NKZ Data Analysis: Multiple regressions

(i) Do gender, self-concept, value of education, teacher expectancies, and self-efficacy (all measured at time1) predict student rated effort (at time 2) controlling for effort at time 1?

(ii) Do students perceptions teacher interest, teacher caring, and teacher competency predict any additional variance above and beyond the students personal characteristics?

**Checklist**

In an effort to examine the data, we decided to clean the data and check for outliers. Thus, we decided to create Case ID numbers. All discussion of outliers will refer to the Case numbers created in this step because data will be continually sorted and re-ordered. This ID is the unique identifier for all future discussion.

**Univariate Outliers**

Upon creating case ID numbers, we computed Z-scores for each continuous variable to determine univariate outliers. Upon computing Z-scores, we reviewed each variable in ascending order to determine outliers. Since 3.29 is the absolute value of roughly three standard deviations, we were looking for scores above 3.29 or below - 3.29. That said, we did not simply delete those outside this range. We looked at the pattern of scores to see if the outlier made sense or seemed to be truly ‘out there.’

As we reviewed the Z-scores for Self-Concept we decided to delete Case number 28 because the score was -3.57, which is outside of the cut-off and rather far from the next lowest score. We deleted case number 222 because the distance between the scores (for Self-Concept) was too large. This was also the case when we decided to delete case number 250 (for Self-Concept).

We recalculated the numbers and decided not to delete any more case numbers but watch out for the following case numbers: #379 (under Effort 2) which had a z-score of -3.59, #407 (under Effort 1) which had a z-score of -3.75, #21 (under Teacher interest) which had a z-score of -3.38, and #150 (under Teacher caring) which had a z-score of -3.38 as well.

**Normality**

Next we checked for normality by running a Histogram for each continuous variable. We entered the variables in the same order (Self-Concept, Value of Education, Teacher Expectancies, Self-Efficacy, Effort 1, and Effort 2, Teacher Interest, Teacher Caring, and Teacher Competency). We decided that the normality was good. While the Normal Curve was not perfect for all variables, we felt they were ‘good’ enough. There were not any that looked highly skewed.

**Multivariate Outliers and Linearity**

We checked for linear relationships and multivariate outliers using scatter plots.We reviewed each continuous independent variable as it related to the dependent variable (Effort 2). We noticed that case #198 had a high effort but no value for education. We decided to keep it but watch it as we conducted other models.We also decided to watch case #409 because the Teacher Expectancy variable was slightly off but not too far off.We also decided to watch case #407 under Effort 1 because it is also off of the scatter plots. \*\*We noted that this is the second time that we are watching case #407.

We also noticed issues with case #21 with Teacher Interest. Again, this is the second time that case #21 is an outlier.We noticed that Teacher Caring is barely linear but not enough to delete any case numbers.We noticed that case #409 had a high caring and low effort outlier. This is the second time that #409 has been raised as an issue. We noticed that case #198 had a high Effort, low Caring and low Value for education. We thought that this was strange. Perhaps, it was a data entry error with “effort”.We noticed that case #21 under Teacher Competency had high Effort and low Competency. This is the third time that #21 has shown an odd pattern.

To further check for multivariate outliers and linear relationships, we decided to run a Mahalanobis Distance. This required entering Case ID as the dependent variable in a multiple regression with all the variables in our MR questions as independent variables. We looked at the degrees of freedom, 10 because of the 10 variables in our MR questions, on the chi-square table at .001 to determine that 29.5 was the cut off score. We sorted the variables in ascending order and noticed that the highest score was 37.257 which is over the cut score of 29.5. We reviewed each score above the cut off score and deleted them one by one, recalculating the model after each deletion. Eventually we deleted the following in order: case #253, case #21, case #415, and case #413 because all had scores that were above the cut score. Note that #21 was a ‘case to watch’ multiple times above, too.

**Correlations and Potential Multicollinearity**

The next step was to review the correlations matrix to make sure that all variables were correlated and check for any that seemed overly correlated, potentially indicating multicollinearity. We noticed that gender was not significantly correlated with other variables, but this is common with gender. Gender should be included, though, based on literature and accepted practice.

We noticed that Teacher Interest and Teacher Caring had a score of .775 which is very high and close to 1. This is not good because this means that they are overlapping and it could indicate multicollinearity. Other variables that showed possible multicollinearity issues were teacher interest and teacher competency (.801), teacher caring and teacher competency (.748), self-efficacy and self-concept (.591), teacher expectancies and self-concept (.604), and self-efficacy and value of education (.608).

*A full SPSS output with the* sequence of data cleaning, assumption testing and regression model is attached to the Wiki.

