What distinguishes high-performing students?

Will Landau

April 25, 2015

1 Introduction

Which factors best separate successful students from those who struggle? How well can we predict academic success using conditions we can observe and control? For insight, I look at data from the Organization for Economic Co-operation and Development (OECD). In 2012, The OECD's Programme for International Student Assessment (PISA) surveyed roughly five hundred thousand, fifteen-year-old students from sixty-five economies across the globe ("Organization for Economic Co-operation and Development" 2015). Questions measured students' reading, math, and science skills with examinations that, according to the OECD website, "are not directly linked to the school curriculum. The tests are designed to assess to what extent students at the end of compulsory education, can apply their knowledge to real-life situations and be equipped for full participation in society" ("Organization for Economic Co-operation and Development" 2015). Students also answered extensive background questionnaire about their study habits, attitudes towards school, circumstances at home, etc., al of which are factors that may influence student success. In the analysis below, I derive a "student success" variable from the reading and math scores and attempt to predict success using information from the background questionnaire.

2 A first cleanup: getting ready to explore

The student-specific 2012 PISA dataset is large and messy, and it needs to be cleaned and subsetted both before and after exploratory analysis. For example, to save computing time, and because overall pedagogy and the survey's implementation are different among different countries, only students from the United States will be examined here.

2.1 Variables for predicting success

There are around 500 variables from the student background questionnaire, and large fraction of the answers are missing. In fact, after removing the few continuous survey variables and the questions with no recorded responses at all, only 256 variables are left. Of those 256, I remove the ones that probably cannot help education policy, such as self-efficacy measures, self-reported prior familiarity and experience with math and reading concepts, and nondescript "ISCED" variables. 210 factor variables remain for prediction, most of which have between 2 and 4 levels each.

2.2 Measuring student success

For each student, the PISA dataset has 5 overall reading scores and 5 overall math scores. Each score is roughly on a continuous scale from around 200 to around 800, and as seen in Figure 1, the scores are highly correlated. Standardized test scores are only rough measures of academic performance, but when properly censored, they do expose the most egregious achievement gaps. To censor the data, I

1. Compute a total score for each student by summing the 10 standardized PISA scores together.
2. Collect the students with total scores above the 75th percentile, and call them "high-performing".
3. Collect the students with total scores below the 25th percentile, and call them "low-performing".
4. Remove the rest of the students from the data.

In the context of prediction, I now have a response variable with two possible values: high and low. I will temporarily suspend my skepticism and treat this factor as the gold standard of student success.

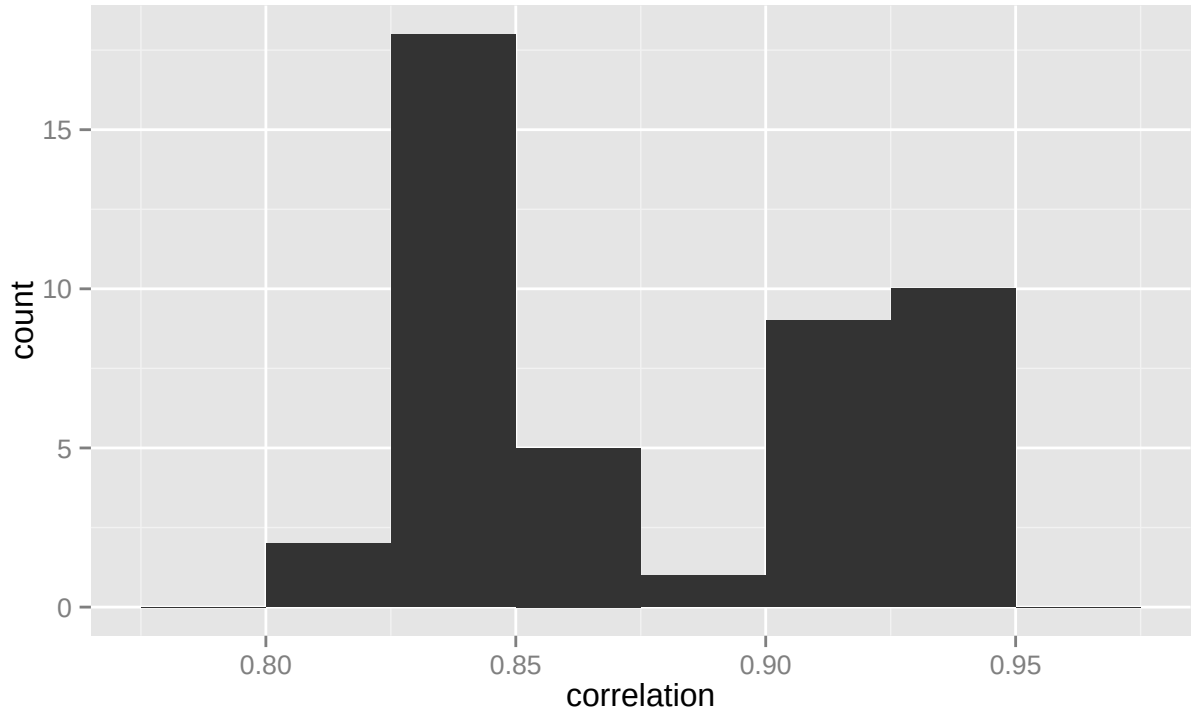

Figure 1: histogram of pairwise correlations among the original 5 reading and 5 math scores from the PISA tests. Correlations are high, so I do not lose much information in summing them up to produce a single total score for each student.

3 The best general issues for predicting success

In this section, I attempt to find the general issues that have the highest potential of distinguishing successful students from those who struggle.

3.1 Ranking individual predictor variables

To get a rough picture of the important issues, I first rank all 210 variables individually. For the rankings, I use a matching heuristic that loosely measures how well a factor can split students by success level. For each factor x, I calculate this heuristic as follows.

