Multivariate analysis of the admission results in the 2021 admission process to selective higher education in Chile (by Bladimir Padilla)

Context

The selective higher education admission process in Chile is, on the surface, simple: Students apply to selective university programs with a score composed of different academic factors, all with the same standardized scale of 150–850 points. These academic factors can be divided into two dimensions: first, the NEM and the Ranking, which are linked to high school grade point average; and second, the admission tests that measure the knowledge in basic areas of the general curriculum (language, mathematics, physics, biology, and chemistry) that will be necessary to study at the university.

Each program may assign different weights to the academic factors, so a student's composite score may vary depending on the programs to which they have applied. Programs can assign a maximum of 50% of the composite score between NEM and Ranking, while the weight of the admission tests is a maximum of 60%.

The admission process has several assumptions; one is that the admission results depend primarily on the academic factors, and that the socioeconomic characteristics of the students do not affect the process. However, different research findings in Chile have shown that admission test performance levels are correlated with student characteristics, such as the sector of the school they attended, the education of their parents, and the socioeconomic level of their family.

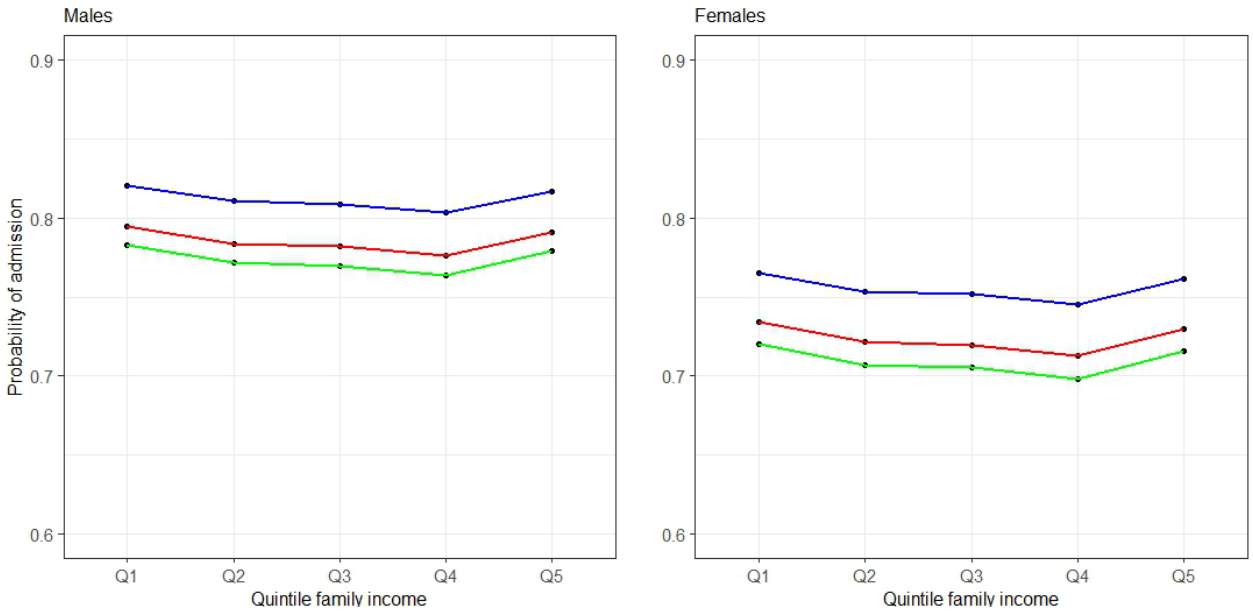
Such findings would extend the socio-economic inequalities of primary and secondary education into higher education and violate the assumption of equality of the admission process (in theory, all students should perform well on the admission tests, since they have learned from the same curriculum) and the assumption that non-academic factors should not predict the probability of admission to selective higher education.

Limitations of the last study

In the last study (using logistic regression), the research question was simple: Do students' socioeconomic characteristics predict the probability of admission to selective higher education in Chile? Using a binary logistic regression model where the dependent variable was admission outcome (0 = not admitted, 1 = admitted), we found that students' socioeconomic and demographic characteristics were significant predictors in the model even after controlling for academic factors and admission test scores. In particular, we found that the estimated probabilities of being admitted to selective higher education in 2021 varied significantly according to gender, sector of the college the student attended, and the socioeconomic status of the family.

