Comparing forms of centering on efficicency of product terms in regression

The following small simulation shows that all forms of centering (mean centering and residual centering) do not have an effect of the main target of interaction modeling—that is the B coefficient of the product term as well as its standard errors and t-value. The coefficients of the first-order effects, of course, change in respect to the centering procedure but that is not a consequence of bias. Instead, centering effects the meaning of the first order effects in presence of a product term, which is presented in detail here:

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

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

That is, the meaning of the B-coefficient of X (or Z) is the effect of X at zero-levels of Z (and vice versa). In the case of an uncentered pair of variables, "zero" in Z has a different meaning as in the centered case where zero means "average". For instance, take the example "work experience" (Z) as a moderator of the work stress (X) effect on tension (Y). In the uncentered case, the B coefficient of workstress is the effect of workstress for persons having no (zero) experience; whereas in the case of mean centered variables, the B cofficient is the effect of work stress for people with average levels of experience.

Accordingly, the standard errors change but again this is no sign of a bias. The paper by Echambadi and Hess explains that in detail.

This whole issue does not change with the form of centering as it was proposed here for the case of residual centering or orthogonalizing.

Lance, C. E. (1988). Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Applied Psychological Measurement*, 12, 163–175.

I am currently not sure of generalizations of the residual centering approach to latent interaction models may be a different issue, as proposed by

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*, 13(4), 497-519.

and myself either

Steinmetz, H., Davidov, E., & Schmidt, P. (2011). Three approaches to estimate latent interaction effects: Intention and perceived behavioral control in the theory of planned behavior. *Methodological Innovations Online*, 6(1), 95-110.

So this might be a nice future topic to investigate.

So, let's now come to the proof. I did a small and easy simulation in R. The code is below and can be run by anyone without further background knowledge. R uses so called packages that encompass functions to calculate or estimate something. In the presentation, I use the packages dplyr and broom that I need to create the product variables. These are automatically downloaded (with "install.packages(.)) and activitated (by library(.)).

The simulation works in three steps:

1. I simulate a simple regression model. In this model X and Z have first-order-effect of B=0.2, and the product of both is .80. The B coefficient of this product, its SE and the resulting t value is the key issue (as you want to find the interaction effect).
2. I created normally distributed data (N=300) that follows from this model
3. Then I conduct three regressions. The first uses the data as it is, the second calculates mean centered X and Z variables and uses these as ingredients of the product term, the third applies the residual centering approach by Lance (1988)

Note. All headers start with a # so that R won't stumble over them.

#Installation of the packages

```
install.packages("dyplr")
install.packages("broom")
library(dplyr)
library(broom)
```

#Step 1 & 2: Model and data

```
set.seed(123)   #For repeting the simulation with my results
X = rnorm(300)  #Creating X as a normally distributed variable
Z = .3*X + rnorm(300, 1.5, 1.2)  #The same for Z; both correlate slightly
Y = .2*X + .2*Z + .8*X*Z + rnorm(300, 1.2, .8)  #The interaction model
data = tibble(X,Z,Y)  #Creating the data set "data"
```

#1) Uncentered regression

```
LR_uncentered <- lm(Y ~ X*Z, data=data)
summary(LR_uncentered)

#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  1.26206    0.07797  16.187  < 2e-16 ***
#X            0.19548    0.08383   2.332   0.0204 *
#Z            0.17120    0.04054   4.223 3.22e-05 ***
#X:Z          0.78090    0.04464  17.492  < 2e-16 ***
```

#→ The coefficients come close to the simulated (population) coefficients. Note and #remember the Std.error and t value of the product "X:Z"

#2) Mean centered regression

#First, we center the data and build the product term

```
(data_cent <- data %>%
  mutate(X_cent = X - mean(X),
         Z_cent = Z - mean(Z),
```

```
         product = X_cent*Z_cent) )
```

Second, here's the regression

```
LR_meancentered <- lm(Y ~ X_cent + Z_cent + product, data=data_cent)
summary(LR_meancentered)

#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  1.57015    0.04869  32.249  < 2e-16 ***
#X_cent       1.38343    0.05153  26.847  < 2e-16 ***
#Z_cent       0.19810    0.04049   4.893 1.63e-06 ***
#product      0.78090    0.04464  17.492  < 2e-16 ***
```

→ As explained above, the first order effects change but not the product coefficient, its std.error and t-value!

#3) Residual centering

The last analysis does the residual centering approach (Lance, 1988). In this approach, the product term is created by multiplying the raw X and Z. Then (step 2), the product is regressed on the raw X and Z and the residuals are saved in the data. In the last step, The regression is done with the resiuals of step 2 as the product information.

#Step 1: Creating raw product

```
data_res <- data %>%
  mutate(product = X*Z)
```

#Step2 Creating residuals and saving them in the data

```
step1 <- lm(product ~ X + Z, data=data_res)

(data_res <-augment(step1, data_res))
```
#Adding residuals to the data

#Step 3: Doing the regression

```
LR_rescent <- lm(Y ~ X + Z + .resid, data = data_res)
summary(LR_rescent)

#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  1.35871    0.07777  17.471  < 2e-16 ***
#X            1.35248    0.05150  26.262  < 2e-16 ***
#Z            0.21082    0.04048   5.208 3.58e-07 ***
#.resid       0.78090    0.04464  17.492  < 2e-16 ***
```

#You see, no changes in the product term's B coefficient, std.error, and t value.

Conclusion: As you see, mean centering does not do anything…It only affects the interpretation of the first order effects. The theoretical question hence is: Is "zero" a meaningfull value in your theoretical model? In the case of using a dummy variable, like sex,

it surely is as zero would refer to the reference category (see Frazer et al., 2004). As noted above, the B for X is the effect of X when Z = 0. In the case of a dummy variable, hence it is the effect of X in the reference category.
Likewise, having experience would mean that B of X is the effect of X for people which are newbies and have no experience at all (if you want that!). However, often zeros are not meaningful, for instance in the case of age as a moderator, as the first order effect of X would be the effect of X for people whose age = 0. Statistically, but not informative.