ON ANALYZING MATH LEARNING AND REASONING USING A DATA MINING APPROACH

Emilio Gerardo Sotto Riveros
emiliosotto@gmail.com
Polytechnic School, National University of Asunción
P.O. Box 2111 SL, San Lorenzo 2160, Paraguay

Santiago Gómez-Guerrero
sgomezpy@gmail.com - Corresponding author
Polytechnic School, National University of Asunción
P.O. Box: 2111 SL, San Lorenzo 2160, Paraguay

Christian E. Schaerer
chris.schaerer@cima.org.py
CIMA, Centro de Investigación en Matemática
Dr. César López Moreira 693 - P.O. Box 1766 - Asunción, Paraguay

Abstract

In this work we explore the use of data mining for understanding aspects that most influence the process of mathematical learning. Math performance data are taken from the answers given to a battery of problems by students in the age range from 11 to 17 years old, grouped into scores that correspond to math subjects like algebra, geometry, statistics and so on. Each score is considered a response variable. We attempt to explain scores with a second dataset consisting of the same students' answers to questions in the areas of school life, study habits, perceptions related to teachers, socio-economic factors and the family environment, looking for relationships between the personal characteristics of the students and the quality of their reasoning in math. Results so far show that the amount of homework and the motivation from school represent the largest variations that have an influence on math outcomes. By contrast, other groups of variables like the perceived usefulness of maths or home facilities for study seem unimportant for math score. With this case study, we show that new roads can be opened in the search for aspects that most influence the process of real learning in mathematics, and in the effort to gain a better understanding of differences among students.

Keywords: *math learning, math reasoning, school math, problem solving, data mining*

1 INTRODUCTION

Mathematics has earned fame to be a difficult subject in all programs of study at school, high school, and even at the university level. Several factors affect the right performance of math students in a manner not well understood, and the quality of reasoning in mathematics remains an open topic. In fact, it is observed in classrooms that some students have specially good or better than expected performance, while others don't, in spite of having started the course with apparently better chances.

Therefore, discovering the actual factors in the learning and reasoning process of students is key to help promote higher levels of success in their future life at university, and as professionals. Studies on educational programs with scientific orientation such as STEM (USDE, 2015), showed that low performing students had significantly higher improvement rates on mathematics scores than high and middle performing students, and in addition, ethnicity and social-economic status were good predictors of academic achievement (Han et al., 2014). In Latin America, this issue is even more dramatic for engineering and science careers (UNESCO, 2015)(OECD, 2015) but factors may be different.

Some data mining algorithms used in engineering problems have shown good potential for extracting rules that help to identify variables in student achievement in specific subjects. This benefits the work of educators and curriculum experts (ElGamal, 2013).

Thus, the inclusion of economic, family, social and cultural data in the analysis can help to identify, classify and determine trends, patterns and factors in the search of solutions to complex problems such as student dropout (La Red Martínez, 2016).

A large variety of available data may seem to be a plus; however, feature selection is usually necessary to identify which variables are dominant or more strongly affect certain educational outcome. In this context we look at principal component analysis as an important tool that allows to find multiple correlations in the dataset, while reducing many variables to a few relevant factors. With this somewhat less classical approach, we begin our search for insights to the problem of learning, reasoning, and achieving good performance in math. Results from this work will allow to address future research for improving student performance.

This article is organized as follows. A brief description of the Principal Component Analysis is presented in Section 2. Section 3 reports details on how data was gathered and then prepared for processing. In Section 4 we perform an exploration of data, select groups of variables that look promising, analyze the top groups of variables and perform a quick profiling of the best students. We finalize with conclusions in Section 5.

2 OVERVIEW OF PRINCIPAL COMPONENTS ANALYSIS

Principal Component Analysis (PCA) belongs to a group of multivariate statistical techniques. It is a descriptive technique that has become attractive for a variety of fields, because it helps to reduce the complexity inherent to having multiple variables.

PCA allows to reduce the dimensionality of the data; it transforms an original set of p variables into another set of q uncorrelated variables where $p \geq q$, called *principal components*. The p variables are measured on each of n individuals or objects, obtaining an array of data of order np, where $p < n$.