Figure 1

Estimated probabilities of being admitted to selective higher education in the 2021 admission process according to Sex, school Sector, and Quintile family income



Although these findings are not new, the effect of family socioeconomic quintile was not as strong as we expected. Figure 1 shows that the differences between different quintiles are small, even though there are many arguments in the literature to assume that this variable tends to have a greater impact on students' educational opportunities. Of course, these effects are conditioned on the rest of the model's predictors, so perhaps the effect of this variable is secondary.

Another limitation is that we left out students who did not apply to selective universities, and perhaps this is not something random. The previous study focused on students who applied and were not admitted versus those who applied and were admitted, so the results and analysis could be contaminated by self-selection bias of not choosing to apply.

To address these two limitations, path analysis has several advantages over conventional regression models.

Current study

As before, this project is focused on the selective higher education admission process in Chile in 2021. The sample selection criteria were threefold: first, students who were participating for the first time in the admission process; second, who had learned the general curriculum for secondary education, which is used for the design of the admission tests; and third, who had valid scores in the academic factors, i.e., scores between 150 and 850.

Although the general context is similar, there are some important differences. In this study, the dependent variable is nominal (ADM) with three categories: does not apply, applies but is not selected, and applies and is selected. Previously we left those who did not apply out of the analysis, as we assumed that they were outside the focus of the research. However, socio-economic factors could also be affecting the participation or progress of this group in the admission process. To identify these students, the application database leaves their record blank, so we only had to replace the missing data (NA) with 0 (after selecting the sample with the three filters explained above). Then, the group of students who applied was divided into selected and non-selected using the application outcome variables. If the student was selected, a 1 was assigned, if not selected, 2. This coding was chosen because in Mplus, the highest category is used as the reference by default. Consequently, the different estimated submodels compared the probability of applying or being selected with the probability of applying but not being selected.

Another difference is the type of model. In this analysis, a multivariate model (path analysis) was used to estimate the mediation of socioeconomic factors in the relationship between socioeconomic characteristics of students and the probability of applying to a selective program (ADM#1) and the probability of being selected in a selective program (ADM#2). As we know, the univariate framework does not allow for this type of analysis, since its regressions only estimate bivariate relationships between one predictor and one outcome at a time. Moreover, for this analysis we only had observed variables, so we had no problems adjusting latent variables as in a SEM framework.

Finally, the socioeconomic variables included in the analysis were reduced. This project only includes the variables First generation student (FIRST_G) and Economic quintile of the family (FAMILY_INCOME). We could have included gender and type of school, significant predictors in the last project, but the former is not relevant to the current research question and the latter would introduce a problem that we have not addressed yet: nested data[1]. FIRST_G is a binary variable that identifies students with a parent who reached higher education and who are the first members of their family to reach higher education. The five-category FAMILY_INCOME was coded using a dummy schema, in which 4 new binary predictor variables were added to the model to represent it. Each new variable represented the difference between students of different economic levels (Q2, Q3, Q4, and Q5) compared to students of the first economic quintile.

Research question

What are the direct and indirect effects of socioeconomic variables on the probability of applying, being rejected, or being selected in selective higher education in Chile?

[1] This problem is being addressed in a parallel project that analyzes admission test results from a multilevel perspective with a clustered multilevel model.

Method

This project used a path model to estimate the effect of socioeconomic variables and academic factors on the probability of admission to selective higher education in the year 2021. Table 1 shows a detail of the model variables. Because the original scale of academic factors produced very small coefficients, these variables were divided by 100 and further centered on the 550 points (55), which is considered an adequate score to compete for selection in a selective program.

Table 1

Description of the variables included in the Path analysis (n = 183,935)

