**André Vargas Gálvez (s4600479)**

**Take-home Exam**

due in class May 30, 2016 (13:45 pm)

This exam involves the use of structural equation models (SEM) to examine part of a theoretical model involving choice of secondary school. Specifically, the exam requires (a) the inspection of the data, (b) the proper specification of SEMs to test a set of hypotheses, and (c) a description of the analytic procedure and results in APA style. The exam consists of 30 items. Provide a printed copy of the answers and upload this document on Blackboard by the due date and time. The background for the data and hypotheses are provided below.

**Background**

There has been special interest in the transition process from the elementary school to the secondary school in The Netherlands. The outcome of this process determines one’s chances in life, job position, income, socio-economic status, etc. Hence, this is a very important choice made at the end of elementary school, approximately at the age of 12. The choice is basically between vocational education and pre-tertiary education, however, with six difficulty levels. In The Netherlands the choice is between VMBO BBL, VMBO GL, VMBO KBL, VMBO TL, HAVO, and VWO. These choices are printed in order of increasing difficulty, so that VWO is the highest and most difficult secondary education one can follow. Because this choice is such an important decision early in life, it is only normal that researchers are interested in the reasons for specific decisions.

Figure 1 presents the theoretical model regarding choice of secondary school. This theory posits that the final choice, which is decided upon by a committee, is predicted by the teacher recommendation (Recommen), the child’s preference (Preferen), and the child’s standardized academic scores (CITO; CitoTest). This makes sense, in that, these three measures are the primary criteria used by the committee when making the final decision. These three measures are all predicted by the achievement of the child (Achieve), and the child’s achievement is predicted by quality of the school environment (Quality) and quality of the home environment (SES).

**Figure 1: Conceptual model of the choice of secondary school**

****

**The rationale for the present study (exam)**

The present study examines the associations between school quality, home environment, student achievement, and CITO test scores. There is some debate regarding the most appropriate measures of quality of the school and home environments. Some scholars argue for specific measures, whereas others argue for more sophisticated measures of these multidimensional constructs. The present study is an initial attempt at creating latent constructs of school quality and home quality, and examining whether these latent constructs predict an observed measure of student achievement. Specifically, school quality describes three indicators: an objective measure of overall rating of the school (*school*), an evaluation of teachers qualifications and ability (*teacher*), and sociometric nominations gathered from classmates (*accept*). The home environment includes two measures of socioeconomic status (*educate* and *income* of parents) and a measure of parental involvement in educational activities (*involve*). Achievement describes teacher reports of child academic and social-emotional development (*achieve*). In addition, this study examines the link between achievement and CITO test scores.

Three research questions are addressed. First, do the selected indicators of school and home quality provide reliable latent constructs? The hypothesis is that the measurement model for the latent constructs (school and home quality) will demonstrate an adequate fit to the overall data. Second, do school and home quality each uniquely predict student achievement? It is expected that both latent constructs will be positively associated with achievement. Third, does achievement predict scores on the CITO test? The hypothesis is that achievement will be positively associated with CITO test scores.

**Method**

*Participants*

The sample included 800 children attending 50 elementary school classrooms (50% female, *M*age = 10.2 years) in the Netherlands. Almost all participants were born in The Netherlands and had two biological parents from The Netherlands (92%). Thirty schools were initially contacted and invited to participate in the study. Twenty-five schools agreed to participate, with the primary reason for non-participation being excessive workload reported by teachers. For those that agreed, students were approached in the classroom, parents and teachers are approached via email and/or regular mail.

*Measures (in the order they appear in the data file)*

*Sex*: coded (female = 1, male = 2)

*Educate*: Parent reports of highest education level obtained by child’s mother and father

*Income*: Parent reports of total household income.

*Involve*: Parent reports of involvement in educational and school-related activities.

*School*: Objective rating of overall school quality

*Teach*: Evaluation of teacher in terms of qualifications and ability

*Accept*: Number of peer nominations of most liked minus most disliked (social preference).

*Achieve*: Child’s score on teacher report of academic and socio-emotional competence.

*CITO*: Child’s score on the standardized academic test

All measures are continuous (except sex), and all have been converted to reflect a common scale, with negative numbers reflecting low values (e.g, lower educated parents) and positive numbers reflecting high values. Please note: these are not standardized or centered scores. These data are available on Blackboard in a file called takehome2016.csv (which includes the variable names). Use this data file and the information provided above to answer the following questions.

**PART A: Remember the three steps?**

Inspect the variables in terms of univariate distributions, outliers, and missing values and answer the following questions. Each item is worth one point.

1. **Are any of the variables skewed or kurtosed? If so, which variable(s)?**

Only one variable is skewed (teach).

1. **Do any of the variables contain any outliers (> 3 *SD*)? If so, which variables and how many?**

The variables educate has 2, school has 2, teach has 1, accept has 2, and achieve has 3

1. **BRIEFLY describe the patterns and prevalence of missing values.**

The variable with most missing values is teach, followed by involve, income and achieve.