1. Remove the missing values from x, along with the corresponding values from the binary vector y of student performances (high and low coded as 1 and 0, respectively).
2. For every subset s of the levels of $x = (x_1, \ldots, x_n)$,
   a. Create the binary vector $z = (z_1, \ldots, z_n)$, where $z_i = I(x_i \in s)$.
   b. Let the matching score of s be

$$\frac{1}{n} \max \left\{ \sum_{i=1}^{n} I(y_i = z_i),\ \sum_{i=1}^{n} I(y_i \neq z_i) \right\}$$

3. Take the matching score of x to be the maximum of all the matching scores calculated in step 2.

One can interpret the matching heuristic as the most optimistic rate of correct classification for a prediction on a single variable. A matching of 1 means that x can predict y perfectly, and a matching of 0.5 means that x is no better than chance.

Figure 2 shows the matching heuristics of the 210 predictor variables. Most individual variables predict better than chance. The two variables with matchings better than 0.7 are "How many books at home" (censored) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late".

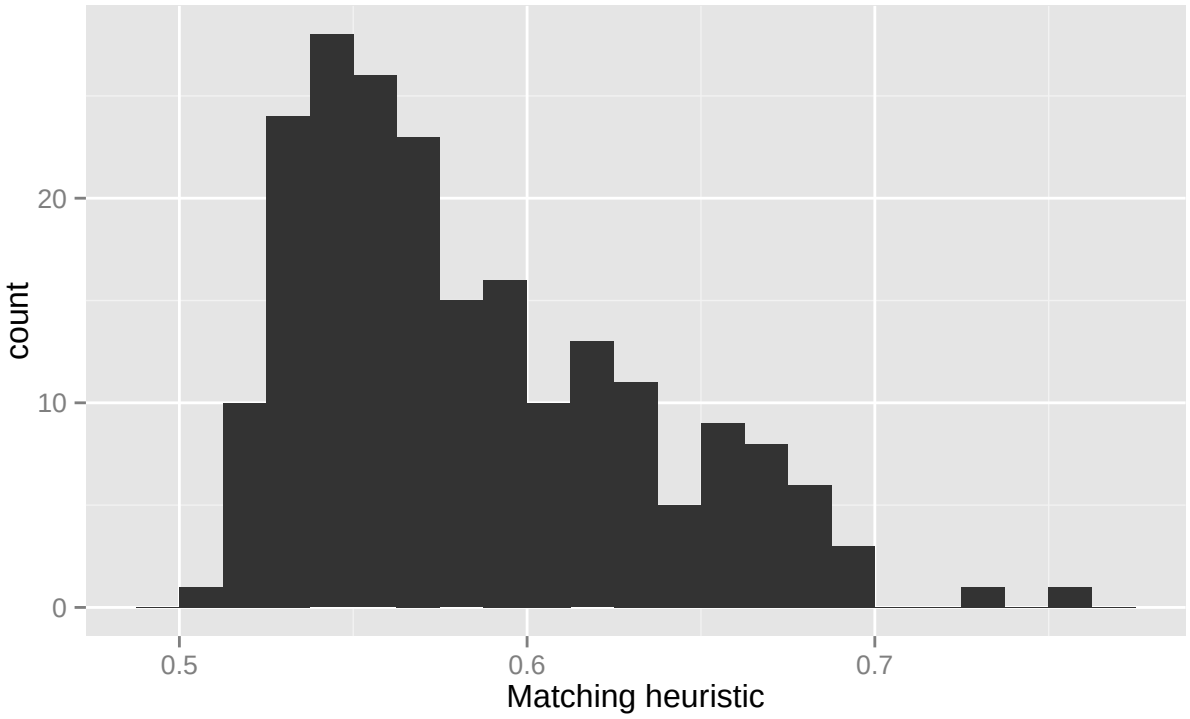

Figure 2: matching heuristics of all 210 predictor variables. Most individual variables predict better than chance. The two variables with matchings better than 0.7 are "How many books at home" (censored) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late".

Figure 3 shows the matching scores of the 210 variables, where the variables are grouped by the general issues they cover, such as possessions, attitudes, teaching, etc. The results are not definitive because the matching scores only apply to separate variables individually. However, we can start to identify potentially useful key issues in education. The three topics with multiple high matching scores are teaching, attitude/interest/motivation, and parental backgrounds. These three issues have high potential for affecting student success, and they are the ones I will continue pursuing in subsequent sections. It is important to recognize here, however, that these issues have the most variables, and their apparent importance could just be due to bias in the design of the PISA survey.

4 Focusing on the key issues

The previous section established that teaching, attitude/interest/motivation, and parental backgrounds could be important areas to investigate. Some matching scores on individual variables in these areas are high. But how important is each key issue overall? Which issues are more important than the others? How does each issue compare to the full predictive potential of the whole PISA dataset? To find out, I build a dataset on each

