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. I use this process to look for the most decisive issues in determining student success, first for the USA, and then for other countries. The goal is to find the next logical steps to improve education policy.

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.

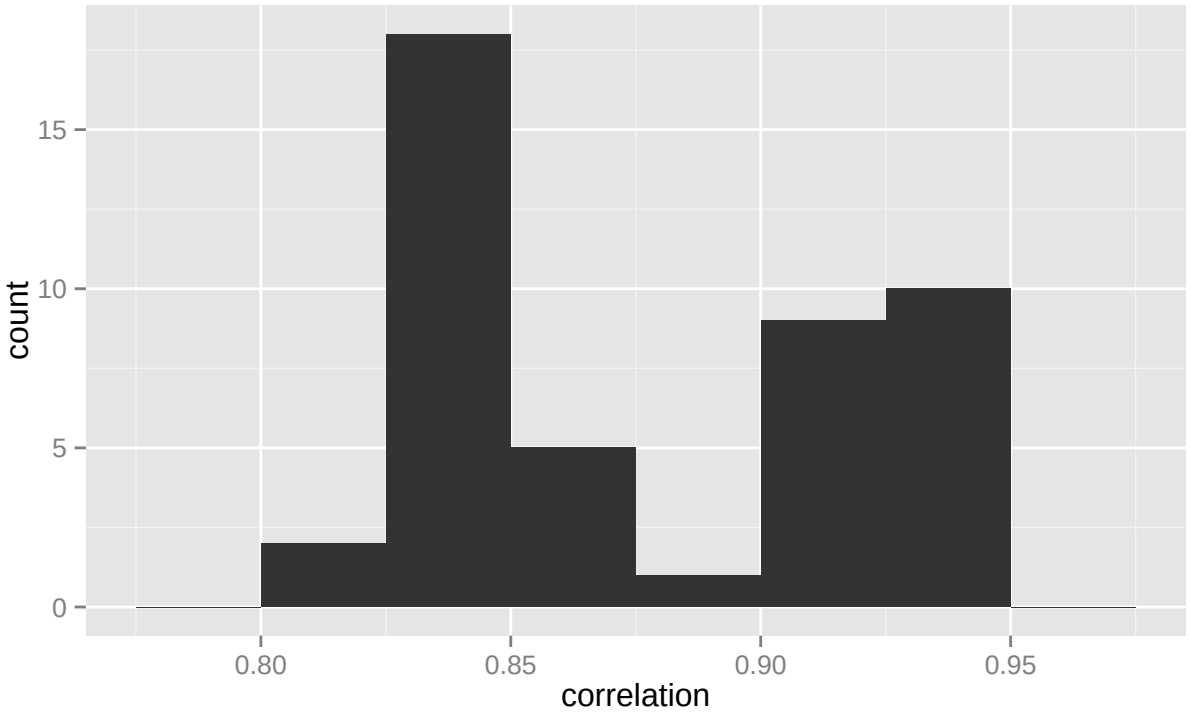

Figure 1: histogram of pairwise correlations among the original 5 reading and 5 math scores from the PISA tests. Correlations are high, so I do not lose much information in summing them up to produce a single total score for each student.

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".