The most repeated pattern is the one containing all of the values.. The second pattern is observations without teach variable and then without educate and the third without teach and CITO.

Males from the sample where missing the teach variable followed by achieve and involve. Females presented more missing values in teach, followed by involve and income.

Males in comparison to females missed more values from the school, income and teach variable, but females missed more values on achieve variable than males.

1. **Are the missing values Missing Completely at Random (MCAR)? Report the test statistic and interpret the statistical significance of this statistic.**

By doing the Little MCAR test we found a p-value < .001, which means that there is a probability of the missing values not being missed completely at random (there might be pattern)

1. **BRIEFLY describe the correlations (based on listwise deletion) in APA style.**

Parent reports of highest education level (educate) and income reports where positively correlated, *r* = .3, *p* < .001, this means that if a kid’s parents have high education levels the person will likely also have a high levels of income. Involvement in educational and school-related activities (involve) is positively correlated with education, *r*= .3 *p* < .001, and income levels, *r* = .8, *p* < .001. This means that if parents are more involved then is more likely to have higher levels of education. The same thing should apply for income levels. Parents education levels is positively correlated with the quality of the school *r* = .13, *p* < .001, and the evaluation of teachers, *r* = .13, *p* < .001. This means that more educated parents would likely have children in schools with higher level and “better” teachers. School quality is positively correlated with parents’ income *(r* = .59, *p* <.001) and involvement, *r* = .6, *p* <.001. This means the better the school quality children are likely to have parents with more involvement and more income.

Evaluation of teachers (teach) is positively correlated with parents income (*r* = .61, *p* <.001), involvement (*r* = .62, *p* <.001), and school quality ( *r*= .82, *p* <.001). This means that teachers with better evaluations would likely come from schools with high quality, where parents have high income and more involvement with children. Children’s social acceptance (accept) is positively correlated with parents income (*r* = .48, *p* <.001), involvement (*r* = .47, *p* <.001), school quality (*r* = .48, *p* <.001) and teachers’ evaluations (*r* = .5, *p* <.001), but not correlated at all with parents education level. This means that children who are more socially accepted are likely to have parents with higher income and involvement. In addition kids that are more socially accepted attend schools with better quality and highly evaluated teachers.

Child’s score on teacher report of academic and socio-emotional competence (achieve) is positively correlated with parents education (*r* = .25, *p* <.001), income (*r* = .71, *p* <.001), and involvement (*r* = .70, *p* <.001), with school quality (*r* = .53, *p* <.001), teachers’ evaluation (*r* = .57, *p* <.001) and more acceptance (*r* = .41, *p* <.001). Finally the CITO score is positively correlated with parents education (*r* = .13, *p* <.001), parents income (*r* = .59, *p* <.001), and involvement (*r* = .59, *p* <.001), school quality (*r* = .82, *p* <.001), teachers evaluation (*r* = .78, *p* <.001), children’s acceptance (*r* = .46, *p* <.001) and achievement (*r* = .53, *p* <.001).

1. **Does the pattern of correlations calculated using listwise deletion differ from those calculated using pairwise deletion?**

Yes they differ because with the pairwise deletion some correlations seem a little bit higher.

This might be the case because there might be a systematic patterns of missingness,and this method leads to biased results when this happens.

1. **Based on these preliminary analyses, are there any reasons to NOT use “normal” maximum likelihood estimation (MLE)? If so, state the reasons why MLE is not appropriate and describe potential alternatives. If you think MLE is appropriate, then state the reasons why MLE is appropriate.**

I think there are very few reasons to not use MLE method, because given that we have missing values MLE is strong (unbiased and uses accurate power) against MAR patterns of missing values which might be the case here. Moreover, we only have one non-normal variable which is kurtosed, but not very strongly, so I think this is not a big issue that if it becomes one it can be resolved with a variable transformation or with bootstrapping.

**PART B: SEM estimation, evaluation, and improvement**

Answer the following items based on the information describing the rationale of the study. Each item is worth one point.