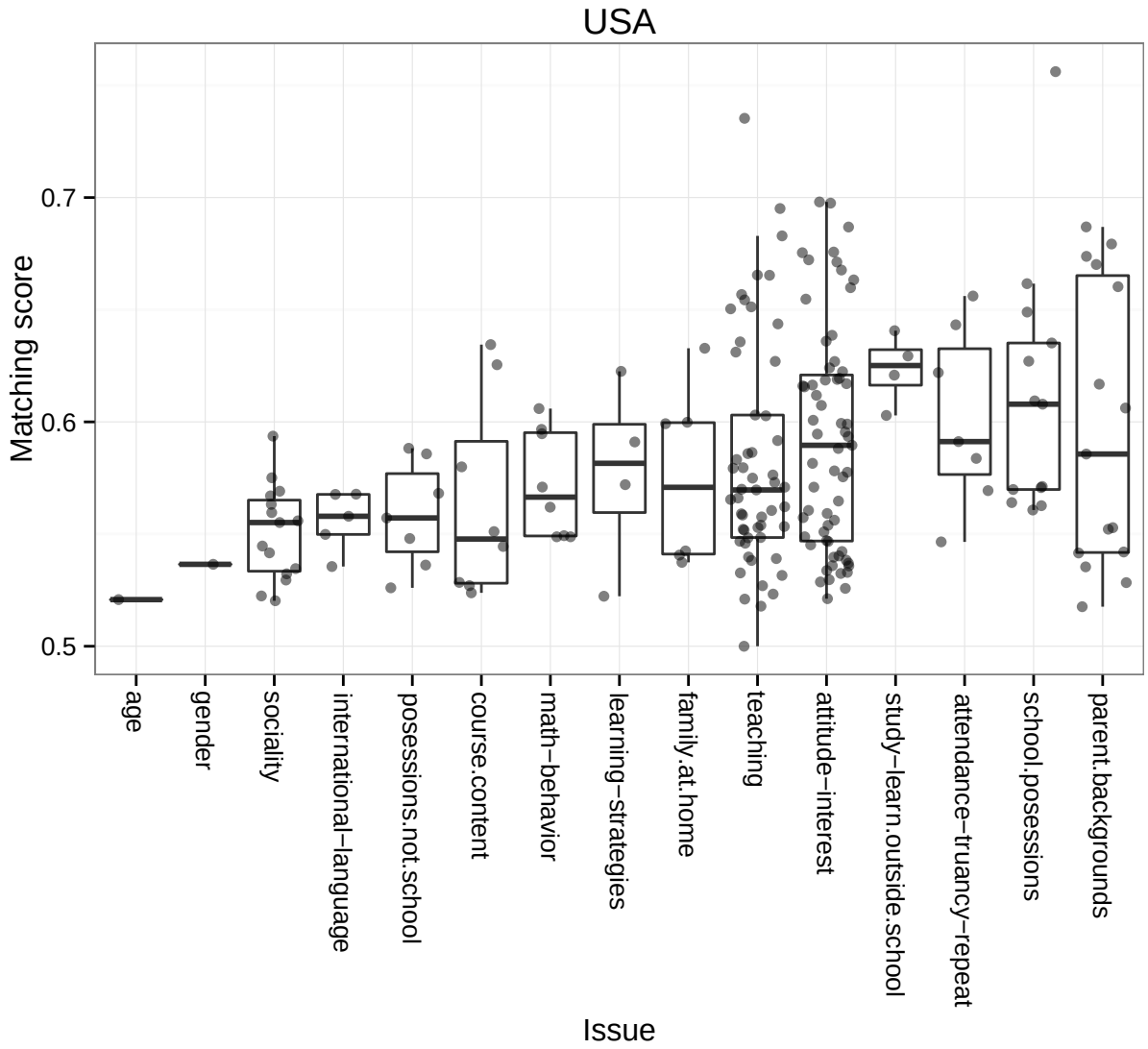

Figure 3: matching scores of the 210 variables, where the variables are grouped by the general issues they cover, such as possessions, attitudes, teaching, etc. The results are not definitive because the matching scores only apply to separate variables individually. However, we can start to identify potentially useful key issues in education. The three topics with multiple high matching scores are teaching, attitude/interest/motivation, and parental backgrounds. These three issues have high potential for affecting student success, and they are the ones I will continue pursuing in subsequent sections. It is important to recognize here, however, that these issues have the most variables, and their apparent importance could just be due to bias in the design of the PISA survey.

issue and attempt to classify students according to academic success. Below, I describe these issue-specific datasets.

4.1 Issue-specific datasets

4.1.1 Teaching

The teaching variables measure many different aspects of teaching style and quality as experienced by the students, such as the frequency of homework, quality of feedback, classroom management, the disciplinary climate of the classroom, student-teacher rapport, assessments, and calculator use. For the teaching dataset, I take the teaching variables with the top 20 matching scores, shown in Figure 4.

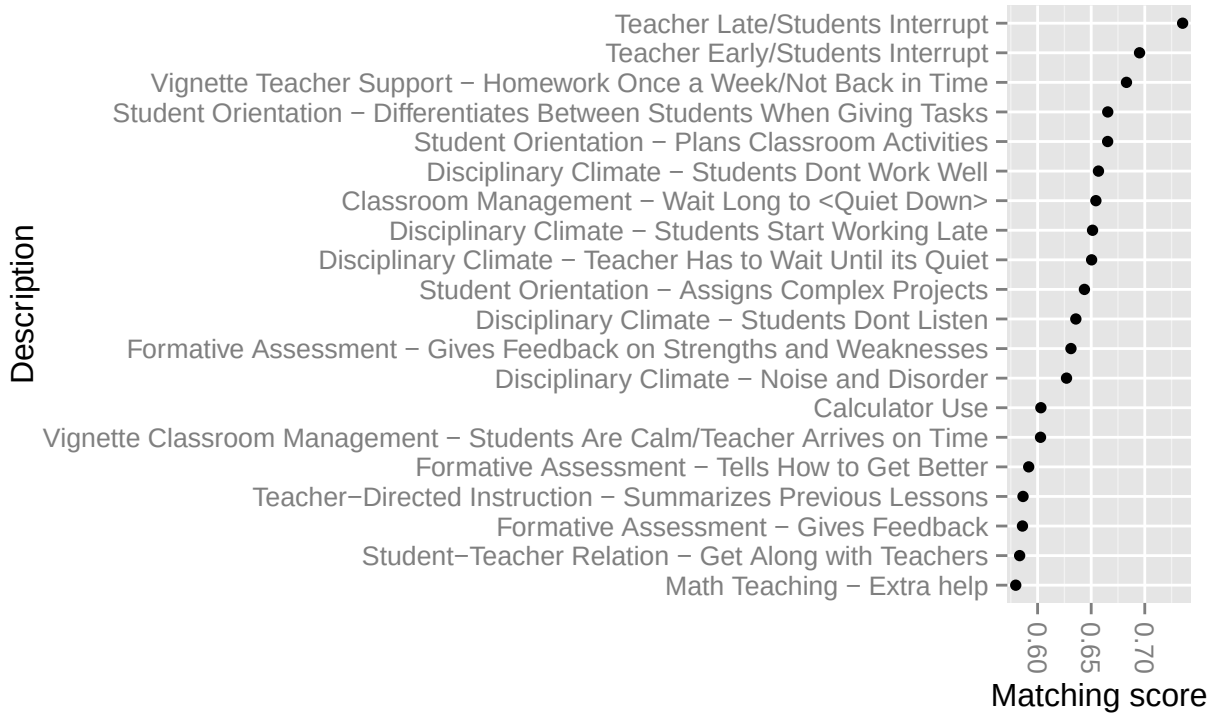

Figure 4: For the USA teaching dataset, I take the teaching variables with the top 20 matching scores, shown here.