2.1 Using Real Principal Components Analysis

When all variables are numeric, real (also called standard) PCA can be applied. The analysis is performed in the space of the variables and in the space of the individuals. The variable points and the individual points are represented graphically using the calculated principal components as new coordinate axes; it is fairly common practice to just use the first two components (PSU, 2017). Often, it can facilitate the interpretation of results to observe the similar location of the points in this new system of coordinates. Although the plane of variable points does not overlap the plane of individual points, it is useful to interpret the proximity of a group of individual points to certain variables.

If X is a random vector of dimension p with finite $p \times p$ variance-covariance matrix $V[X] = \sum$, then the principal component analysis (PCA) solves the problem of finding the directions of the greatest variance of the linear combinations of observations. In other words, it seeks the orthonormal set of coefficient vectors $a_1, \ldots, a_k$ such that

$$\begin{aligned}
\mathbf{a}_1 &= \underset{\mathbf{a}:\|a\|=1}{\arg\max}\ \mathbb{V}[\mathbf{a}'\mathbf{x}] \\
&\vdots \\
\mathbf{a}_k &= \underset{\substack{\mathbf{a}:\|a\|=1 \\ a \perp a_1,\ldots,a_{k-1}}}{\arg\max}\ \mathbb{V}[\mathbf{a}'\mathbf{x}] \\
&\vdots
\end{aligned}$$

where $\|a\|$ is the norm of a. The maxima are those of a convex function on a compact set, and thus exist, and are unique if no perfect collinearity exists in the data, up to the change of the sign of all elements of $\mathbf{a}_k$. The linear combination $a_k' X$ is referred to as the k-th principal component (PC).

The motivation behind this problem is that the directions of greatest variability give "most information" about the configuration of the data in multidimensional space.

Let the random vector of p variables be $X^t = [X_1 X_2 \ldots X_p]$ with variance-covariance matrix $\sum$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ and $a_1, a_2, ..., a_p$ are the eigenvalues and eigenvectors corresponding to $\sum$. Consider the following linear combinations:

$$\begin{aligned}
Y_1 &= a_1^t X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\
Y_2 &= a_2^t X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\
&\vdots \\
Y_p &= a_p^t X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p
\end{aligned}$$

It can be proven that

$$\begin{aligned}
Var(Y_i) &= a_i^t \sum a_i & i &= 1, 2, ..., p \\
Cov(Y_i, Y_j) &= a_i^t \sum a_j & i, j &= 1, 2, ..., p
\end{aligned}$$

The main components are the linear combinations $Y_1, Y_2, \ldots, Y_p$ that are not correlated with each other and whose variances satisfy $Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p) \geq 0$. The main components are then defined as follows:

- The first principal component is the linear combination $Y_1 = a_1^t X$ which maximizes $Var(a_1^t X)$ subject to $\langle a_1, a_1 \rangle = 1$.
- The second main component is the linear combination $Y_2 = a_2^t X$ that maximizes $Var(a_2^t X)$ subject to $\langle a_2, a_2 \rangle = 1$ and $Cov(Y_1, Y_2) = 0$.
- In general the i-th main component is the linear combination $Y_i = a_i^t X$ that maximizes $Var(a_i^t X)$ subject to $\langle a_i, a_i \rangle = 1$ and $Cov(Y_i, Y_k) = 0$ for $k < i$.

Thus, under this method we have the following result: Consider the variance-covariance matrix $\sum$ associated with the vector $X^t = [X_1 X_2 \ldots X_p] \in R^p$, and let $(\lambda_1, a_1), (\lambda_2, a_2) ... (\lambda_p, a_p)$ be eigenvalues and eigenvectors corresponding to the matrix $\sum$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Then the i-th main component is given by:

$$Y_i = a_i^t X = a_{i1} X_1 + a_{i2} X_2 + ... + a_{ip} X_p \qquad i = 1, 2, ..., p$$

subject to the following conditions:

$$Var(Y_i) = a_i^t \sum a_i = \lambda_i \qquad i = 1, 2, ..., p$$

$$Cov(Y_i, Y_j) = a_i^t \sum a_j = 0 \qquad i \neq j$$

The largest proportion of the total variance of the population explained by the main components is given by

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_k} \qquad k = 1, 2, ...p \tag{1}$$

The first principal component will have the greatest variance and extract the largest amount of information from the data; the second component will be orthogonal to the first one, and will have the greatest variance in the subspace orthogonal to the first component, and extract the greatest information in that subspace; and so on. Also, the principal components minimize the L_2 norm (sum of squared deviations) of the residuals from the projection onto linear subspaces of dimensions 1, 2, etc.