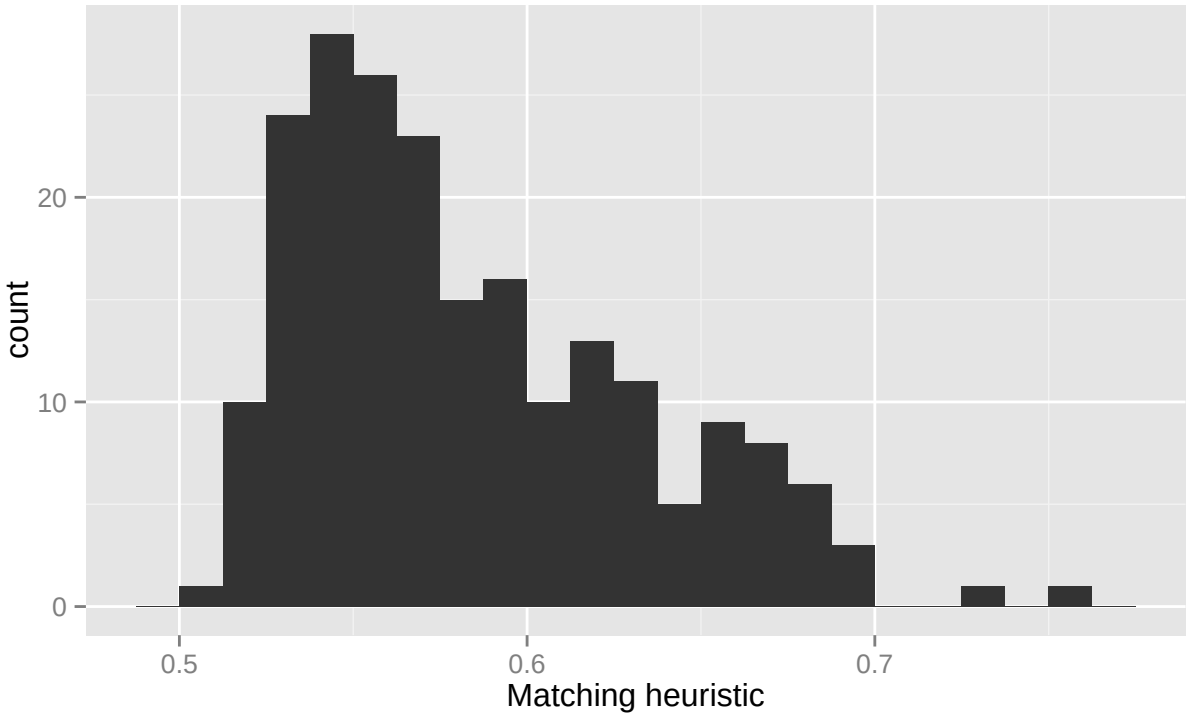

Figure 2: matching heuristics of all 210 predictor variables. Most individual variables predict better than chance. The two variables with matchings better than 0.7 are "How many books at home" (censored) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late".

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 topics with multiple high matching scores are teaching, attitude/interest/motivation, and parental backgrounds. The "number of books at home" variable has the highest matching score of any variable, so school-related possessions may be important as well. These four issues have high potential for affecting student success, and they are the ones I will continue pursuing in subsequent sections.