4.1.2 Attitude/interest/motivation

These variables are student self-reported measures of perceived control, work ethic, motivation, attitude towards school, anxiety, attributions to failure, and perseverance. For the attitude/interest/motivation dataset, I take the variables in this area with the top 20 matching scores, shown in Figure 5.

4.1.3 Parental backgrounds

The parental background variables measure the educational levels, job statuses, and "ISCED qualifications" of the parents of each student. (It's a shame that PISA does not explain what these ISCED qualifications really mean. Many nondescript "ISCED" variables are poorly documented.) I use all 15 of these variables for the parental backgrounds dataset, shown in Figure 6.

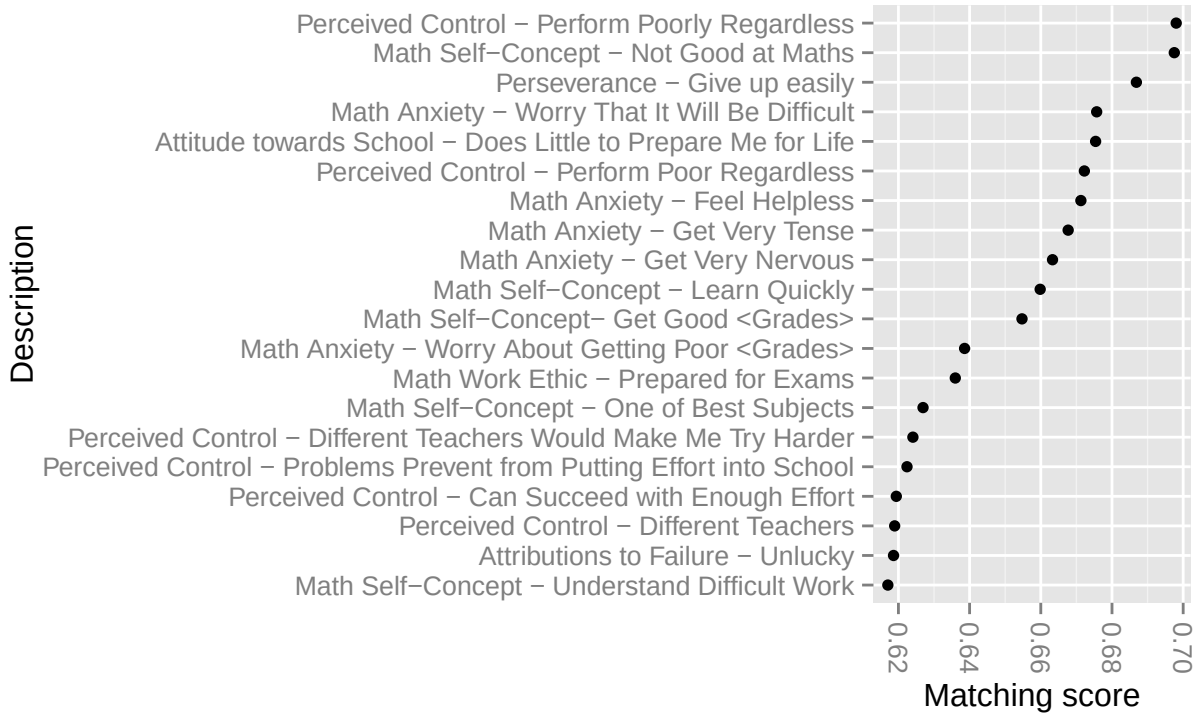

Figure 5: For the USA attitude/interest/motivation dataset, I take the variables in this area with the top 20 matching scores, shown here.