The first PC gives a line such that the projections of the data onto this line have the smallest sum of squared deviations among all possible lines. The first two PC define a plane that minimizes the sum of squared deviations of residuals, and so on.

The principal components analysis can be carried out for both the theoretical distributions and the actual data. In the latter case, one would analyze the empirical covariance matrix. Plotting several first components against each other can often give good insight into the structure of the data, presence of clusters, nonlinearities, outliers, etc.

There are a number of practical choices that researchers have to make when performing the principal component analysis. The first one is to choose what variables to include in the analysis. The desirable choice is that all variables describe a common phenomenon. As far as PCA was originally developed for the multivariate normal distribution and samples from it, the PCA will work best on the variables that are continuous and at least approximately normal.

2.2 Using Categorical Principal Components Analysis (CatPCA)

We frequently encounter discrete data in the analysis of observations from educational, social or health studies. Often the discrete data are binary, that is, a variable that can only take one of two values, such as gender (male/female), or ownership of a house (yes/no). Other times, discrete variables show several categories, as in type of seat in an airplane, industry of a firm, or geographical region.

If there are several categories of a discrete variable, they may have some natural ordering and the variable is referred to as ordinal, because categories can be listed using a monotone relation between them (for example, *antecedes* or *goes before*). Two examples might be: the level of education of a person (primary, secondary, higher, professional or advanced degree), and the opinion on certain political issue being debated (fully agree, partially agree, undecided, somewhat disagree, completely disagree). Perhaps in the latter example we could use codes such as (2, 1, 0, -1, -2); but notice these numeric labels are just an arbitrary choice. However, care needs to be taken so as to avoid using codes as a continuous variable.

Yet another type of discrete data is tally or count data, such as the number of take-home assignments given to students in a week, or the number of successful students in a given subject.

When the observed discrete x_k data or labels are used as they come, analyzing them with standard principal component analysis has at least two implications.

Firstly, discrete data, specially those having few categories, only have a probability mass function (no density), hence the distributional assumptions including normality are violated. Also, even with their finite range, the discrete data can exhibit high skewness and kurtosis, especially if most data points are concentrated in a single category.

Secondly and more important, a consequence of the discreteness is that the covariances or correlations computed between the discretized versions X_1^*, X_2^* of any two variables of interest do not reflect the "true" covariances or correlations of the (unobserved or unknown) underlying variables X_1, X_2. But in rigor, covariances of discretized versions cannot be computed, unless we use category labels as if they were values of continuous variables.

An additional challenge is that natural ordering of categories is not generally kept by the principal component analysis, so the only way to identify such ordering would be the use of ordinal variables for which the higher values really mean higher scores in the response variable. But many times, we may be unaware of which is the "correct order" before the first few runs of the algorithm.

The analysis of categorical main components is a multivariate statistical technique used for the reduction of data, which makes no assumptions of normality, in which we study p observable variables, $x_1, x_2, ..., x_p$, through which we will generate other k unobservable variables, $k < p$.

The real principal component analysis assumes linear relationships between the numeric variables, whereas the Categorical PCA allows to scale the variables to different levels. Categorical variables are optimally *quantified* in the specified dimensionality. As a result, non-linear relationships between variables can be modeled.

The CatPCA statistical technique is used for the reduction of any mix of p nominal, ordinal and numeric data, without making any normality assumptions. CatPCA performs the optimal scaling of categorical variables, by assigning appropriate numeric values to the differ-

ent responses. It also reduces dimensionality to q new dimensions, where $p \geq q$, by applying standard PCA to all variables transformed to numeric scale.