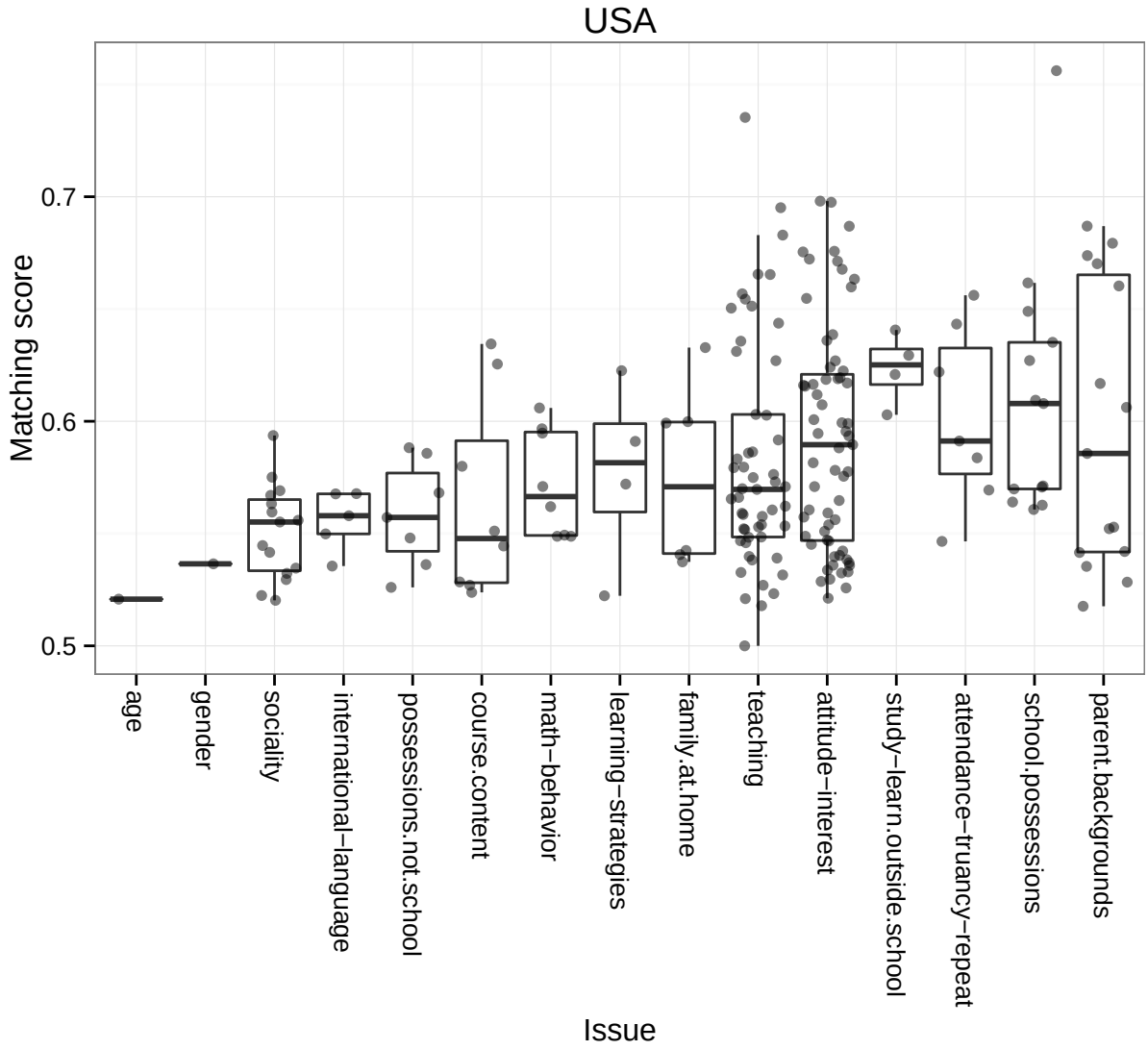

Figure 3: matching scores of the 210 variables, where the variables are grouped by the general issues they cover, such as possessions, attitudes, teaching, etc. The results are not definitive because the matching scores only apply to separate variables individually. However, we can start to identify potentially useful key issues in education. The topics with multiple high matching scores are teaching, attitude/interest/motivation, and parental backgrounds. The "number of books at home" variable has the highest matching score of any variable, so school-related possessions may be important as well. These four issues have high potential for affecting student success, and they are the ones I will continue pursuing in subsequent sections.

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 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.