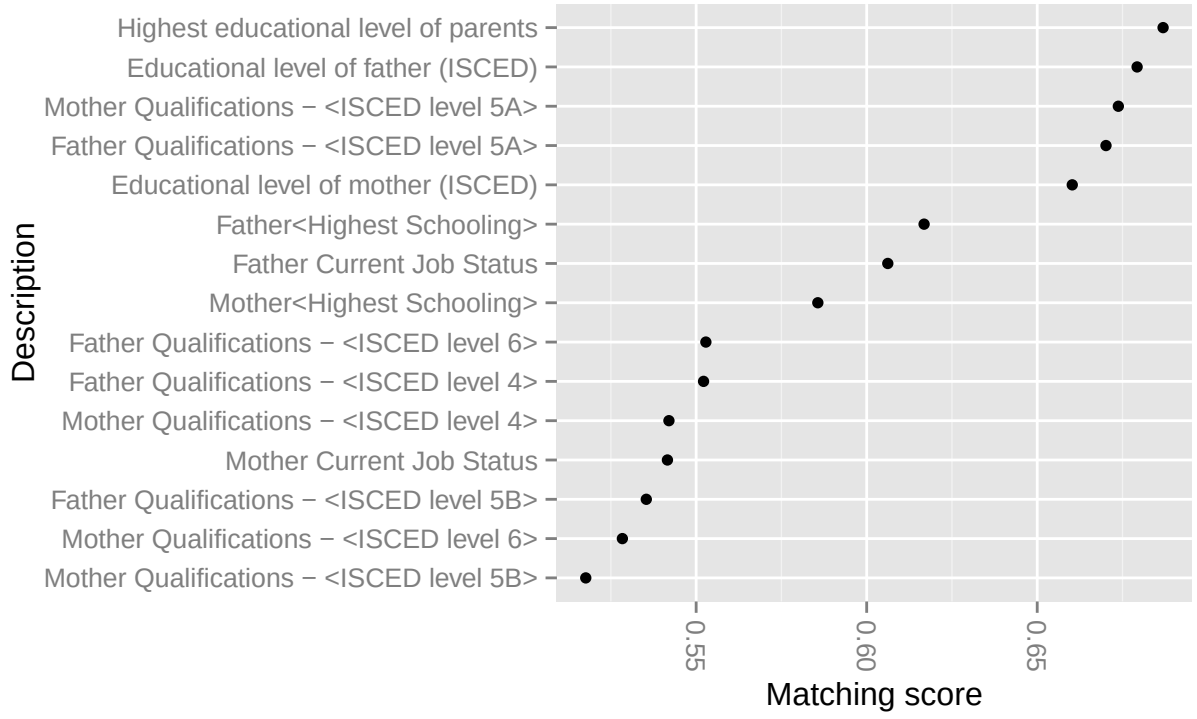

Figure 6: The parental background variables measure the educational levels, job statuses, and "ISCED qualifications" of the parents of each student. I use all 15 of these variables for the parental backgrounds dataset, shown here.

4.1.4 School-related possessions

The school-related possessions variables measure things like the number of books at home (highest matching score by far), number of computers, number of textbooks, access to internet, and access to study space. I use all 13 of these variables, shown in Figure 7.

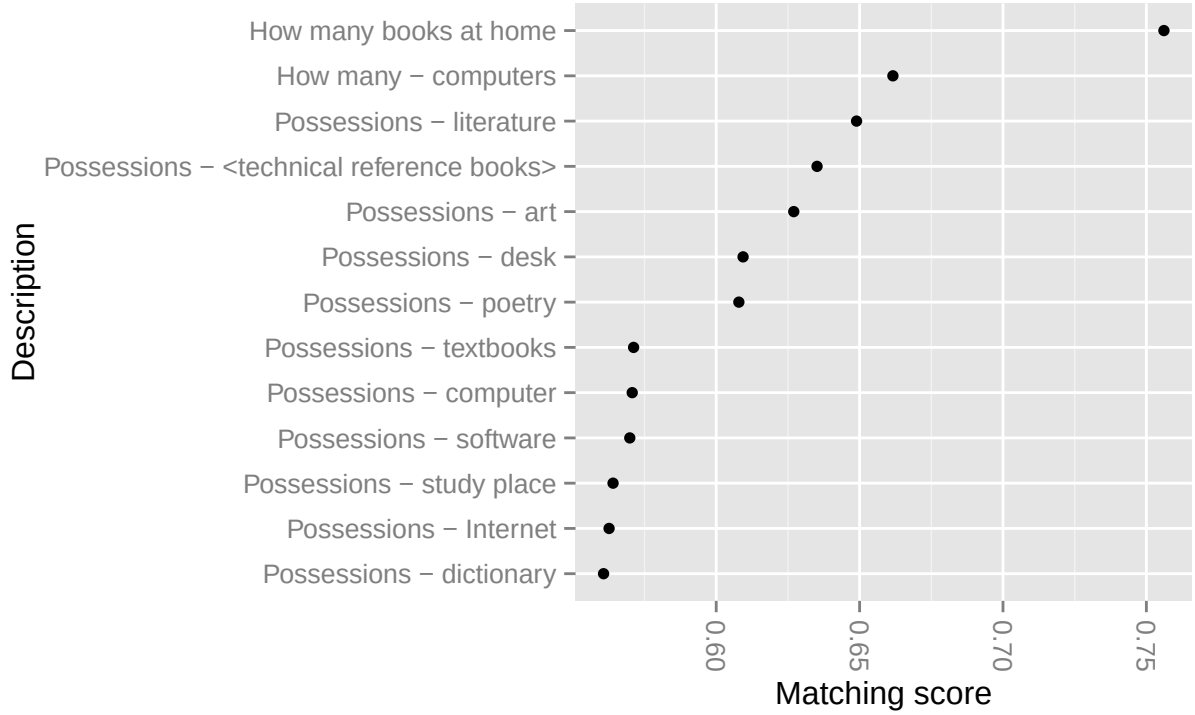

Figure 7: The school-related possessions variables measure things like the number of books at home (highest matching score by far), number of computers, number of textbooks, access to internet, and access to study space. I use all 13 of these variables, shown here.

4.1.5 Top 20 variables

For the sake of comparison, I collect the variables with the top 20 matching scores out of all the usable 210 factors from the USA PISA student dataset, shown in Figure 8.

4.2 Imputation

Each of the four datasets above has missing values, before I can classify students, I need to impute them. Here, I carry out an 10-nearest-neighbors imputation for the "top 20 variables" USA dataset. I do not show the imputation for the other 3 datasets here because these cases are similar.

Figure 9 shows that many variables have an entire third of their values missing, so the imputation process is messy. Some students have an unmanageably high number of missing values, as seen in Figure 10, and these students need to be removed. (For some other countries, variables with over 70% missing cases also needed to be removed, but there were no such variables for the USA.) I make the choice to remove the small percentage of students (less than 4%) with over 15 missing values. This cutoff of 15 has the advantage of keeping as many students as seems practical, but it may be too high: imputing too many missing values could distort the data. Hence, I check that the summary information for the imputed data is the same as the summary information of the original data. Figure 11 shows that the ranges are the same between the