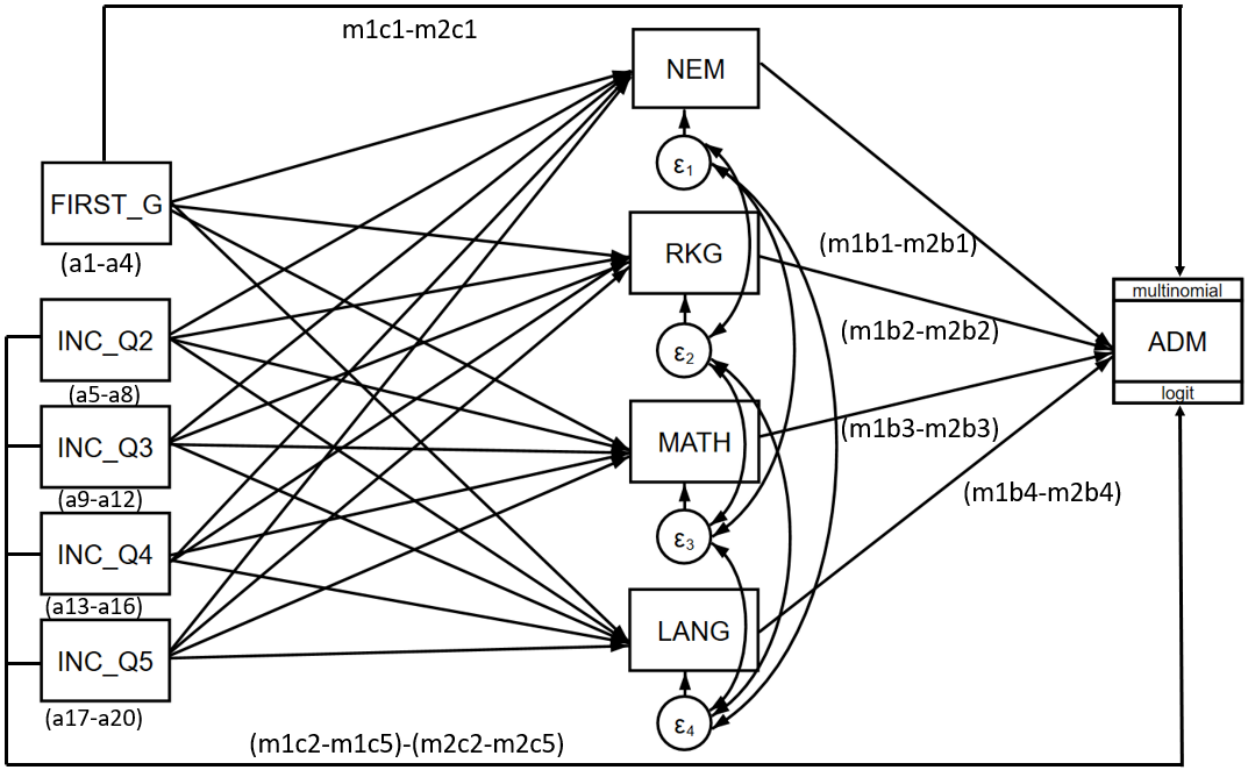
Variable	Description	Type
NEM	Standardized high school grade point average (GPA)	Continuous range = −2.99 – 2.90 mean = .452; sd = 1.053
RKG	Relative position of the student's GPA in his or her school of graduation	Continuous range = −2.99 – 3.00 mean = .672; sd = 1.231
LANG	Standardized score in the Language admission test	Continuous range = −4.00 – 3.00 mean = −.516; sd = 1.070
MATH	Standardized score in the Mathematics admission test	Continuous range = −4.00 – 3.00 mean = −.512; sd = 1.085
FIRST_G	First family member who could enter higher education	Binary 0 (No) = 40.4% 1 (Yes) = 59.5%
FAMILY_INCOME	Quintile of per capita income to which the student's family belongs (1 USD = 752.9 CLP)	Ordinal Q1 (REF.) = 30.2% INC_Q2 = 24.0% INC_Q3 = 14.3% INC_Q4 = 15.2% INC_Q5 = 16.2%
ADM	Student application status	Nominal 0 (does not apply) = 43.3% 1 (applies and is selected) = 43.9% 2 (applies but is not selected) = 12.8%

Results

Figure 2 shows a graphical representation of the path analysis. In parenthesis, you can see the labels used to identify the different regressions fitted. FIRST_G and the set of INC_Q predictors were treated as exogeneous predictors, thus their parameters (means, variance and covariances) were not included in the model. The admission factors (NEM and ranking; language and mathematics admission test scores) were both outcomes and predictors, and so their intercepts, and residual variances and covariances were estimated in the model. The multinomial distribution was chosen for the outcome residuals and the link function was logit.

Figure 2

Intended analysis



Note. The diagram labels in parenthesis are the same as those that appear in the files used in Mplus.