1. **Express the portion of the theoretical model examined in this study as a set of four equation(s). That is, describe two equations that specify the measurement model (creation of latent constructs of school quality and home quality); and describe two equations that specify the structural model (the latent constructs predicting achievement, and achievement predicting CITO test scores).**

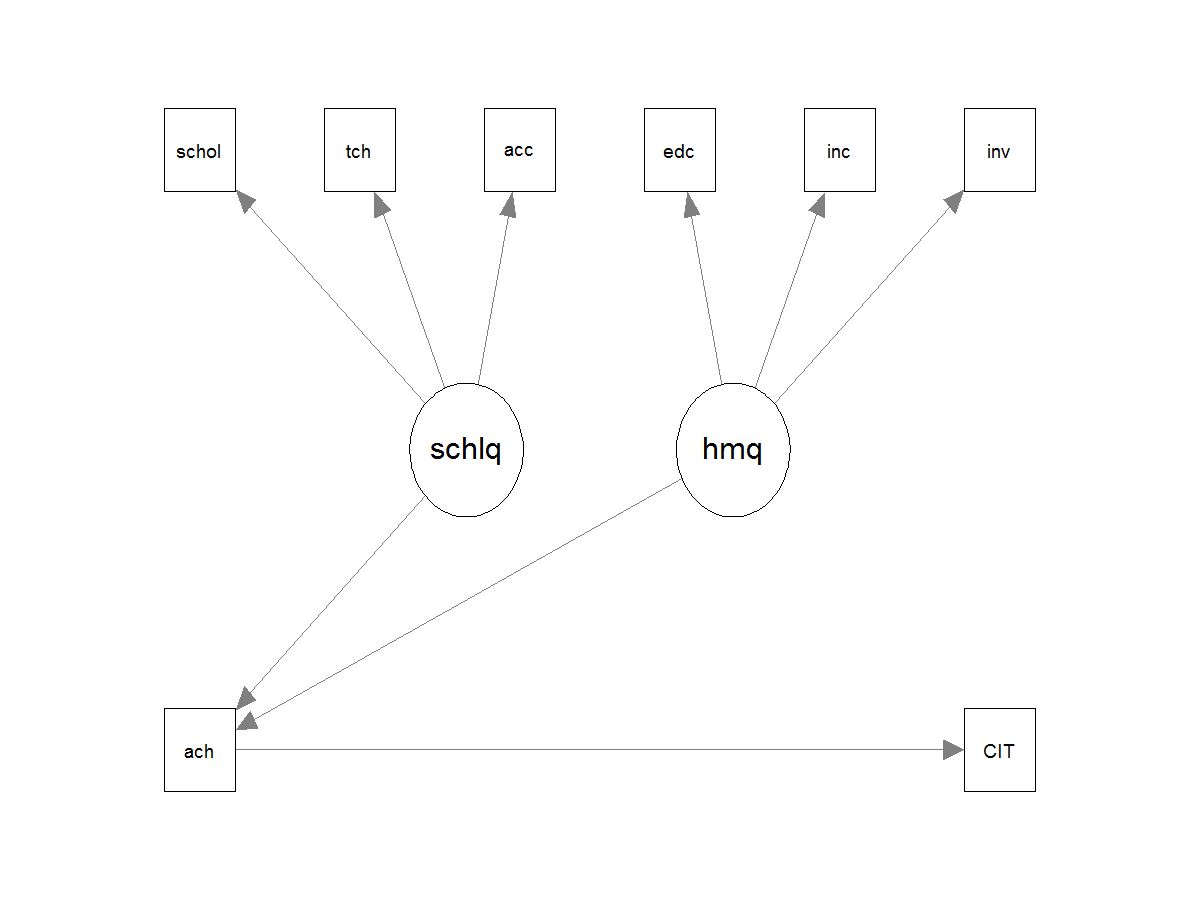
EQ1: schoolq =~ school + teacher + accept

EQ2: homeq =~ educate + income + involve

EQ3: achieve ~ schoolq + homeq

EQ4: CITO ~ achieve

1. **Create a path diagram that depicts the theoretical model using semPlot().**

****

1. **Provide a R script that specifies, fits, and summarizes the measurement model using the *lavaan()* function. This script should include the correct specification of the latent constructs, the intercepts and variances of the manifest indicators, the variances of the latent constructs, and the covariance between the latent constructs. Use “normal” MLE to estimate the model and full-information maximum likelihood to account for missing data. Hint: this model should include two of the equations described in #8.**

modeleq <- "schoolq =~ 1\*school + teach + accept

homeq =~ 1\*educate + income + involve

homeq~~homeq

schoolq~~schoolq

schoolq~~homeq

educate~~educate

income~~income

involve~~involve

school~~school

teach~~teach

accept~~accept

accept~1

involve~1

school~1

teach~1

eductae~1

income~1

involve~1"

fit <- lavaan(modeleq, data = takeh, missing="FIML", estimator = "ML")

1. **Did the measurement portion of the model adequately fit the observed data? If so, briefly describe what criteria were used to support this decision? If not, which modifications to the model needed to be estimated to adequately fit the model to the data?**

The model presents (under ML) a p-value < .001, which indicates that the model is not a good fit, but because chi square can be biased by the n sample size (bigger sample sizes increase the chance of inflating estimates and therefore p values), we examined other goodness of fit described above.

The CFI = .980, and TLI = .963, which indicate a “good” model fit. Furthermore, the RMSEA p-value < .001, and the SRMR = .038, indicating the model does fit the data.