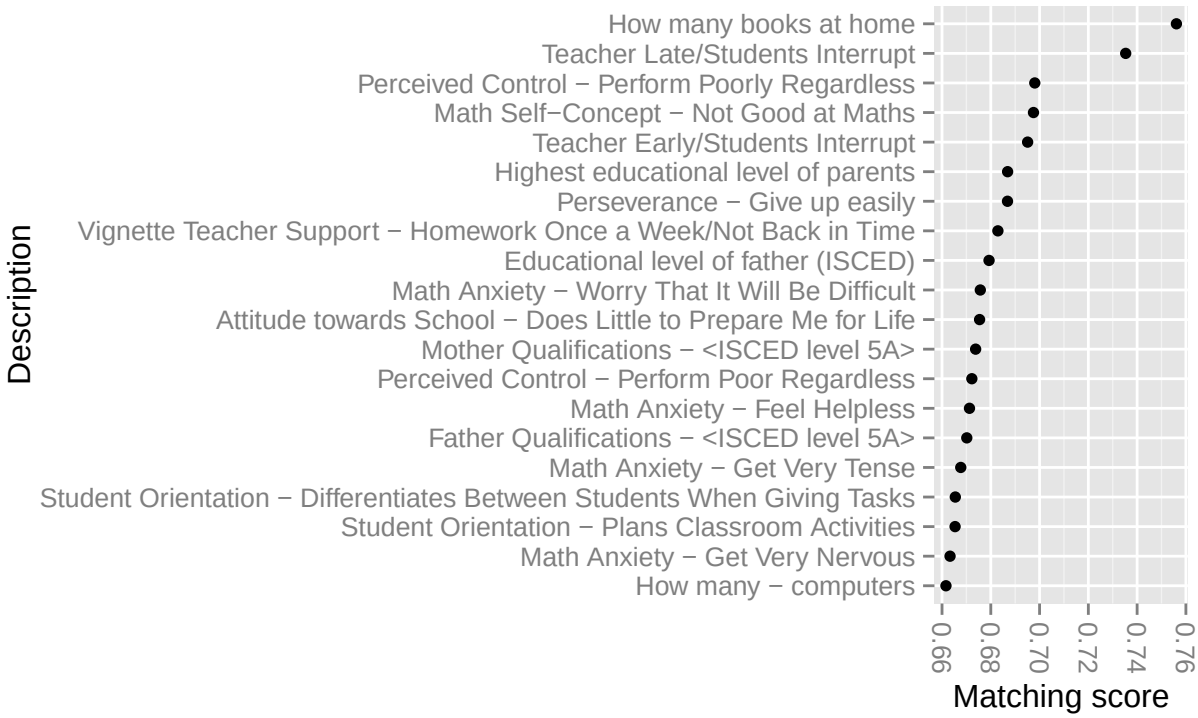

Figure 8: For the sake of comparison, I collect the variables with the top 20 matching scores out of all the usable 210 factors from the USA PISA student dataset, shown here.

imputed and non-imputed versions, and that most of the quartiles are the same. The 10-nearest-neighbors imputation was a success.

Although it would be cumbersome to show here, I imputed and checked the other three datasets analogously. For the teaching and attitude/interest/motivation datasets, I removed students with more than 15 missing values. For the parent backgrounds dataset, which had 15 predictor variables, I removed students with more than 12 missing values.

4.3 So which issue is most important?

I use each of the four datasets above to attempt to classify students according to high or low academic success. I try several different classifiers, including

- logistic regression.
- a random forest with 500 trees (default in the `randomForest` package in R).
- bagging (random forest with 500 trees and all variables used per tree).
- neural networks with 2, $m/4$, and $3m/4$ nodes in the hidden layer, where m is the number of predictor variables.
- support vector machines with linear, cubic, and radial kernels.
- linear discriminant analysis.
- quadratic discriminant analysis.
- K nearest neighbors with $K = 5$, 10, and 25.

Before applying the classifiers, I divided each dataset at random into training and test sets (75% training cases, 25% test cases). Figure 12 shows the test error rates of these classifiers, grouped by key issue, after training on the training sets. As expected, all the classifiers performed best on the dataset with the

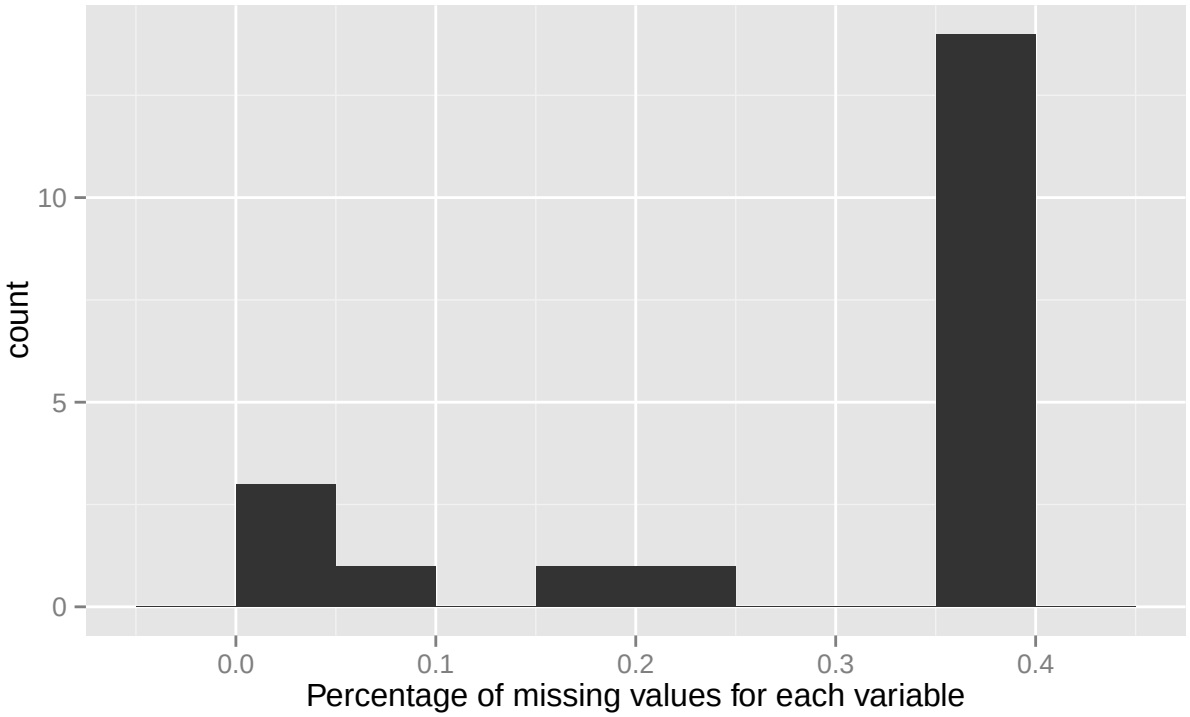

Figure 9: percentage of missing values for each variable in the "Top 20 variables" USA dataset.

