**Reflections on Y750 Semester Project (Shimon Sarraf)**

This semester project was a very interesting and rewarding educational experience for me. Not only did I learn about an advanced statistical model, latent growth curve model, but I was able to apply it to research questions that many in higher education assessment are interested in, namely have survey response rates declined, if so by how much, and what sorts of things can colleges and universities do to possibly help improve their response rates.

For this reflection, I would like to focus on the problems I encountered while completing this project and how I addressed them. The first major issue worth noting is the confusion I experienced when estimating my final model using FIML to address missing values for both continuous and categorical data. In my earlier models all cases were being used (over 1,000) but as soon as I entered in my covariates the count dropped to about 60. After a bit of reflection and recollecting something I had read and/or been told by Prof. Rutkowski, I realized that FIML cannot address categorical missing data. I had a choice to either forget about the categorical missing data relating to schools using their learning management systems and survey incentives (very important controls in a response rate model, at least in theory) or to dive into multiple imputation with R’s *Amelia* package. The lavaan google group was a great source of information for multiple imputation in lavaan; Prof. Terence Jorgensen was extremely helpful with answering different related questions. The decision to ultimately use multiple imputation was the right one I believe, but it led me to the door of my next major issue—estimating model fit statistics with imputed data.

Estimating model fit statistics with multiple imputed data sets was challenging at first. I really struggled because I kept bumping into error messages in R. I spoke with Prof. Rutkowski and she provided very helpful advice. She advised me to get parameter estimates using all multiple imputed datasets together, but to use individual imputed data sets to run model fit statistics and then average the results across the data sets. I really consumed a lot of time trying to figure out on my own how to get model fit statistics based on all imputed data sets at once, so it was a relief to hear some “outside the box” ideas that work. Having an experienced researcher to turn to close by is a luxury, and I realize I will need to think outside the box a bit more in the future and not feel so compelled to find and follow some prescribed set of rules (though sometimes this may not be possible).

Although a smaller hurdle to jump over, figuring out whether or not to transform my variables to address skewness/kurtosis as well as their varying scale took a lot of thinking. I had heard from Prof. Rutkowski and read on the Mplus website that using items with very different scales can cause convergence issues so I decided to standardize my variables from the start. This did not solve my non-normal distributed data issue, but using an appropriate estimator (WLSMV) I expected this issue would not bias my results. I thought to use the MLR estimator to address kurtosis, but lavaan’s growth command said in an error message I could only use WLSMV.