**Homoscedasticity**

As we reviewed the residual plot, we noticed two outliers. Case #409 is an outlier in the corner. Case #409 was a ‘case to watch’ multiple times above. Thus, we deleted that case and recalculated the regression and residual plot. We decided that case #407 has been an outlier multiple times and is still an outlier on the residual plot, so we decided to delete it. The assumption of Homoscedasticity was verified, the residuals are approximately equal for all predicted DV scores.

**Running the Regression**

Model 1 is significant, *R* = .56, *R2* = .31, adjusted *R2* = .31, *F*(1, 415) = 188.42, *p*<.05. This indicates that 31% of the variation in Effort 2 is explained by Effort 1. It would seem that Effort 1 and Effort 2 have a strong enough relationship to justify the use of effort 1 as a control. It is clearly a pre- and post- test situation, so this makes sense.

Model 2 is significant, *R* = .59, *R2* = .35, adjusted *R2* = .34, *F*(6, 410) = 36.27, *p*<.05. The change in R Square indicates 3.5% more of the variation in Effort 2 is explained when you add student characteristics (p<.05). Model 3 is significant,  *R* = .6, *R2* = .36, adjusted *R2* = .34, *F*(9, 407) = 24.89, *p*>.05. However the change in R Square, just under 1%, is not significant (p>.05). No additional variance can be explained with the addition of the teacher variables.

We decided that we could have a suppression effect in model 3, the variables are highly correlated and the constructs overlap in the literature. We believe that suppression may exist because the signs for Teacher Caring and Teacher Competency flip from zero-order to partial. We also noticed that there was a weird pattern with the Value of Education pattern, partial goes up in model 3 from model 2. So this variable explains more when the teacher variables are added to the model. That could be a mediation effect, we think. We noticed that Teacher Expectancy also has the weird pattern that Value of Education has, with partial correlation going up in model 3 from model 2. However, it is not a significant predictor in either model. We noticed that Value of Education is not significant in model 2 but it is significant in model 3. We think this could point to a mediation effect.

 **Questions**

**(a)** **Diagnostics**

* Is multicollinearity a problem? **Yes, we noticed that Teacher Interest and Teacher Caring had a correlation of .7.67 which is very high and close to 1. This is not good because this means that they are overlapping and it could indicate multicollinearity. Other variables that showed possible multicollinearity issues were teacher interest and teacher competency (.802), teacher caring and teacher competency (.745), self-efficacy and self-concept (.582), teacher expectancies and self-concept (.591), and self-efficacy and value of education (.595).**
* Examine the patterns in zero-order, part and partial correlations. Briefly comment on any information these values provide. **Multicollinearity and suppression/moderation detected due to the sign change between zero-order self-efficacy and value of education (. 595). The largest correlations existed between teacher interest and teacher competency at .802 and teacher caring and teacher competency (.745).**
* Are there any influential data points?**We definitely had influential points that changed the pattern of IVs. These were detected using the residual graph.** **Two outliers: Case #409 and #407 were deleted seperately, the residual graph was rerun after each deletion.**



RESIDUAL SCATTERPLOT A RESIDUAL SCATTERPLOT B (RERUN)



RESIDUAL SCATTERPLOT C (RERUN)

Concise responses: Output from the descriptive data and regression models

1. Considering only the research question and the descriptives (means/sd/simple correlations), identify at least 2 concerns that are evident before running the main analysis (MR).





Considering the research questions and the descriptives and simple correlation we were concerned with multicollinearity, and data entry errors (missing data) that influenced our mean and standard deviation. Specifically we worried that certain variables appeared too closely related. For multicollinearity, we focused specifically on the correlation of teacher competency & teacher teacher interest which showed a .802 relationship; that of teacher interest & teacher caring (.767); and that of teacher caring and teacher competency (.745). In the variable set for students, the variables of self concept and self efficacy correlated at .582; that of teacher expectancies and self-concept (.591); and that of self-efficacy and value of education (.595).

For data entry errors we focused on the mean and standard deviation of variables like self-efficacy with a mean of 2.82 and an SD of .82 and all three teacher variables with Means ranging from 2.5-2.9 with SD ranging from .71 - .92. These were resolved by running zscores and mahalanobis to compare thresholds and patterns between scores.

The research questions are also somewhat concerning. First, there are a lot of variables included which can limit the strength and accuracy of the model. The more variables you have, the more at risk you are for issues of multicollinearity, suppression, moderation, or mediation. While additional variables that can explain more variance are important, it might be better to start with fewer variables and see what they can do with the intention of watching how it changes with more variables. In a sense that’s what we did with the hierarchical approach, but we could have broken it down into more steps with fewer variables in each if the research questions allowed it. Second, the teacher variables are identified as a separate step, beyond the student characteristics, but teacher expectancies is in among the student characteristics. It should probably be moved.