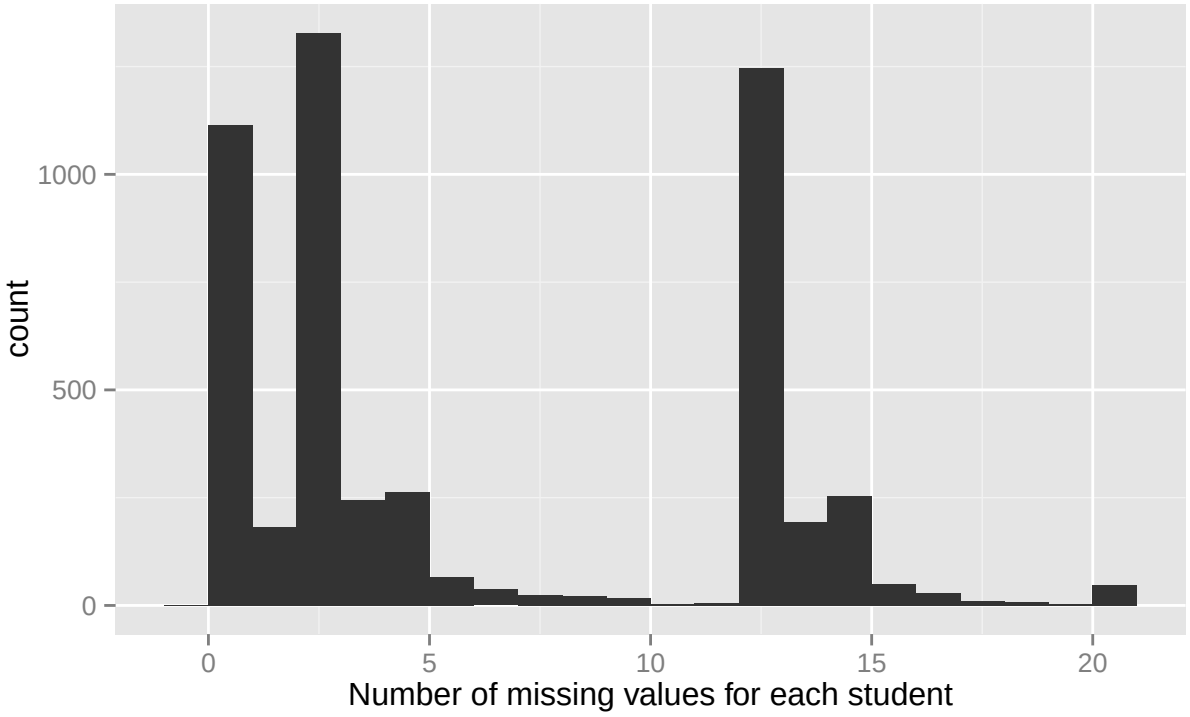

Figure 10: number of missing values per student in the "Top 20 variables" USA datset.