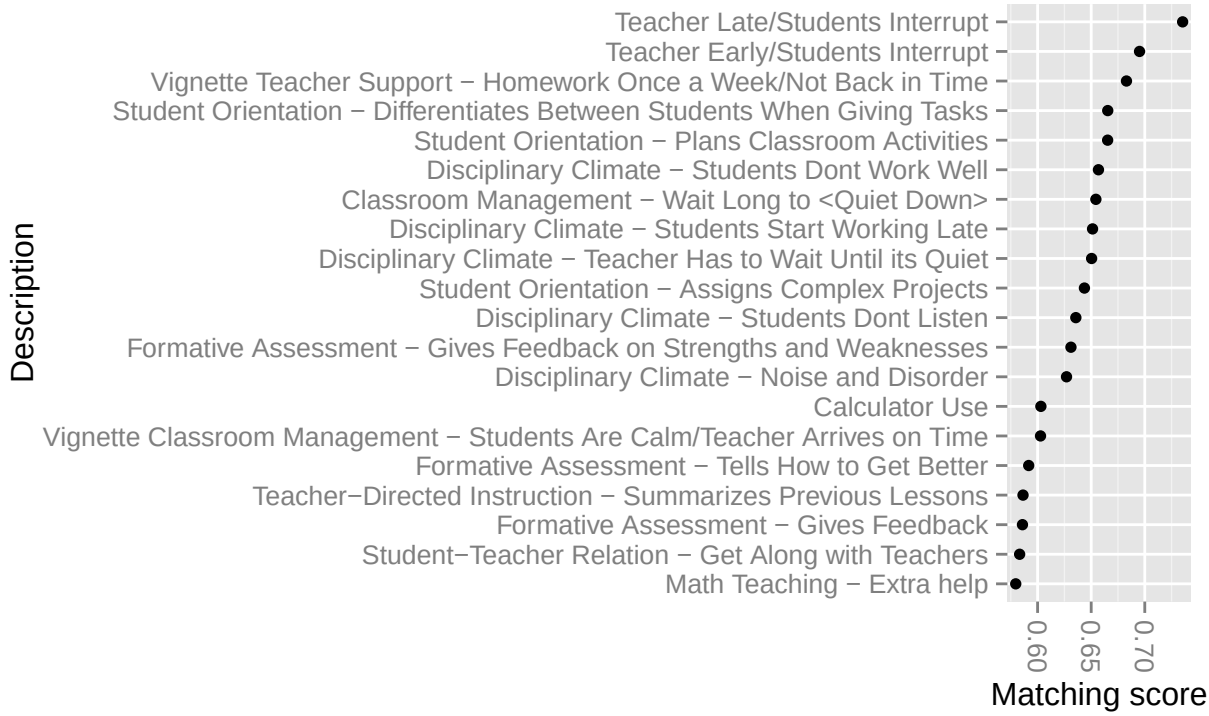

Figure 4: For the USA teaching dataset, I take the teaching variables with the top 20 matching scores, shown here.

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.

