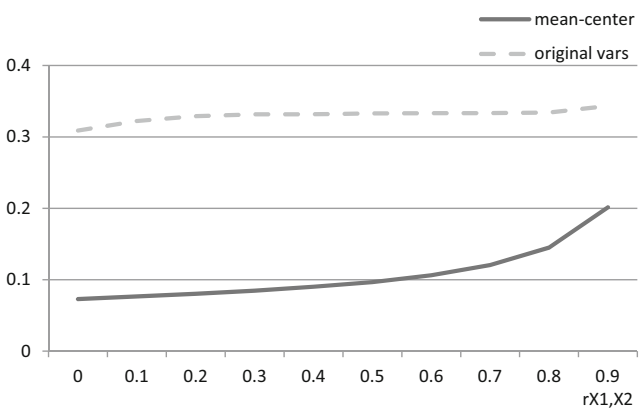


Fig. 3 Standard error of β_1 or β_2 as a function of multicollinearity $r_{X1,X2}$

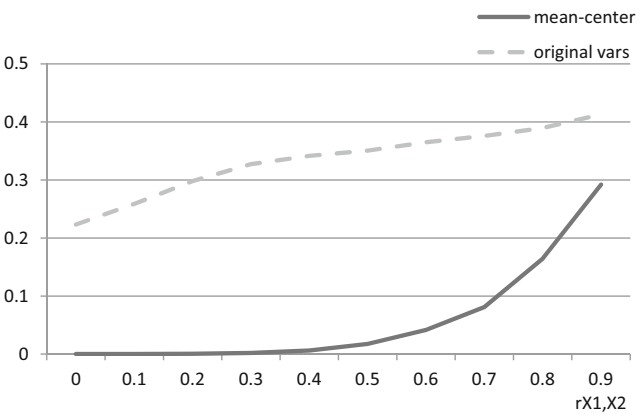


Fig. 4 P-value of β_1 or β_2 as a function of multicollinearity $r_{X1,X2}$

When all is said and done, should a researcher mean center the X_1 and X_2 variables before computing a product term X_1X_2 to include in a moderated multiple regression? It depends. Mean centering is advisable when: (1) the predictor variables are measured on scales with arbitrary zeros and the researcher seeks to enhance the interpretation of the regression results vis-à-vis the variables' means rather than the arbitrary zero points, or (2) the research questions involve testing the main effect terms in addition to the interaction term and the researcher seeks to obtain these statistical tests without the interference of the so-called nonessential multicollinearity. On the other hand, mean centering may be bypassed when: (1) the research question involves primarily the test of the interaction term, with no regard for the lower order main effect terms, or (2) the research question involves primarily the assessment of the overall fit of the model, the R^2, with no interest in apportioning the explained variability across the predictors, main effects or interaction alike.

Appendix 1

Recall, as claimed in the paper after Eq. (8), the expected value, written $\varepsilon(X_1)$ (or population mean, μ_1) of a deviation score such as $\left(X_{1i}-\overline{X}_1\right)$ is zero. Essentially,

$$\varepsilon\left(X_{1i}-\overline{X}_1\right) = \varepsilon X_{1i}-\varepsilon\left(\overline{X}_1\right) = \mu_1-\mu_1 = 0.$$

The relationship can be seen in the sample values as well:

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}\left(X_{1i}-\overline{X}_1\right) &= \\ &= \frac{1}{n-1}\sum X_{1i} - \frac{1}{n-1}n\,\overline{X}_1 \\ &= \frac{1}{n-1}\sum X_{1i} - \frac{1}{n-1}\sum X_{1i} \\ &= \frac{1}{n-1}\left(\sum X_{1i}-\sum X_{1i}\right) = 0\end{aligned}$$

Appendix 2

To prove the equality of the correlations with or without mean centering for terms that do not include the product term (as referenced after Eq. (8)), let us begin with the simple equation for the correlation between X_1 and X_2:

$$r_{X_1,X_2} = \frac{1}{n-1}\sum\left(\frac{X_{1i}-\overline{X}_1}{s_{X1}}\right)\left(\frac{X_{2i}-\overline{X}_2}{s_{X2}}\right) \quad \text{(B1)}$$