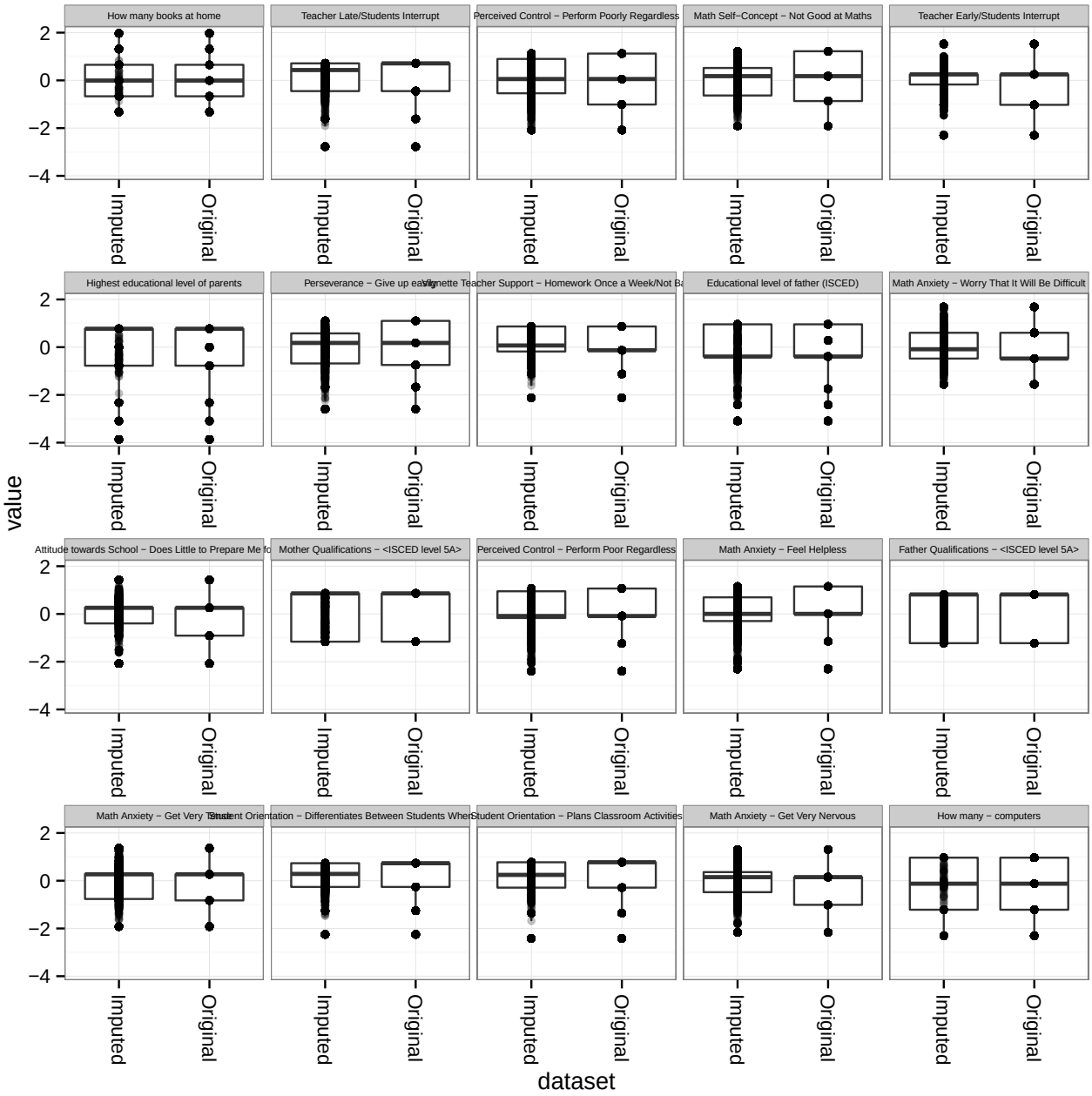

Figure 11: Here, I check that the summary information for the imputed data is the same as the summary information of the original data ("Top 20 variables" USA data). The ranges are the same between the imputed and non-imputed versions, and most of the quartiles are the same. The 10-nearest-neighbors imputation was a success.

variables with the top 20 matching scores. After that, the teaching dataset was the most useful, then the attitude/interest/motivation dataset. The parental backgrounds dataset is the least useful.

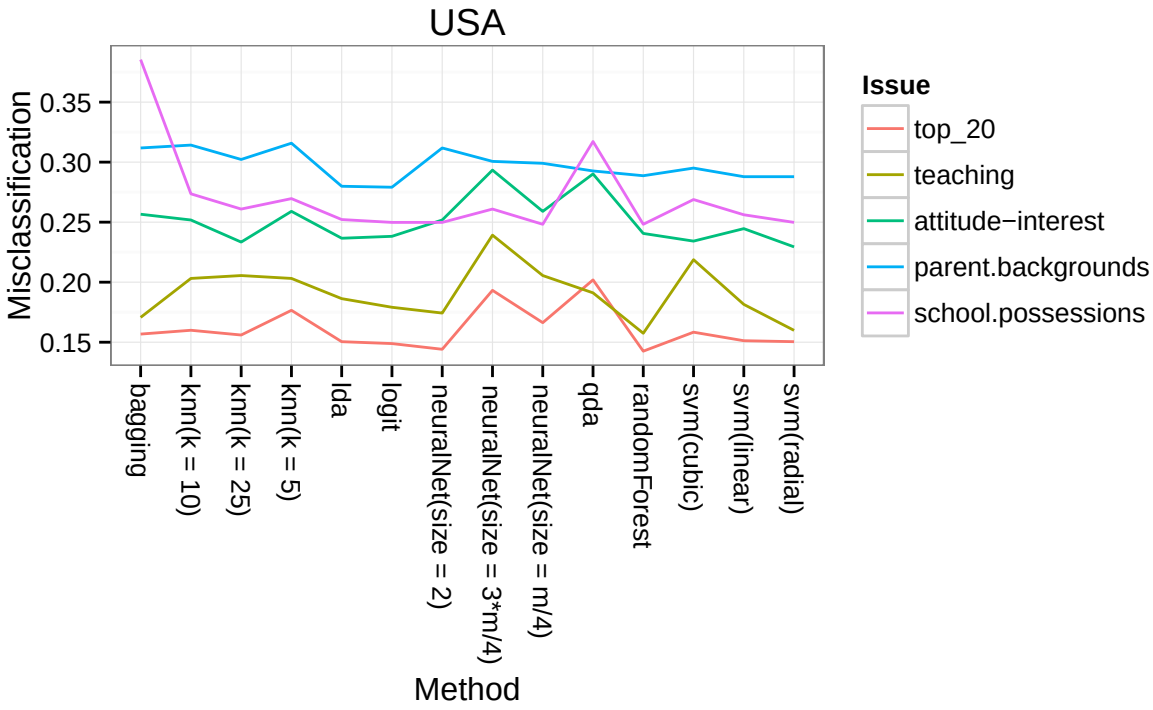

Figure 12: for the USA, test error rates of these classifiers, grouped by key issue, after training on the training sets. As expected, all the classifiers performed best on the dataset with the variables with the top 20 matching scores. After that, the teaching dataset was the most useful, then the attitude/interest/motivation dataset. The parental backgrounds dataset is the least useful.

It is tempting to conclude that the quality and style of teaching is more important to academic success than student attitudes and that parental influences matter even less. However, these findings could easily be artifacts of the PISA survey design. From Figure 3, teaching and attitude/motivation/interest had far more variables than any other issue, including the matter of parental backgrounds. Similarly, some survey questions may be more effective than others because some topics in education research may be more advanced than others. Alternatively, OECD employees may have simply wanted to focus on the issues they already thought were most salient. A confirmation bias is usually inevitable, but it does overemphasize the importance of hot-button issues and preclude the discovery new ones.

5 Other countries

6 Acknowledgements

I would like to thank Dr. Cook for steering me in the right direction. The PISA data is messy and cumbersome, and the guidance is very appreciated.

References

"Organization for Economic Co-operation and Development." 2015. http://www.oecd.org/.