Table 2 shows the regressions, or multiple direct effects, estimated in the path model. The set of coefficients predicting ADM#1 correspond to the effect of the predictors on the probability of not applying versus the probability of applying but not being selected. The effects of ranking and admission tests were negative, so they would reduce the probability of not applying to any selective program, which is not surprising, especially if students have good scores on these factors. With respect to the family's economic quintile, the positive effects indicate that, compared to the first quintile, the probability of not applying

to a selective program increases as the family's economic resources increase. This is an interesting finding, but there are authors in the literature who suggest that students from wealthy families sometimes prefer to prepare for the next admission process if their results in academic factors do not allow them to enter high-demand programs. For a year, these students pay for remedial and preparatory programs to participate again in the admission process and obtain better results.

The second set of slopes (on ADM#2) estimate the effect of the predictors on the probability of applying and being selected versus applying but not being selected. Except for the NEM, the academic factors had a significant positive impact on the probability of being selected, something that meets the assumptions of the admission process (favorable results in these factors increase the probability of admission). Although the slopes are not standardized, they are on the same scale (15 to 85), so we could compare them and highlight the difference between the admission tests and the other two academic factors (good results in these tests seem more relevant than good results in NEM and ranking, although, in theory, these four factors would be closely correlated). Regarding family economic income, only the difference between the first and last quintile was significant, with a small difference in favor of students with families with more money.

Table 2

Path analysis results in Mplus (n = 180,295)

Variables	Label	Estimate	S.E.
ADM#1 ON			
NEM	m1b1	-0.515	0.041**
RKG	m1b2	0.155	0.033**
LANG	m1b3	-0.268	0.010**
MATH	m1b4	-0.027	0.009**
FIRST_G	m1c1	0.072	0.018**
INC_Q2	m1c2	0.073	0.020**
INC_Q3	m1c3	0.112	0.024**
INC_Q4	m1c4	0.233	0.025**
INC_Q5	m1c5	0.253	0.028**
ADM#2 ON			
NEM	m2b1	-0.079	0.043
RKG	m2b2	0.087	0.035*
LANG	m2b3	0.734	0.011**
MATH	m2b4	0.943	0.011**
FIRST_G	m2c1	-0.028	0.019
INC_Q2	m2c2	0.016	0.022
INC_Q3	m2c3	0.011	0.026
INC_Q4	m2c4	0.000	0.027
INC_Q5	m2c5	0.069	0.030*

NEM ON			
FIRST_G	na1	-0.300	0.005**
INC_Q2	na2	0.098	0.007**
INC_Q3	na3	0.137	0.008**
INC_Q4	na4	0.082	0.008**
INC_Q5	na5	0.393	0.008**
RKG ON			
FIRST_G	ra1	-0.277	0.006**
INC_Q2	ra2	0.104	0.008**
INC_Q3	ra3	0.137	0.009**
INC_Q4	ra4	0.066	0.009**
INC_Q5	ra5	0.351	0.009**
LANG ON			
FIRST_G	la1	-0.512	0.005**
INC_Q2	la2	0.166	0.006**
INC_Q3	la3	0.234	0.008**
INC_Q4	la4	0.183	0.008**
INC_Q5	la5	0.469	0.008**
MATH ON			
FIRST_G	ma1	-0.507	0.005**
INC_Q2	ma2	0.158	0.006**
INC_Q3	ma3	0.225	0.008**
INC_Q4	ma4	0.203	0.008**
INC_Q5	ma5	0.649	0.008**

Note. ** = p-value > 0.01, * = p-value > 0.05

The effect of being a first-generation student on both models also deserves some attention. Regarding the firs submodel (ADM#1), it had a significant positive effect, which indicates that, on average and conditional on the other predictors (the academic factors), first-generation students were more likely not to apply to any selective program than to apply but not be selected. This can be linked to at least three scenarios: that they decide not to apply for economic reasons (lack of benefits to cover tuition, fee or living costs), academic (low scores in academic factors), or personal (they believe that they would not be selected in the program they wanted, or lack of interest in higher education).

In the second submodel (ADM#2), FIRST_G had a negative effect; thus, on average, first-generation students were more likely to be rejected than admitted to a selective program in 2021 (after controlling for the academic factors). This can be linked to multiple reasons, especially with the number and type of educational choices made by first generation students. Perhaps their lower probability of admission is linked to the fact that they make fewer applications than the rest, or that other students with better composite scores also apply to the programs they apply for, or that in general they do not achieve scores that allow them to compete with the rest of the students.