2. If we dropped time 1 effort from the final model, how much would change would you expect see in R2 ?

Dropping time 1 effort would result in a change of 14%, the unique contribution that effort 1 offers the overall model.

3. Is there any evidence of suppression effects? Why or why not?

The possible existence of suppression/moderation was detected due to the change in sign between the partial and zero-order correlation of self-efficacy. There was a change in sign for teacher caring and teacher competency but these were not statistically significant. In addition, the residual scatterplot which showed a predictability in the higher range was better than in the low alerted our group to the possibility of a suppression/moderation effect.

Finally, because the Partial order correlation of teacher competency, teacher caring, teacher interest, and teacher expectancy were close to zero, we believed a mediation effect could be present.

Our group decided that multicollinearity caused variables to compete to explain variance, i. e., research indicates Self-efficacy should have been a stronger predictor of effort . The same is true of teacher caring. Teacher competency was thought to play an important part in effort but its effect on effort in the model may have been moderated by other variables (see above).

4. What do the diagnostic analyses you conducted indicate?









For relationship of IVs the diagnostic analysis (coefficient table) was used to find the best predictors and the variables with statistical significance. Model 1: is significant, *R* = .56, *R2* = .31, adjusted *R2* = .31, *F*(1, 415) = 188.42, *p*<.05. This indicates that 31% of the variation in Effort 2 is explained by Effort 1. Model 2 is significant, *R* = .59, *R2* = .35, adjusted *R2* = .34, *F*(6, 410) = 36.27, *p*<.05. The change in R Square indicates 3.5% more of the variation in Effort 2 is explained when you add student characteristics (p<.05). Model 3 continues to be significant,  *R* = .6, *R2* = .36, adjusted *R2* = .34, *F*(9, 407) = 24.89, *p*>.05. However the change in R Square, just under 1%, is not significant (p>.05).

A first run of the table was alarming as we had several variables that were not significant which our group felt (according to what we know of the constructs) were being affected by outliers which led us to cleaning the data further. After cleaning the data and running the final diagnostic analysis our group agreed that the coefficient table (below) displayed an accurate analysis of our variables: for this model the best predictor variables are self-efficacy and self-concept in both M2 and M3, when you assume that effort 1 is a control variable and not a predictor. Value of education was not significant in M1 but significant in M2, which may point to a mediating variable in the teacher variables. Teacher expectancies, however, remained not statistically significant in both models. Although this was the best data cleaning we could perform, our group felt that had we been able to delete variables we would have achieved a better variance effect, i.e., self-concept which we felt was competing with self-efficacy to explain variance.

During the diagnostic analysis, the deletion of outliers was made using the information from the residual scatter plot (see above step-by-step process), evaluation of z-scores, multicollinearity evidence in the correlation tables, and the presence of abnormal patterns in IVs (mahalanobis). Nine cases in all were deleted. A rerun of the coefficient and correlation tables at each deletion confirmed that deletions were necessary as the variance increased with each deletion.

5. If you were to run another regression with this data or collect data for a similar study in the future are there any changes you would make in your analysis plan (dropping IV’s, testing a different hypothesis, collecting additional data, etc.) Why or why not?

Limitations of this study included using IVs that were too highly correlated (see above) and/or not well designed to answer the first question. Our group felt that self-concept could have been omitted from the study as it competed with self-efficacy to explain variance. Without self-concept in the model we felt that the variance of the model would have increased and self-efficacy would have explained more variance. We also hypothesize that deleting teacher expectancy and/or teacher interest from the model would have added more variance from other variables. Teacher interest was at the heart of the two highest correlations that could be indicating multicollinearity (teacher interest & teacher competency showed a .802 relationship; teacher interest & teacher expectations showed a .767 relationship).

Based on our concerns above (question #1) regarding the research questions, specifically the high number of variables and the teacher expectancy variable in the first question instead of with the other teacher variables, we set out to revise the research questions for a second run at this data set. Taking into account the concerns with self-concept and the teacher variables, we removed self-concept and teacher interest. We felt that teacher caring and teacher interest were closely related in construct, and with teacher interest so highly correlated with other variables, we could justify leaving out interest while keeping caring. We moved teacher expectancy to the third model so it would be with the other teacher characteristics.

Our new questions are:

(i) Do gender, value of education, and self-efficacy (all measured at time 1) predict student rated effort (at time 2) when controlling for effort at time 1?

(ii) Do students perceptions of teacher expectancies, teacher caring, and teacher competency predict any additional variance above and beyond the students’ personal characteristics?