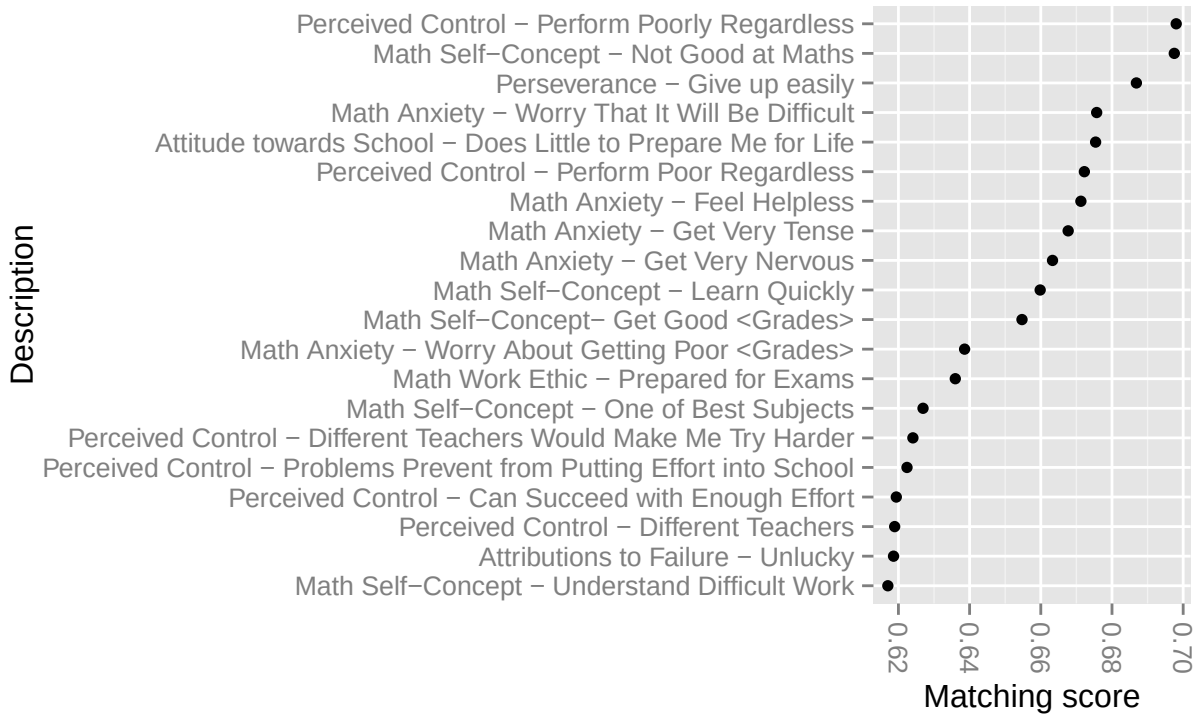

Figure 5: For the USA attitude/interest/motivation dataset, I take the variables in this area with the top 20 matching scores, shown here.

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.

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.

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 (`DMwR` R package (Torgo 2010)). 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

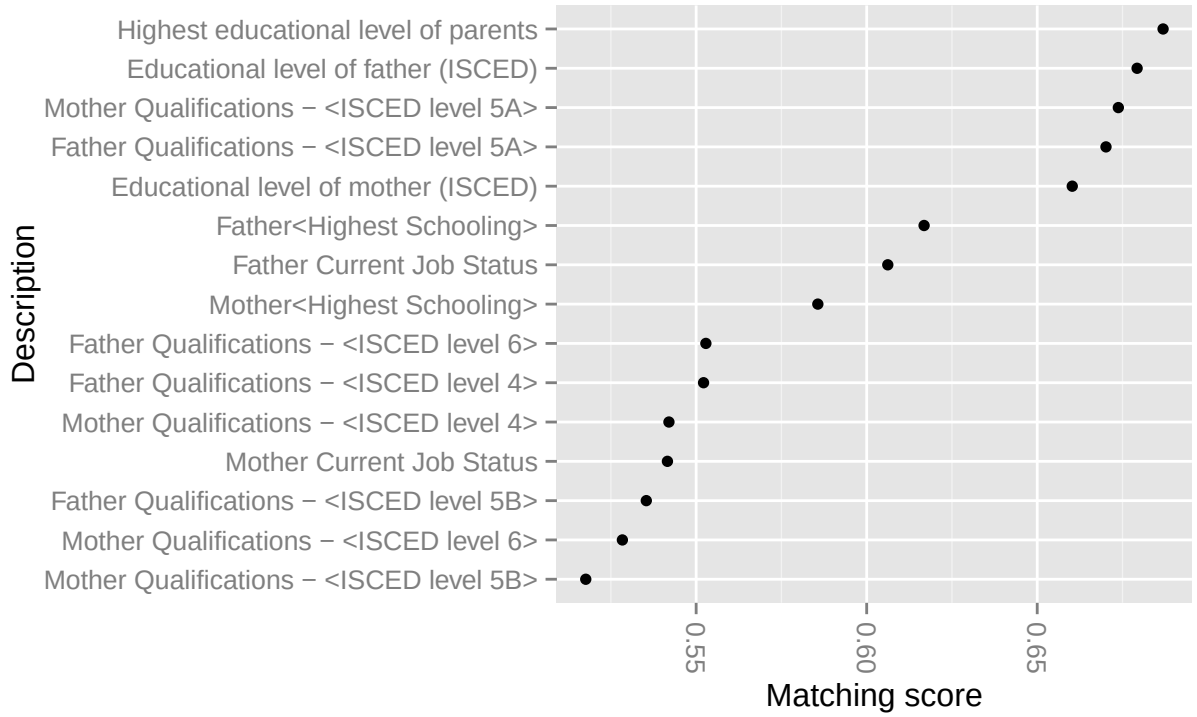

Figure 6: The parental background variables measure the educational levels, job statuses, and "ISCED qualifications" of the parents of each student. I use all 15 of these variables for the parental backgrounds dataset, shown here.