As with PCA, having a random vector p, we obtain the matrix of variances and covariances from which the principal components are extracted. These components are linear combinations that are not correlated with each other. Then the first component is designated as the linear combination with maximum variance, the second component is the linear combination with the second maximum variance, and so on, proceeding in the same way to find the other components.

Again, the number of principal components chosen will depend on what percent of total variance one needs to explain. The technique is most useful when an extensive number of variables prevents an effective interpretation of the relationships between objects (cases or units). Under a reduced dimensionality, a small number of components is interpreted instead of an extensive number of variables (Alvarez, 2005). In this work we use the IBM SPSS (IBM, n.d.) implementation of Categorical PCA.

3 DATA GATHERING AND PREPROCESSING

Here we give a brief account of how data were gathered and prepared prior to the PCA (Principal Components Analysis) processing.

Student population: The Young Talents project is managed in Paraguay by OMAPA, a not for profit organization of educators and other professionals running Math Olympics tests and a variety of other educational activities every year, seeking improvement in the students' math performance. Students from public, private and subsidized schools throughout Paraguay who consistently participate in preparation towards Math Olympics, can be enrolled as Young Talents and are monitored during their school years.

For a student to be part of the Young Talents project, it is necessary that he/she has been summoned to be among the students with the best scores in the National Mathematical Olympics (OMAPA, 2017), a student competition held each year. The current enrollment in the program is over 90,000 students.

Instruments: A 15-question test in mathematics was administered to a group of 110 students in the age range of 11 to 17 years old, from 26 public and private educational institutions, who are part of the Young Talents project in the Greater Asuncion geographical area. The difficulty level for each item was set in accordance with Bloom's taxonomy (Bloom, 1956). Each question in the test was pre-classified into one of the 5 following areas: arithmetics, algebra, geometry, logic, and statistics. The global math grade was also computed as the sum of all individual grades.

In addition, a second instrument was given to the same students. This was a questionnaire in a survey-style design, with 61 questions on various areas of the respondent's life. Each question was single-choice in a Likert ordinal scale of four (sometimes five) opinions or options (Likert, 1932). These questions were classified according to their topics as:

- Parental follow-up,
- Teachers' commitment,
- Student's commitment,

- Work assigned for home,
- Home facilities for study,
- Self-appraisal in math,
- Motivation for maths,
- Usefulness of maths,
- Preference for the school,
- Relationship with classmates.

Each student's responses to this instrument were joined to his/her responses to the math olympics test instrument, thus enabling for analysis of the whole dataset or selected portions thereof.

Administration of the instruments: Both the math exam and the student survey were administered in the same school to all selected Young Talents students. The administration of instruments was completed in one day.

Data preprocessing: Dirty data such as multiple responses to a survey question where only one response was expected, were identified and deemed as missing data. Data on ordinal Likert scales were assigned increasing numeric values going from the "bad" to the "good" perceptions; this was possible because of the way all questions were presented. The math test results, only numeric variable in the study, were transformed into categories from 1 to 7 for easier interpretation of visual results.

No missing values existed in the math exam instrument; however, there were values missing in the survey responses. These missing values were replaced by the mode (most frequent response) of the corresponding variable, just before the run. An exception to this rule was when all answers from one respondent were missing, in which case the record itself was deleted just before the run.

Standardization is considered a need when not all variables are measured in the same units. This applies to our study, therefore all data were standardized at the start of each run.

4 EXPLORING AND ANALYZING

We ran a few exploratory analyses with PCA using all variables in the survey, with one of the six math results (arithmetics, algebra, geometry, logic, and statistics plus the global result). However, the total variance explained by the first two or three principal components stayed under 50%, hindering our goal of reducing dimensionality.

In a series of new attempts we only took one group of questions from the survey with the global math result each time. In this effort we use our prior "knowledge of the field" by isolating certain aspect of the survey, say for example parental follow-up, and seeing how it correlates with math outcomes. The following table gives the percent of total variance explained by the first two components, in this series of runs.