Looking at the mi and epc:

educate should be added as an independent variable to explain school quality, it has a high mi, but the epc is negative.

Accept should be added as an independent variable to explain home quality.

The covariance between school and teach should also be added .

The covariance between accept and educate should also de added.

The covariance between income and involve also should be added.

1. **Estimate the model using a robust estimator. Describe which robust estimator is most appropriate in this situation and describe whether the use of the robust estimator alters the pattern of results.**

ML robust estimator is good with missing values or incomplete data, and doesn’t deal with least squares estimation problems. Regarding the model fit parameters the chi square p value stayed the same, the CFI = .979, and TLI = .961 decreased. Furthermore, the RMSEA p-value = .001 and the SRMR stayed the same indicating that the model does fit the data.

1. **Estimate the initial model (i.e., the model using “normal” MLE) using listwise deletion. Does the exclusion of those with missing values alter the pattern of results?**

It doesn’t change the estimates their significance. Regarding the model fit parameters the chi square p value < .001, the CFI = .980, and TLI = .963 increased compared to the last model. Furthermore, the RMSEA p value = .001, and the SRMR = .038, indicating a “good” model fit.

1. **Provide a R script that specifies, fits, and summarizes the measurement AND structural portions of the model using the *lavaan()* function. Use “normal” MLE to estimate the model and full-information maximum likelihood to account for missing data. Hint: this model should include all four equations described in #8.**

modelmax <- "schoolq =~ 1\*school + teach + accept

homeq =~ 1\*educate + income + involve

homeq~~homeq

achieve ~ schoolq + homeq

CITO ~ achieve

schoolq~~schoolq

schoolq~~homeq

educate~~educate

income~~income

involve~~involve

school~~school

teach~~teach

accept~~accept

accept~1

involve~1

school~1

teach~1

educate~1

income~1

CITO~1

achieve~1

achieve~~achieve

CITO~~CITO"

fitm <- lavaan(modelmax, data = takeh, missing="FIML", estimator= "ML")

1. **Did the SEM adequately fit the observed data? Briefly describe what criteria were used to support this decision.**

The model presents (under ML) a p-value < .001, which indicates that the model is not a good fit, but because chi square can be biased by the n sample size (bigger sample sizes increase the chance of inflating estimates and therefore p values), we examined other goodness of fit described above. The CFI = .829, and TLI = .520, which indicates a mixed result. Furthermore, the RMSEA p-value < .001, and the SRMR = .119, indicating a mixed result, but taking all in account we assume the model does not fit the data.

1. **Based on the modification indices, what adjustments improve the model the most? That is, which parameters have the highest expected parameter change? Does it make “theoretical” sense to include any of these parameters?**

Based on the mi:

Achieve should be added as an independent variable to school quality

Achievement as a predictor of home quality.

Also CITO as the predicto of school quality.

The 3 highest epc are:

The home quality as a predictor of school quality.

Achievement as a predictor of school quality.

And adding achievement as an independent variable to home quality.

1. **Based on your understanding of the hypothesized model and the associations among study measures, what are the two adjustments that make the most theoretical sense to include in order to improve the fit of the model?**

School quality as a predictor of CITO might be a good idea because kids with better scores should be on schools with better quality, or better quality schools could have better CITO scores in general.

Also home quality as a predictor of CITO might be a good idea because the quality of the parenting style might also influence kids grades and how he is developing.

Both relations (CITO, school) (CITO, school) both have positive correlations, including a relative high mi and epc.

1. **Provide a R script that specifies, fits, and summarizes the “improved” model using the *lavaan()* function. This SEM should include the two additional parameters described in #17.**

modelimp <- "schoolq =~ 1\*school + teach + accept

homeq =~ 1\*educate + income + involve

achieve ~ schoolq + homeq

CITO ~ achieve + schoolq +homeq

homeq~~homeq

schoolq~~schoolq

educate~~educate

income~~income

involve~~involve

school~~school

teach~~teach

accept~~accept

accept~1

involve~1

school~1

teach~1

educate~1

income~1

CITO~1

achieve~1

CITO~~CITO

achieve~~achieve

homeq~~schoolq"

fitimp <- lavaan(modelimp, data = takeh, missing="FIML", estimator= "ML")

1. **Did the “improved” model adequately fit the observed data? Briefly describe what criteria were used to support this decision.**

The model presents (under ML and FIML) a p-value < .001, which indicates that the model is not a good fit, but because chi square can be biased by the n sample size (bigger sample sizes increase the chance of inflating estimates and therefore p values), we examined other goodness of fit described above. The CFI = .984, and TLI = .972, which indicates a positive result. Furthermore, the RMSEA p-value = .004, and the SRMR = .035, indicating a positive result, so we assume the model does fit the data.