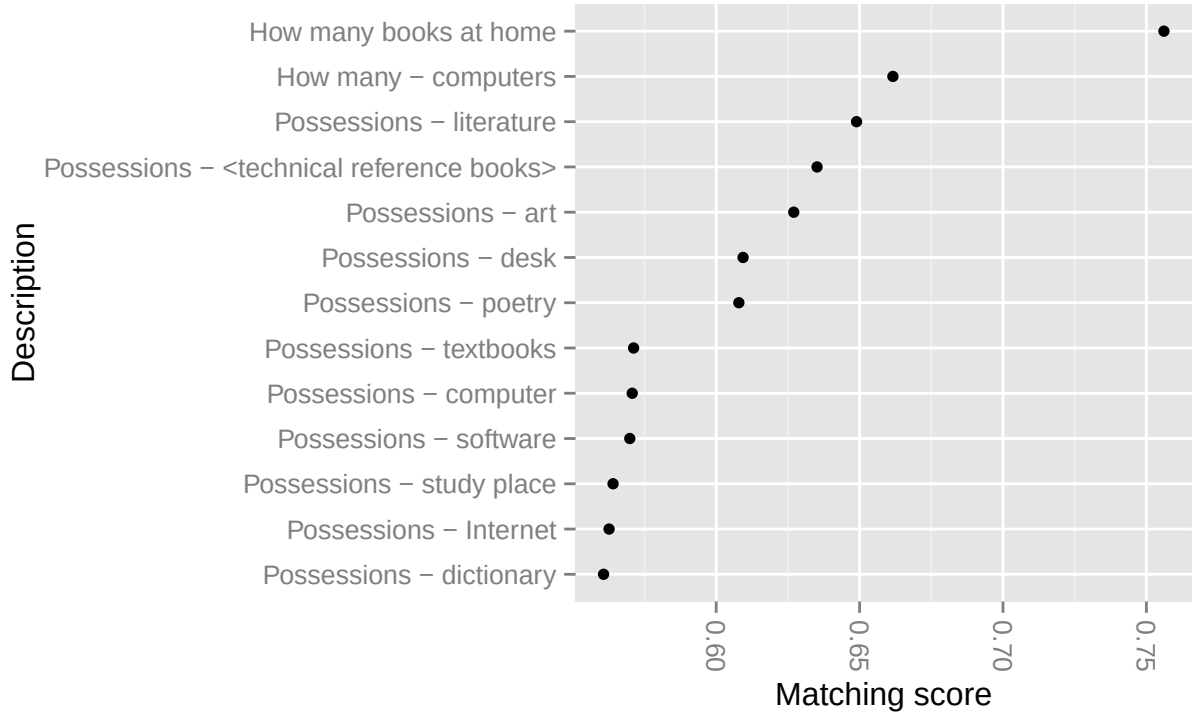

Figure 7: The school-related possessions variables measure things like the number of books at home (highest matching score by far), number of computers, number of textbooks, access to internet, and access to study space. I use all 13 of these variables, shown here.