As preliminary results, Table 1 shows that the most promising groups of variables, as far as explaining reasonably high percentages of variance, are:

Table 1: Percent of Total Variance Explained by First 2 Dimensions

Group	*Percent Variance*
Work assigned for home and Global MAT score	84.988%
Motivation for maths and Global MAT score	72.372%
Preference for the school and Global MAT score	69.823%
Self-appraisal in math and Global MAT score	60.969%
Teachers' commitment and Global MAT score	58.489%
Relationship with classmates and Global MAT score	55.201%
Parental follow-up and Global MAT score	52,356%
Usefulness of maths and Global MAT score	52.178%
Student's commitment and Global MAT score	51.169%
Home facilities for study and Global MAT score	33.937%

- Work assigned for home vs Global MAT score
- Motivation for maths vs Global MAT score
- Preference for the school vs Global MAT score

We now analyze the results obtained from the three datasets with highest percentage of explained variance.

4.1 Work assigned for home vs Global MAT score

We analyze questions related to how much homework is assigned by math teachers and the time spent by the student to resolve assignments, together with the Global MAT score. The amount of Variance explained by the first two Dimensions is 50.146% and 34.842%, accumulating a total of 84.988% between the two.

Table 2: Saturations on Components:
Two variables on homework with MAT score

Variable	*Dim 1*	*Dim 2*
Frequency of homework	0.873	-0.167
Time I need to resolve homework	0.860	0.237
MAT	-0.59	0.981

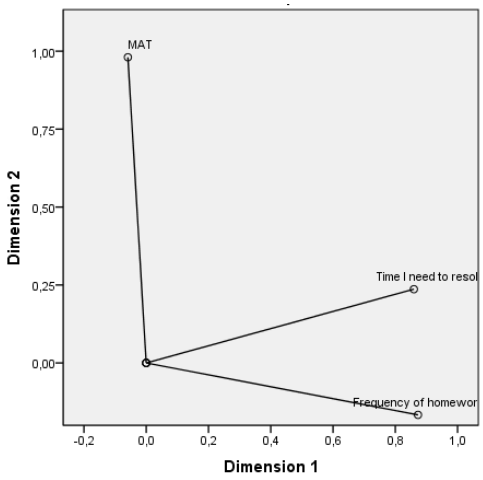

Figure 1: Time devoted to math at home, orthogonal with MAT score

In the first dimension, we observe higher weight in *Frequency of homework* and *Time I need to resolve homework*, both with positive sign. Together they may be regarded as *Time devoted to math at home*. Big values on this dimension indicate students with a lot of homework who resolve it in relatively short time.

For the second dimension the only variable is the MAT test score, with positive sign, so this is just the *Global MAT score* dimension. High values here are of course indicative of a good result in MAT.

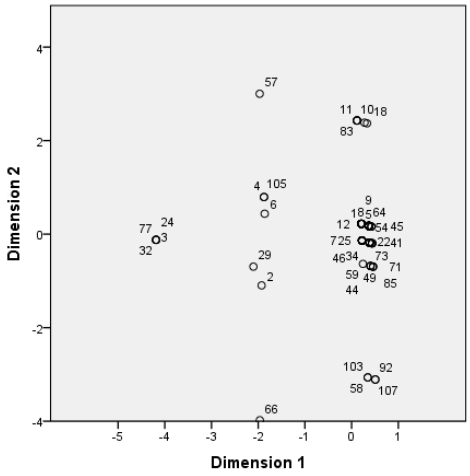

Figure 2: Students projected on the two dimensions. Labels represent student numbers.

The orthogonality of both dimensions suggests that the amount of homework assigned is almost unrelated to the global MAT score. This is clearly verified by observing the symmetry of the previous figure with respect to the Dimension 1 axis, and also by comparing the two groups of students {10,11,18,57,83} and {58,66,92,103,107}: both are on the "lot of time for math at home" side; the first group gets high math grades whereas the second one gets low grades.

Looking at the quadrants, in the upper right we have students who get lots of homework assigned, resolve their task quickly and made high grades in the test; the greatest concentration of points is found there. By contrast, in the lower right quadrant we have students with lots of

homework, resolve their tasks quickly and made low grades in the test; this quadrant has the least number of points.

4.2 Motivation for maths vs Global MAT score

Another important outcome is the analysis of questions referring to the enthusiasm, motivation and interest that the student feels towards maths, joined with the global MAT score. This produced 46.696% of explained Variance on Dimension 1, and 25.676% of explained Variance on Dimension 2, accumulating a total of 72.372%.