To show that the correlation between these two predictors is not affected by centering, we replace X_1 and X_2 with $\left(X_{1i}-\overline{X}_1\right)$ and $\left(X_{2i}-\overline{X}_2\right)$:

$$r_{\left(X_1-\overline{X}_1\right),\left(X_2-\overline{X}_2\right)} = \frac{1}{n-1}\sum\left(\frac{\left(X_{1i}-\overline{X}_1\right)-0}{s_{\left(X_1-\overline{X}_1\right)}}\right)\left(\frac{\left(X_{2i}-\overline{X}_2\right)-0}{s_{\left(X_2-\overline{X}_2\right)}}\right) \quad \text{(B2)}$$

The zeros in the numerators of Eq. (B2) are a result of the proof in Appendix 1. We can also show that the standard deviations are unaffected by centering, that is, $s_{X1} = s_{\left(X_1-\overline{X}_1\right)}$. The equation for the standard deviation of X_1 is:

$$s_{X1} = \sqrt{\frac{\sum\left(X_{1i}-\overline{X}_1\right)^2}{n-1}} \quad \text{(B3)}$$

The equation for the standard deviation of the centered version of the variable, $\left(X_{1i}-\overline{X}_1\right)$ is:

$$\begin{aligned}s_{\left(X_1-\overline{X}_1\right)} &= \sqrt{\frac{\sum\left(\left(X_{1i}-\overline{X}_1\right)-0\right)^2}{n-1}} \\ &= \sqrt{\frac{\sum\left(X_{1i}-\overline{X}_1\right)^2}{n-1}} = s_{X_1}\end{aligned} \quad \text{(B4)}$$

Thus, the correlations in Eqs. (B1) and (B2) are equal, $r_{X_1,X_2} = r_{\left(X_1-\overline{X}_1\right),\left(X_2-\overline{X}_2\right)}$. The same equality holds for $r_{X_1,Y} = r_{\left(X_1-\overline{X}_1\right),Y}$, and $r_{X_2,Y} = r_{\left(X_2-\overline{X}_2\right),Y}$.

In contrast to these relationships remaining unchanged with or without mean centering, the correlation between the product score and any of these non-product variables will be different depending on whether the product score was calculated on the original variables, X_1X_2, or their centered counterparts, $\left(X_1-\overline{X}_1\right)\left(X_2-\overline{X}_2\right)$.

References

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

Allison, P. D. (1977). Testing for interaction in multiple regression. *American Journal of Sociology, 83*(1), 144–153.

Arnold, H. J., & Evans, M. G. (1979). Testing multiplicative models does not require ratio scales. *Organizational Behavior and Human Performance, 24,* 41–59.

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association, 64*(328), 1439–1442.

Bradley, R. A., & Srivastava, S. S. (1979). Correlation in polynomial regression. *The American Statistician, 33*(1), 11–14.

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.

Dunlap, W. P., & Kemery, E. R. (1987). Failure to detect moderating effects: Is multicollinearity the problem? *Psychological Bulletin, 102*(3), 418–420.

Echambadi, R., & Hess, J. D. (2007). Mean centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science, 26*(3), 438–445.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26*(4), 797–833.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association, 55*(292), 708–713.

Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research, 38*(February), 100–109.

Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research, 25*(4), 467–478.

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement, 58*(1), 42–67.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics, 12*(3), 591–612.

Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association, 75*(369), 87–91.

Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables: Power and sample size considerations. *Organizational Research Methods, 12*(3), 510–528.

Shieh, G. (2010). On the misperception of multicollinearity in detection of moderating effects: Multicollinearity is not always detrimental. *Multivariate Behavioral Research, 45*(3), 483–507.

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

Smith, K. W., & Sasaki, M. S. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods & Research, 8*(1), 35–56.

Stone, E. F., & Hollenbeck, J. R. (1984). Some issues associated with the use of moderated regression. *Organizational Behavior and Human Performance, 34,* 195–213.