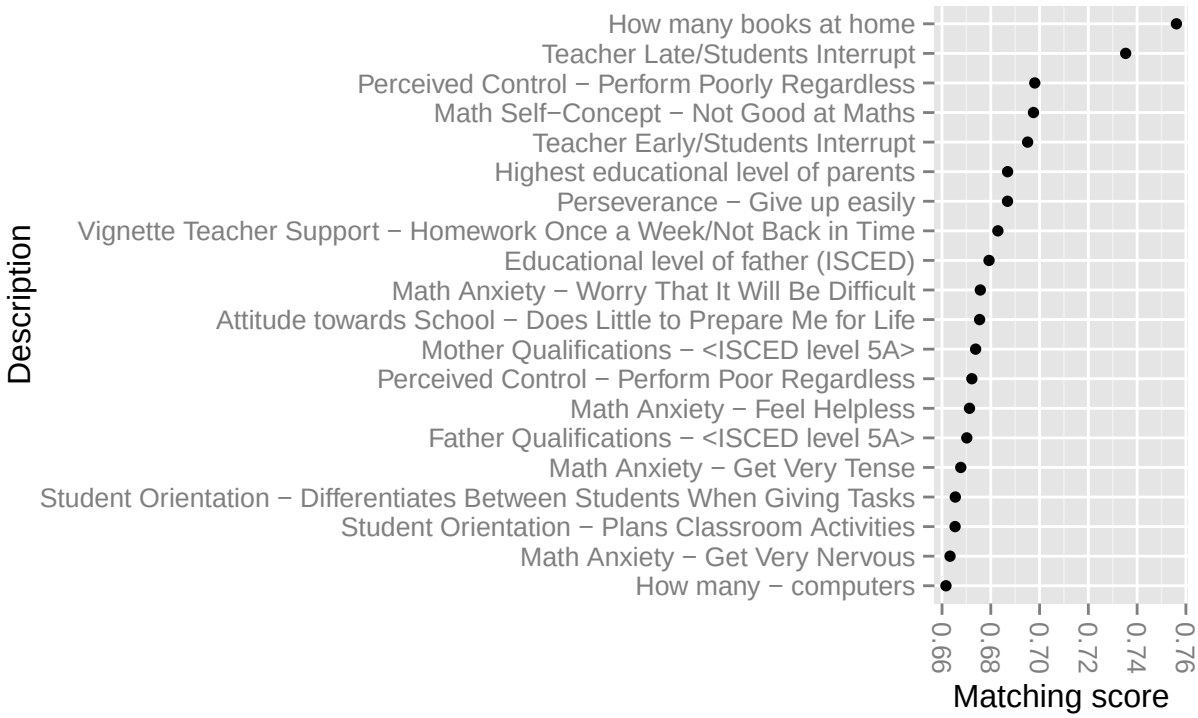

Figure 8: For the sake of comparison, I collect the variables with the top 20 matching scores out of all the usable 210 factors from the USA PISA student dataset, shown here.

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 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. For the school-related possessions dataset, I used all 13 predictor variables and removed students with more than 9 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 (`glm` function in core R (R Core Team 2014)).
- a random forest with 500 trees (`randomForest` R package (Liaw and Wiener 2002)).
- neural networks with 2, $m/4$, and $3m/4$ nodes in the hidden layer, where m is the number of predictor variables (`nnet` R package (Venables and Ripley 2002a)).
- support vector machines with linear, cubic, and radial kernels (`e1071` R package (Meyer et al. 2014)).
- linear discriminant analysis (`MASS` package in R (Venables and Ripley 2002b)).