Table 3: Saturations on Components:
Motivation and Thrill for math together with MAT score

Variable	*Dim 1*	*Dim 2*
The math subject is boring (flipped)	0.285	0.812
I don't want to study math (flipped)	0.141	0.805
I enjoy learning math	0.941	0.004
The things I learn are interesting	0.821	0.247
It's important to do well in maths	0.875	-0.321
I like maths	0.911	-0.230
MAT	0.116	-0.522

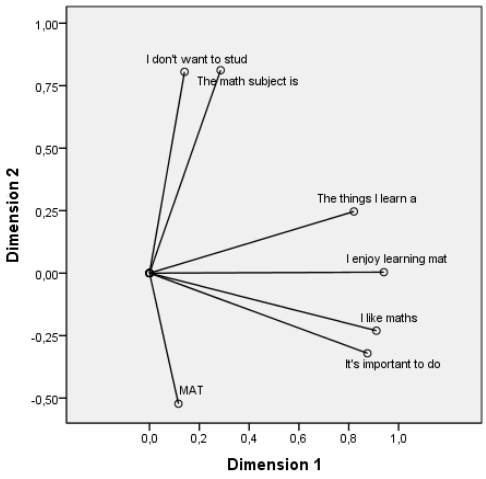

Figure 3: Motivation for maths and Thrill for math,
made up of several attributes

In the first dimension the greater weight is observed on questions *I enjoy learning math*, *I like maths*, *It's important to do well in maths* and *The things I learn are interesting*, all with positive sign. This reveals a *Motivation for maths* dimension.

The second dimension is like the other face of the first. At data input time we flipped the ordinal scale of answers, lining them up with the rest of the survey questions. Thus the

opinions should be read as *The math subject is not boring* and *I want to study math* in a natural "higher is better" sense. These opinions conform the second dimension together with the *Global MAT score* which is negatively related. In our interpretation this dimension expresses *excitement* for math or *Thrill for math*, showing that students can be attracted by school maths but not necessarily perform well in math olympics. Some students might even entertain the idea of studying math. Here, big values indicate excitement for maths with light tendency to get lower MAT grades, while small values suggest disenchantment towards math subjects with a surprising slight tendency for higher MAT grades − an apparent contradiction that many of us have experienced in life.

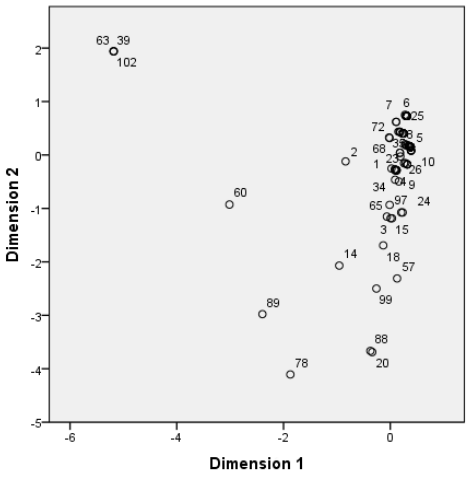

Figure 4: Students projected on Motivation and Thrill for math dimensions.

In the upper right quadrant, we have students who feel motivated towards math and hold high expectations about it, with a relatively low global MAT score; this is the most populated quadrant. The upper left quadrant shows students less motivated for math and with high expectations; this is the least populated quadrant.

4.3 Preference for the school vs Global MAT score

In this group we analyze questions on how students feel inside the school, together with the global MAT grade. The amount of Variance explained on Dimensions 1 and 2 are 40.269% and 29.553% respectively, adding up to 69.823% on two principal components.