The direct effects of socioeconomic predictors on academic factors (the set of *a* effects) followed a similar pattern to what we found in the previous study. Differences in average

scores between students from different economic quintiles grow between extremes, especially when looking at admission test scores (for example, students from families in the fifth income quintile would score, on average, 47 [$p < 0.001$] points higher than their peers from families in the first quintile on the language test and 65 [$p < 0.001$] points higher on the math test). Similarly disadvantaged would be first-generation students, who, on average, scored notably lower on the academic factors compared to those students whose one of the parents had attained higher education.

Taking advantage of path analysis, we calculated some indirect effects between the socioeconomic predictors and the dependent variable, through the academic factors, by multiplying the coefficients ‘a’ and ‘b’ of each path. Although it may seem obvious, these coefficients are not interpreted in the same way as the coefficients we have seen in class, since their 'effect on the dependent variable for every one unit of variation’ is not the same as in the linear slopes we have analyzed. Since both slopes (a and b) are on different scales, the indirect effects of our model can only help us to assess whether there is a difference between the unconditional regression of X on Y (c) and its conditional version (c') by the mediator. In other words, Table 3 will help us to know whether the academic factors are indeed mediators of the effect of the socioeconomic predictors on the students' probability of admission in 2021.

Table 3

*Indirect effects (unstandardized a*b effects)*

First generation student (FIRST_G)	Label	Estimate	S.E.
via NEM to ADM#1	m1b1*na1	0.154	0.012**
via RKG to ADM#1	m1b2*ra1	0.137	0.005**
via LANG to ADM#1	m1b3*la1	0.024	0.013
via MATH to ADM#1	m1b4*ma1	-0.376	0.007**
via NEM to ADM#2	m2b1*na1	-0.043	0.009**
via RKG to ADM#2	m2b2*ra1	0.014	0.005**
via LANG to ADM#2	m2b3*la1	-0.024	0.010*
via MATH to ADM#2	m2b4*ma1	-0.478	0.007**
Family income (Q5)			
via NEM to ADM#1	m1b1*na5	-0.202	0.016**
via RKG to ADM#1	m1b2*ra5	-0.126	0.005**
via LANG to ADM#1	m1b3*la5	-0.031	0.017
via MATH to ADM#1	m1b4*ma5	0.345	0.008**
via NEM to ADM#2	m2b1*na5	0.054	0.012**
via RKG to ADM#2	m2b2*ra5	-0.017	0.006**
via LANG to ADM#2	m2b3*la5	0.031	0.012*
via MATH to ADM#2	m2b4*ma5	0.612	0.011**

Note. ** = p-value > 0.01, * = p-value > 0.05

All indirect effects, except for FIRST_G and Q5 on ADM#1 via LANG, were significant, so we can conclude that, in our sample, academic factors are intermediary in the effects of socioeconomic predictors on the probabilities of admission.

Before concluding, it is important to mention that the two included socioeconomic variables were able to significantly explain part of the variance of the academic factors (Table 4), especially the admission tests.

Table 4

R-square estimates

Variable	Estimate
NEM	0.048
RKG	0.029
LANG	0.102
MATH	0.126

Finally, it is important to review our research question (and how we have overcome some of the limitations of the previous study). The results of this project show that the effect of socioeconomic variables varies across the nominal submodels. In general, being a first-generation student and belonging to a low-income family (Q1) instead of a high-income family (Q5) would increase the probability of not applying to any selective program (ADM#1) in the 2021 admission process and would decrease the probability of being selected once they have decided to apply (ADM#2).