1. **Are there any reasons to question the results and conclusions drawn from the final model? If so, briefly describe your concerns.**

I don’t believe there is any true concern about this model so far, maybe it could be “upgraded” looking at the mi or epc even further, or adding more theoretical explanations which may turn into more covariances between variables, but so far the model shows a good model fit overall.

**PART C: Write it up!**

Report the results of the model estimation procedure in APA style (i.e., similar to the results section of a journal article). You may include a figure or table to describe the results, or you should simply report all relevant values in the text. Each of the items are worth 2 points each. Be sure to include a description of:

1. **The hypothesized model, the estimation procedure and statistical software.**

The model presented, tries to use several scores (variables) to explain the transition process from the elementary school to the secondary school in the Netherlands. The hypothesized model was created with a structural equation modeling approach (SEMs), to test for the effects among some constructs. The SEM was estimated using R (R Core Team, 2015) and the package lavaan (Rosseel, 2015). First the model was used to estimate a measurement model, which then was used to estimate a structural model, that lead to a modified version of the whole model.

The model consists of four equations, which are divided into two sections (measurement model, and structural model). The measurement model has the aim to generate latent constructs using independent variables to explain them. Specifically, school quality (schoolq) was calculated with an objective measure of overall rating of the school (school), an evaluation of teachers qualifications and ability (teacher), and sociometric nominations gathered from classmates (accept). In addition, home quality (homeq) was measures using the socioeconomic status (education and income of parents) and a measure of parental involvement in educational activities (involve).

The second part of the model or structural part of the model consisted of two more equations. The first equation used Achievement as dependent variable which measures the child’s score on teacher report of academic and socio-emotional competence, as a function of the two latent constructs home quality and school quality. Finally, the last equation was made of CITO score (child’s score on the standardized academic test) as a dependent variable as a function of the achievement variable.

1. **The procedure used to inspect the prevalence, pattern, and mechanism of missing values**

The data was analyzed beforehand in order to check for missing data using the VIM and mvnle packages (Gross & with help from Douglas Bates, 2012; Templ, Alfons, Kowarik, & Prantner, 2015). Rubin (1976) proposed a classification scheme for missing data mechanisms that was used to classify the missing data as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

First a general inspection of the whole data set was made. It was found that the variable with most missing values is teach, followed by involve, income and achieve. The most repeated pattern is the one containing all of the values. The second most prevalent pattern is based on observations that lack the variable teach, followed by observations without educate, and the third most prevalent pattern were observations without, both teach and CITO variables.

Furthermore, a second analysis was made, by subseting the sample with, generating one sample for males and one sample for females, in order to check for differences or noticeable patterns depending on the gender. In the sample with only males, the variables that were missing the most were teach followed by achievement and involve. In the case of females, the patterns presented more missing values in the variables teach, followed by involve and income.

The most relevant patterns where that males in comparison to females, missed more values from the variable school, income and teach, but females missed more values on the achievement variable than males. Finally in order to classify the pattern of missing values a Little MCAR test was made with the BaylorEdPsych package (Beaujean, 2012). It was found that there is a probability of the missing values not being missed completely at random (Little's MCAR test: χ2 (300) = 485.46, *p* < .001), suggesting there might be pattern similar to MAR.

1. **How missing values were addressed (i.e., imputation strategy).**

In order to deal with the possible pattern of missing value in our data, three different estimations were made. The first was using a maximum likelihood estimation using a full information maximum likelihood imputation (FIML) strategy(Enders & Bandalos, 2001). The second estimation used a FIML imputation strategy and a robust estimator, namely the maximum likelihood robust estimator (Yuan-Bentler Scaled χ2 and SEs), which deals with missing values or incomplete data in an efficient way without the least squares estimation problems on missing data. Finally, the model was estimated using the maximum likelihood estimation and listwise deletion strategy, which may lead to biased parameter estimates under MAR (McDonald & Ho, 2002).

1. **The overall fit of the hypothesized model to the observed data (use multiple criteria).**

The hypothesized model that was estimated with the maximum likelihood estimation and FIML imputation strategy, presented a χ2 (18) = 797.66, *p* < .001, indicating that the model is not a good fit. Nonetheless, chi square test are sensitive to large sample sizes and non-normality (bigger sample sizes increase the chance of inflating estimates and therefore p values), thus we examined other goodness of fit described above. Bentler’s comparative fit index (CFI) was examined (CFI = .829), and Tucker-Lewis index (TLI = .520) too. Using a cut-off value bigger or equal than .95, both indicate that he model is not a good overall fit. Furthermore, the Root Mean Square Error of Approximation (RMSEA) was analyzed, and the Standardized Root Mean Square Residual (SRMR) too. Both RMSEA (RMSEA = 0.233, *p* < .001) and SRMR (SRMR = .119), indicating a mixed result, but taking all the other criteria in account we assume the model does not fit the data.