Table 4: Saturations on Components:
Preference for school together with MAT score

Variable	*Dim 1*	*Dim 2*
I feel safe at school	0.359	0.714
I like being at school	0.743	-0.498
I prefer school better than another place	0.855	-0.193
MAT	-0.448	-0.622

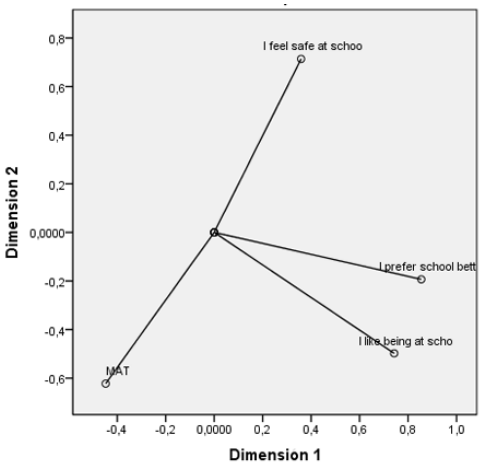

Figure 5: Liking school and Finding protection at school dimensions.

The first dimension shows greater weight at *I prefer school better than another place* and at *I like being at school*, both with a positive sign, hence it can be considered as a *Liking school* dimension. High values in this dimension indicate that the student likes school.

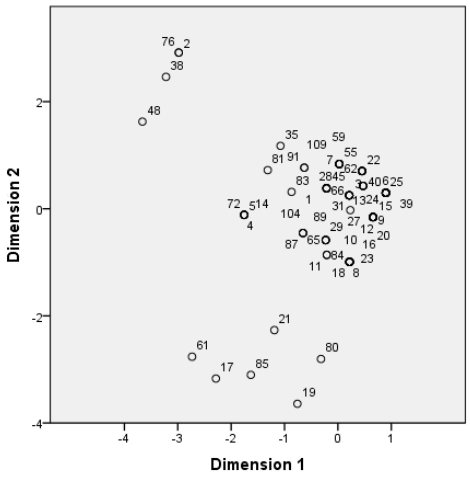

Figure 6: Students projected on the Liking school and Finding protection at school dimensions.

In the second dimension, the variables with higher weight are *I feel safe at school* with positive sign and *Global MAT score* with negative sign. This is a dimension about *Finding protection at school*, and it carries the MAT score negatively associated with the student's security feeling. High values in this dimension indicate that the person values safety at school and tends to get lower math results.

The majority of object points remain relatively close to the new origin of coordinates. The top-right quadrant corresponds to students who like school and feel safe at school; this is where most points are concentrated. In the top-left quadrant we see students who don't like school but feel secure there and make relatively lower grades; this is the least populated quadrant.

4.4 A profile of the top scoring students

As an additional insight, in the following table we offer a profile that characterizes the 30 students who scored above 60% in global MAT. For each question stated in the first column, in the second column we record the most frequent answer (the statistical mode) chosen by that subset of the students.

Table 5: Top scoring students - Work assigned for home

Question	*Answers*
Work assigned for home	1 per week
Time I need to resolve homework	1-15 min

Table 6: Top scoring students - Motivation for maths

Questions	*Answers*
Math subject is boring	Strongly disagree
I don't want to study math	Strongly disagree
I enjoy learning math	Strongly agree
The things I learn are interesting	Strongly agree
It's important to do well in math	Strongly agree
I like maths	Strongly agree

Table 7: Top scoring students - Preference for the school

Questions	*Answers*
I feel safe at school	Strongly agree
I like being at school	Strongly agree
I prefer school better than another place	Partially agree

5 CONCLUSIONS AND FUTURE WORK

We have presented a case study that allowed the discovery of some trends related to student performance in math tests, using a categorical Principal Component Analysis (CatPCA) tool. The study was based on a survey of family and personal study aspects of students' life, and on results of the OMAPA test looking at math olympics, taking into account both the global test grade and the influence that the school and human environment may have. We also illustrated the use of CatPCA for data exploration in the presence of some prior knowledge: we split the

data into several runs, by grouping related questions from the survey together with the math results, and the tool helped to identify which groupings had higher contributions to the total variance. Brief conclusions on the main results follow.

Homework assigned: We observe that frequent assignment of homework and speed to resolve tasks are not associated to high grades; there exists a wide dispersion of low and high math results in students with little homework from school as well as in students with a lot of assignments given.

Motivation and sympathy of students for school math, is weakly related with good performance in a young talents math test — perhaps contrary to general belief. Expectations related to the math subjects manifested themselves as a component separated from motivation, that is, many students could be motivated towards math but at the same time may have little desire for studying it in the way school presents math. These aspects will need to be better clarified in further studies.