Mplus code:

```
TITLE:  Second example on how to use models in research projects - Path analysis
DATA:   FILE = datADM.csv;      ! Can just list file name if in same folder;
        FORMAT = free;          ! FREE (default) or FIXED format;
        TYPE = individual;      ! Individual (default) or matrix data as input;
VARIABLE:
! List of ALL variables in original wide data file, in order;
! Mplus names must use 8 characters or fewer (so rename as needed);
  NAMES = NEM RKG LANG MATH FIRST_G INC_Q2 INC_Q3 INC_Q4  INC_Q5 ADM;

! List of ALL variables used in model;
  USEVARIABLE = NEM RKG LANG MATH FIRST_G INC_Q2 INC_Q3 INC_Q4  INC_Q5 ADM;

! Missing data identifier;
  MISSING = ALL (-999); !You can use any number not included in any of your variables;

! Categorical outcomes;
  NOMINAL = ADM; ! our dependent variable
  ! (1 = did not apply, 2 = selected, 3 = applied but not selected as nominal reference);

ANALYSIS: ESTIMATOR = MLR;                ! Robust full-information maximum likelihood;

OUTPUT:   CINTERVAL;                      ! Print confidence intervals;
          STDYX;                          ! Print fully standardized solution, too;
          SAMPSTAT;                       ! Print descriptive statistics;
          SVALUES;                        ! Print start values (help with convergence);

MODEL: ! * --> Estimated parameter (all listed below for clarity);

! Regressions: y outcomes ON x predictors (label to do math on later, * implied);

  ADM#1  ON NEM RKG LANG MATH FIRST_G INC_Q2 INC_Q3 INC_Q4 INC_Q5 (m1b1-m1b4 m1c1-m1c5);
  ADM#2  ON NEM RKG LANG MATH FIRST_G INC_Q2 INC_Q3 INC_Q4 INC_Q5 (m2b1-m2b4 m2c1-m2c5);
  NEM    ON FIRST_G INC_Q2 INC_Q3 INC_Q4 INC_Q5 (na1-na5);
  RKG    ON FIRST_G INC_Q2 INC_Q3 INC_Q4 INC_Q5 (ra1-ra5);
  LANG   ON FIRST_G INC_Q2 INC_Q3 INC_Q4 INC_Q5 (la1-la5);
  MATH   ON FIRST_G INC_Q2 INC_Q3 INC_Q4 INC_Q5 (ma1-ma5);

! Brief explanation about the models !!!!!!!!!!!!!!!

! Because we are using a multinomial logistic regression model, two submodels were estimated.
```

```
! By default, Mplus defines the last variable category as the reference (if anyone knows
! how to change this, please let me know). In this case, ADM is a categorical variable with three
! levels: 1 = does not apply, 2 = selected, 3 = applied but not selected. The set of coefficients
! predicting ADM#1 are linked to the comparison between students who decided not to apply (category 1)
! and those who applied and were not selected (category 3). The other set of coefficients predicting
! ADM#2 are linked to the comparison between students who applied and were selected (category 2)
! and those who applied and were not selected (category 3).

! Following the traditional way of labeling effects in path analysis, coefficients labeled 'b' correspond to
! direct effects between the mediator and the outcome; coefficients labeled 'a' correspond to direct
! effects between the exogenous predictor and the mediator; coefficients labeled 'c' correspond to the direct
! effect between the exogenous predictor and the outcome.

! Estimated residual covariances for continuous outcomes;
  NEM  WITH RKG*;
  NEM  WITH LANG*;
  NEM  WITH MATH*;
  RKG  WITH LANG*;
  RKG  WITH MATH*;
  LANG WITH MATH*;

! Like ESTIMATE in SAS or LINCOM in STATA;
  MODEL CONSTRAINT:
! List names of estimated effects on NEW;
  NEW(m1fn m1fr m1fl m1fm m2fn m2fr m2fl m2fm m1qn m1qr m1ql m1qm m2qn m2qr m2ql m2qm);

  m1fn = m1b1*na1; ! FIRST_G --> NEM  --> ADM#1
  m2fn = m2b1*na1; ! FIRST_G --> NEM  --> ADM#2
  m1fr = m1b2*ra1; ! FIRST_G --> RKG  --> ADM#1
  m2fr = m2b2*ra1; ! FIRST_G --> RKG  --> ADM#2
  m1fl = m1b3*la1; ! FIRST_G --> LANG --> ADM#1
  m2fl = m2b3*la1; ! FIRST_G --> LANG --> ADM#2
  m1fm = m1b4*ma1; ! FIRST_G --> MATH --> ADM#1
  m2fm = m2b4*ma1; ! FIRST_G --> MATH --> ADM#2

  m1qn = m1b1*na5; ! QUINTILE 5 --> NEM  --> ADM#1
  m2qn = m2b1*na5; ! QUINTILE 5 --> NEM  --> ADM#2
  m1qr = m1b2*ra5; ! QUINTILE 5 --> RKG  --> ADM#1
  m2qr = m2b2*ra5; ! QUINTILE 5 --> RKG  --> ADM#2
  m1ql = m1b3*la5; ! QUINTILE 5 --> LANG --> ADM#1
  m2ql = m2b3*la5; ! QUINTILE 5 --> LANG --> ADM#2
  m1qm = m1b4*ma5; ! QUINTILE 5 --> MATH --> ADM#1
```

```
m2qm = m2b4*ma5; ! QUINTILE 5 --> MATH --> ADM#2
```