1. **The modifications made to improve the fit of the hypothesized model**

In order to present a model that fits the observed data the model was modified and improved using the most common approach. This approach uses the Modification Index (MI) and the expected parameter change statistic (EPC) (Kaplan & Kaplan, 1989). Furthermore once the best fitting combinations were observed, the theoretical implications of adding new variables to the model were taken into account as to make the model theoretically sound.

Based on the MI, the three combinations of variables that had the highest score where: Achievement as an independent variable to school quality (MI = 1384.3), achievement as a predictor of home quality (MI = 1069.86), and finally CITO as the predictor of school quality (MI = 285.54). Based on the expected parameter change the highest scores where: home quality as a predictor of school quality (EPC = 2.316), achievement as a predictor of school quality (EPC = .962), and adding achievement as an independent variable to home quality (EPC = 2.92).

Other relevant variables with relative high MI and EPC scores where: School quality as a predictor of CITO (MI = 247.31, EPC = .521), and home quality as a predictor of CITO (MI = 88.42, EPC = 1.35). These two variables could be added because we can hypothesize that school quality as a predictor of CITO might explain that kids with better scores should be on schools with better quality, or better quality schools could have better CITO scores in general. In addition, home quality as a predictor of CITO might explain that the quality of the parenting style may also influence kids’ grades and how they develop.

1. **The amount of variance explained by the predictors in the two regression equations.**

The amounts of variance explained by the independent variables of the two regression equations are the following. For achievement as a predictor of CITO score: Estimate = .615, SE = .038, *p* < .001, for school quality as one of the predictor of achievement Estimate = 2.62, SE = .16, *p* < .001, and for and home quality as the other predictor of achievement Estimate = .14, SE = .026, *p* < .001.

1. **The overall fit of the final (improved) model to the observed data (use multiple criteria).**

The “improved” model was estimated with the maximum likelihood estimation and FIML imputation strategy, presented a χ2 (16) = 86.75, *p* < .001, indicating that the model is not a good fit. Nonetheless, chi square test are sensitive to large sample sizes and non-normality (bigger sample sizes increase the chance of inflating estimates and therefore p values), thus we examined other goodness of fit. Bentler’s comparative fit index (CFI) was examined (CFI = .984), and Tucker-Lewis index (TLI = .972) too. Using a cut-off value bigger or equal than .95, both indicate that he model is a good overall fit. Furthermore, the Root Mean Square Error of Approximation was analyzed, and the Standardized Root Mean Square Residual too. Both RMSEA (RMSEA = .074, *p* = .004) and SRMR (SRMR = .035), indicating a positive result. Taking all the examined criteria in account we assume the model does fit the data.

1. **The relevant statistics describing all parameter estimates (i.e., factor loadings and regression paths) in the final (improved) model.**

The measurement model, as described above, included the estimation of two constructs school quality and home quality. For school quality, the evaluation of teachers was significant (*b* = .928, *SE* = .023, *p* < .001), meaning that an increase of one in the teachers evaluation would increase .928 the school quality. Moreover, by default, the factor loading of the variable school (rating of school) is fixed to 1 (*b* = 1), thereby fixing the scale of the latent variable, but the completely standardized solution of school on the latent construct was = .932. Finally the variable accept, measuring the social preference of kids, was also significant (*b* = .477, *SE* = .026, *p* < .001), meaning that an increase of one unit in acceptance would increase school quality.

For home quality, parent’s education level (educate), was fixed to one by default (*b* = 1), thereby fixing the scale of the latent variable (home quality), the completely standardized solution of school on the latent construct was = .382. Furthermore, the household income was significant (*b* = 4.12, *SE* = .38, *p* < .001), meaning that an increase of one unit of income would increase home quality by 4.12. Finally, the parent’s involvement (involve) was also significant (*b* = 4.05, *SE* = .375, *p* < .001), meaning that for one unit in the involvement of the parents their home quality would increase by 4.05.

The structural model consisted of two regression equations. The first regression consisted of school quality and home quality to explain children’s achievement. School quality was not relevant to explaining this variable (*b* = .042, *SE* = .038, *p* = .264), but home quality was significant (*b* = 2.84, *SE* = .31, *p* < .001), meaning that an increase on the level of home quality (which was defined above), would increase the score on achievement by 2.84. The second regression equation consisted of achievement, school quality, and home quality as a function of CITO score. Achievement (*b* = .001, *SE* = .046, *p* = .982) and home quality (*b* = -.045, *SE* = .242, *p* = .853) were not significant to explain the CITO score. Nonetheless, school quality was significant (*b* = .983, *SE* = .044, *p* < .001), meaning that an increase of one in school quality (which was defined above), would increase CITO score by .983.