Preference and security: A majority of students prefer or likes their school, together with the protection that the institution gives them. However, those who say they feel secure, tend to achieve lower performance in math as a young talent.

A quick profiling of the top 30 students and their most frequent answers to these three groups of survey questions, shows opinions and characteristics that are generally expected from students who perform well.

In other topics such as parental follow-up, bullying or physical facilities for the student, often seen as fundamental, we haven't found any dimensions with high percentage of variance explained. Nevertheless, those factors may be important for success in areas of study outside of mathematics, which were not considered here.

These analyses are all exploratory, but they suggest to go for more studies on motivation, amount of assigned homework, and preference for the school, by employing methods more oriented for detecting cause-effect relationships. Likely, the dissection of these aspects of school life will allow to find characteristics of students with high performance in math and also characteristics of those who do not achieve that level.

The identification of variables involved in student math performance using data mining and analysis, may help educators to steer curricular changes and to evaluate the effects of implemented changes, and possibly, to improve the generalization of solutions.

ACKNOWLEDGEMENTS

This work is partially supported by CONACyT-Paraguay project 14-INV-186 CABIBESKRY. CES acknowledges the CONACyT-Paraguay PRONII program. SG acknowledges support given by CONACyT-Paraguay project PINV15-706 COMIDENCO. Authors thank Gabriela Gómez Pasquali and Claudia Montanía from OMAPA, for fruitful discussions on topics of the work.

REFERENCES

Álvarez Figueroa, Lorena María, 2005. *Análisis Estadístico Multivariado del Impacto de la Emigración de los Familiares en los Estudiantes Adolescentes de la ciudad de Guayaquil (Ecuador)*. In: *Escuela Superior Politécnica del Litoral*, pp. 141-144.

Bloom, Benjamin S., 1956 (Ed.). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. David McKay Company, Inc., pp. 201-207.

ElGamal, A.F., 2013. *An Educational Data Mining Model for Predicting Student Performance in Programming Course*. Department of Computer Science, Mansoura University, Egypt.

Han, Sun Young; Capraro, Robert M.; Capraro, Mary Margaret, 2014. *How science, technology, engineering and mathematics STEM project-based learning affects high, middle and low achievers differently: The impact of student factors of achievement*. In: *International Journal of Science and Mathematics Education 2014*.
URL: https://www.researchgate.net/publication/271658486.

IBM Knowledge Center. *Categorical Principal Components Analysis (CATPCA)*.
URL: https://www.ibm.com/support/knowledgecenter/en/

La Red Martínez, David; Karanik, Marcelo; Giovannini, Mirtha; Báez, María Eugenia; Torre, Juliana, 2016. *Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos*. Universidad Tecnológica Nacional, Argentina.
URL: http://uajournals.com/ojs/index.php/campusvirtuales/article/view/128

Likert, Rensis, 1932. *A Technique for the Measurement of Attitudes*. In: *Archives of Psychology*, 140: pp. 155.

OECD, Organisation for Economic Cooperation and Development, 2015. *Programme for International Student Assessment (PISA)*. Paris, PISA 2015 Results.
URL: http://www.oecd.org/pisa/pisaenespaol.htm

OMAPA, Organización Multidisciplinaria de Apoyo a Profesores y Alumnos, 2017. *Olimpiadas Nacionales de Matemática*.
URL: http://www.omapa.org/?post_projects=olimpiadas-nacionales-de-matematica

PSU, Pennsilvania State University Eberly College of Science, 2017. *Stat 505 Applied Multivariate Statistical Analysis*.
URL: https://onlinecourses.science.psu.edu/stat505/node/49

UNESCO, Oficina Regional de Educación para América Latina y el Caribe - Santiago, 2015. *Tercer Estudio Regional Comparativo y Explicativo: Factores Asociados*.
URL: http://unesdoc.unesco.org/images/0024/002435/243533s.pdf

USDE, United States Department of Education, 2015. *Science, Technology, Engineering and Math: Education for Global Leadership*.
URL: https://www.ed.gov/stem