1. **The interpretation of the results in terms of the hypotheses.**

This results show an attempt to measure home and school quality with underlying observable variables that may proxy and explain the construct. In this case we observed that this observable variables (e.g. school score, parents income) strongly explain school and home quality, thus confirming our first hypothesis. Furthermore, our model explains the relationships between both latent constructs (school and home quality) with children’s achievement and CITO score using a different path than the original model. In this case home quality only explains achievement, and school quality only explains the CITO scores. Therefore showing a different prediction that our second hypothesis.

1. **A concluding sentence explaining of how these results extend the theory related to secondary school choice.**

By understanding the underlying variables that explain important constructs like home and school quality, we might now what to improve in order for kids to have more choice opportunities and thus, more motivation to acheive higher degrees in education levels.

**R code:**

takeh <- read.csv2("C:\\Users\\s4600479\\Desktop\\takehome2016.csv", sep = ";")

takeh2 <- read.csv2("C:\\Users\\s4600479\\Desktop\\takehome2016.csv", sep = ";")

#run teakeh2 and don't run with normalized variables... or littile mcar will not work

takeh <- read.csv2("C:\\Users\\André\\Google Drive\\Master\\period 4\\sem\\midterm\\takehome2016.csv", sep = ";")

takeh2 <- read.csv2("C:\\Users\\André\\Google Drive\\Master\\period 4\\sem\\midterm\\takehome2016.csv", sep = ";")

summary(takeh)

#packages

install.packages("knitr")

library(knitr)

install.packages("lavaan")

library(lavaan)

install.packages("semPlot")

library(semPlot)

install.packages("VIM")

library(VIM)

install.packages("mvnmle")

library(mvnmle)

install.packages("BaylorEdPsych")

library(BaylorEdPsych)

install.packages("psych")

library(psych)

install.packages("pastecs")

library(pastecs)

install.packages("lattice")

library(lattice)

options (scipen = 12)

####################PART A########################

#skewed kurtosed

densityplot(takeh$sex)

stat.desc(takeh$sex, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$educate, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$income, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$involve, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$school, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$teach, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$accept, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$achieve, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

stat.desc(takeh$CITO, basic=TRUE, desc=TRUE, norm=TRUE, p=0.95)

#outliers

takeh$zsex <- scale(takeh$sex, center = TRUE, scale = TRUE)

describe(takeh$zsex)

#nrow(subset(manova\_1, abs(scale(ESTEEM)) >3)) just to remember that i can know exactly which pp

takeh$zeducate <- scale(takeh$educate, center = TRUE, scale = TRUE) #this one

describe(takeh$zeducate)

length(which(abs(takeh$zeducate)>3))

takeh$zincome <- scale(takeh$income, center = TRUE, scale = TRUE)

describe(takeh$zincome)

takeh$zinvolve <- scale(takeh$involve, center = TRUE, scale = TRUE)

describe(takeh$zinvolve)

takeh$zschool <- scale(takeh$school, center = TRUE, scale = TRUE) #this one

describe(takeh$zschool)

length(which(abs(takeh$zschool)>3))

takeh$zteach <- scale(takeh$teach, center = TRUE, scale = TRUE) #this one

describe(takeh$zteach)

length(which(abs(takeh$zteach)>3))

takeh$zaccept <- scale(takeh$accept, center = TRUE, scale = TRUE) #this one

describe(takeh$zaccept)

length(which(abs(takeh$zaccept)>3))

takeh$zachieve <- scale(takeh$achieve, center = TRUE, scale = TRUE) #this one

describe(takeh$zachieve)

length(which(abs(takeh$zachieve)>3))

takeh$zCITO <- scale(takeh$CITO, center = TRUE, scale = TRUE)

describe(takeh$zCITO)

######################patterns in NA-s#########################

#missing values

takeh$sex <- factor(takeh$sex, levels = c("1","2"), labels = c("female", "male"))

takehf <- subset(takeh, sex == "female")

takehm <- subset(takeh, sex == "male")

miss1<- aggr(takeh)

miss1

missmales<- aggr(takehf)

missfemales<- aggr(takehm)

#identify missing values between continuous and categorical measures

d<- takeh[ ,c("sex","educate")]

barMiss(d)

d<- takeh[ ,c("sex","income")]

histMiss(d)

d<- takeh[ ,c("sex","involve")]

histMiss(d)

d<- takeh[ ,c("sex","school")]

histMiss(d)

d<- takeh[ ,c("sex","teach")]

histMiss(d)

d<- takeh[ ,c("sex","accept")]

barMiss(d)

d<- takeh[ ,c("sex","achieve")]

barMiss(d)

d<- takeh[ ,c("sex","CITO")]

barMiss(d)

###Little's MCAR test

MCAR<-LittleMCAR(takeh2)

MCAR$chi.square

MCAR$df

MCAR$p.value

MCAR$missing.patterns

MCAR$amount.missing

####correlations without nas listwise deletion

#pairwise VS LISTWISE deletion

corr.test(takeh[,2:9], use = "complete") #this is listwise

corr.test(takeh[,2:9], use = "pairwise")

###################PART B#######################

#9

modelp <- "schoolq =~ school + teach + accept

homeq =~ educate + income + involve

achieve ~ schoolq + homeq

CITO ~ achieve"

semPaths(modelp, residuals = F)

#specify the equations and the model

#school and home are latent vars because they are not observed, manifest vars are vars in the data file

modeleq <- "schoolq =~ 1\*school + teach + accept

homeq =~ 1\*educate + income + involve

homeq~~homeq

schoolq~~schoolq

schoolq~~homeq

educate~~educate

income~~income

involve~~involve

school~~school

teach~~teach

accept~~accept

accept~1

involve~1

school~1

teach~1

educate~1

income~1"

fit <- lavaan(modeleq, data = takeh, missing="FIML", estimator = "ML")

summary(fit, standardized = T, rsquare=T, fit.measures= T)

#plots

#semPaths(fit)

#semPaths(fit, residuals = F)

#semPaths(fit, what="std", residuals = F)

#semPaths(fit, what= "std", layout = "tree", rotation = 2, intercepts =T , residuals =F, curve = 2, nCharNodes = 0, edge.label.cex = 1, edge.color = "black", sizeMan = 8, sizeMan2 = 1.7)

##11

#goodness of fit

summary(fit, standardized = T, rsquare=T, fit.measures=T, modindices=T)

##12

#with robust estimator MLR

fit2 <- lavaan(modeleq, data = takeh, missing="FIML", estimator= "MLR")

summary(fit2, standardized = T, rsquare=T, fit.measures=T)

##13

#listwise deletion is default

fit3 <- lavaan(modeleq, data = takeh, missing = "ML", estimator = "ML")

summary(fit3, standardized = T, rsquare=T, fit.measures=T)

##14

modelmax <- "schoolq =~ 1\*school + teach + accept

homeq =~ 1\*educate + income + involve

achieve ~ schoolq + homeq

CITO ~ achieve

homeq~~homeq

schoolq~~schoolq

educate~~educate

income~~income

involve~~involve

school~~school

teach~~teach

accept~~accept

accept~1

involve~1

school~1

teach~1

educate~1

income~1

CITO~1

achieve~1

CITO~~CITO

achieve~~achieve

homeq~~schoolq"

fitm <- lavaan(modelmax, data = takeh, missing="FIML", estimator= "ML")

##15

#goodness of fit

summary(fitm, standardized = T, rsquare=T, fit.measures=T)

##16

summary(fitm, standardized = T, rsquare=T, fit.measures=T, modindices=T)

##18

modelimp <- "schoolq =~ 1\*school + teach + accept

homeq =~ 1\*educate + income + involve

achieve ~ schoolq + homeq

CITO ~ achieve + schoolq +homeq

homeq~~homeq

schoolq~~schoolq

educate~~educate

income~~income

involve~~involve

school~~school

teach~~teach

accept~~accept

accept~1

involve~1

school~1

teach~1

educate~1

income~1

CITO~1

achieve~1

CITO~~CITO

achieve~~achieve

homeq~~schoolq

"

fitimp <- lavaan(modelimp, data = takeh, missing="FIML", estimator= "ML")

#19

summary(fitimp, standardized = T, rsquare=T, fit.measures=T)

semPaths(fitimp, what= "std", layout = "tree", rotation = 2, intercepts =T , residuals =F, curve = 2, nCharNodes = 0, edge.label.cex = 1, edge.color = "black", sizeMan = 8, sizeMan2 = 1.7, "Estimates")

# citations

setwd("C:\\Users\\André\\Desktop")

write\_bib(x = .packages(all.available = T), file = "R-packages.bib",tweak = T)

Bibliography

Beaujean, A. A. (2012). BaylorEdPsych: R Package for Baylor University Educational Psychology Quantitative Courses. Retrieved from https://cran.r-project.org/package=BaylorEdPsych

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models Equation Models. *Educational Psychology Papaers and Publications*, 430–457.

Gross, K., & with help from Douglas Bates. (2012). mvnmle: ML estimation for multivariate normal data with missing values. Retrieved from https://cran.r-project.org/package=mvnmle

Kaplan, D., & Kaplan, D. (1989). Model Modification in Covariance Structure Analysis : Application of the Expected Parameter Change Statistic Model Modification in Covariance Structure Analysis : Application of the Expected Parameter Change Statistic, *24*(908311115), 285–305. http://doi.org/10.1207/s15327906mbr2403

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. http://doi.org/10.1037/1082-989X.7.1.64

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rosseel, Y. (2015). lavaan: Latent Variable Analysis. Retrieved from https://cran.r-project.org/package=lavaan

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. http://doi.org/10.1093/biomet/63.3.581

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2015). VIM: Visualization and Imputation of Missing Values. Retrieved from https://cran.r-project.org/package=VIM