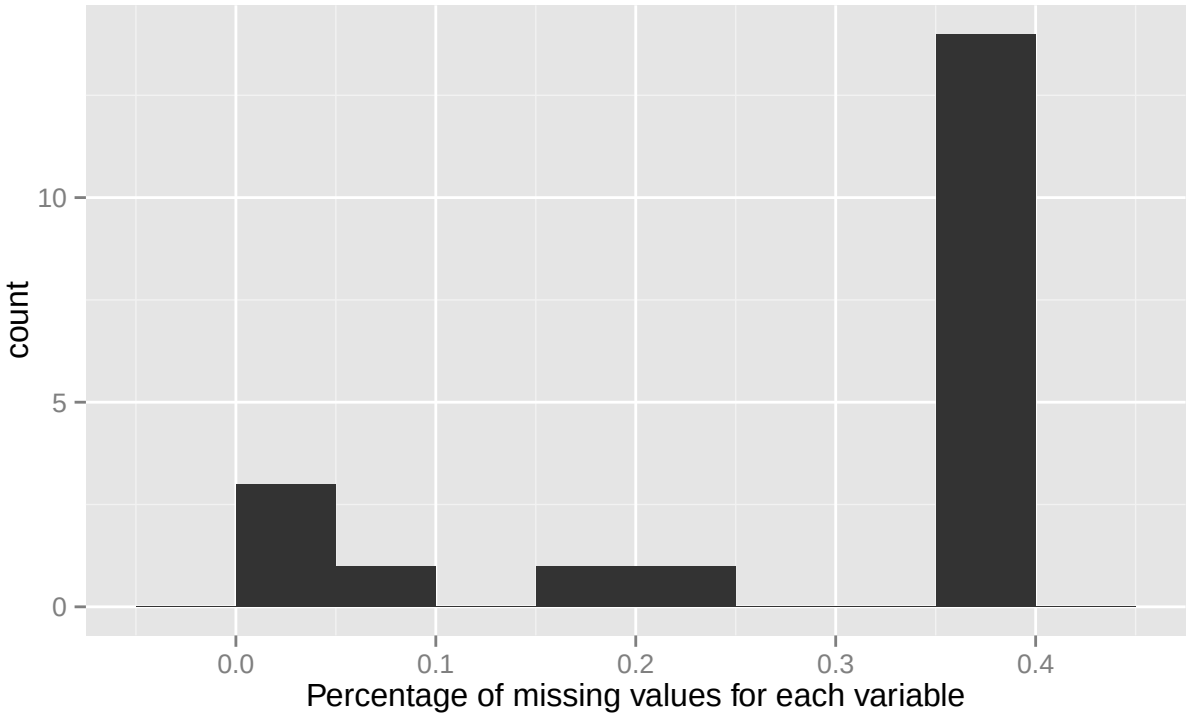

Figure 9: percentage of missing values for each variable in the "Top 20 variables" USA dataset.

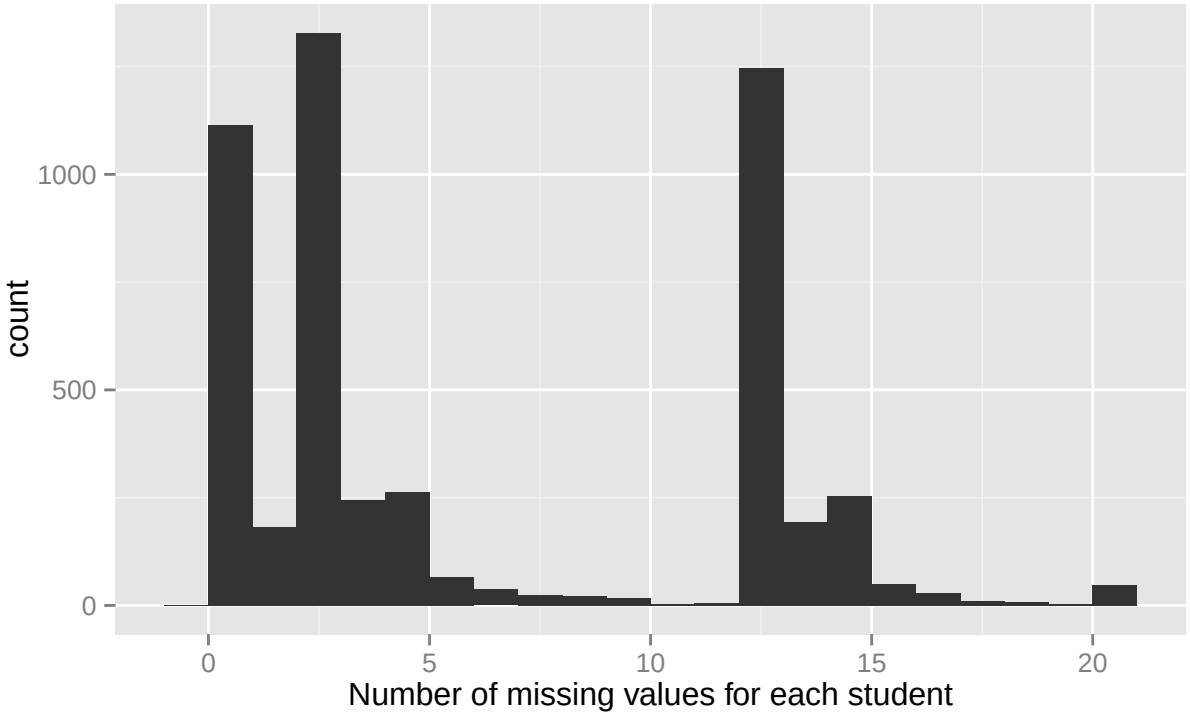

Figure 10: number of missing values per student in the "Top 20 variables" USA datset.

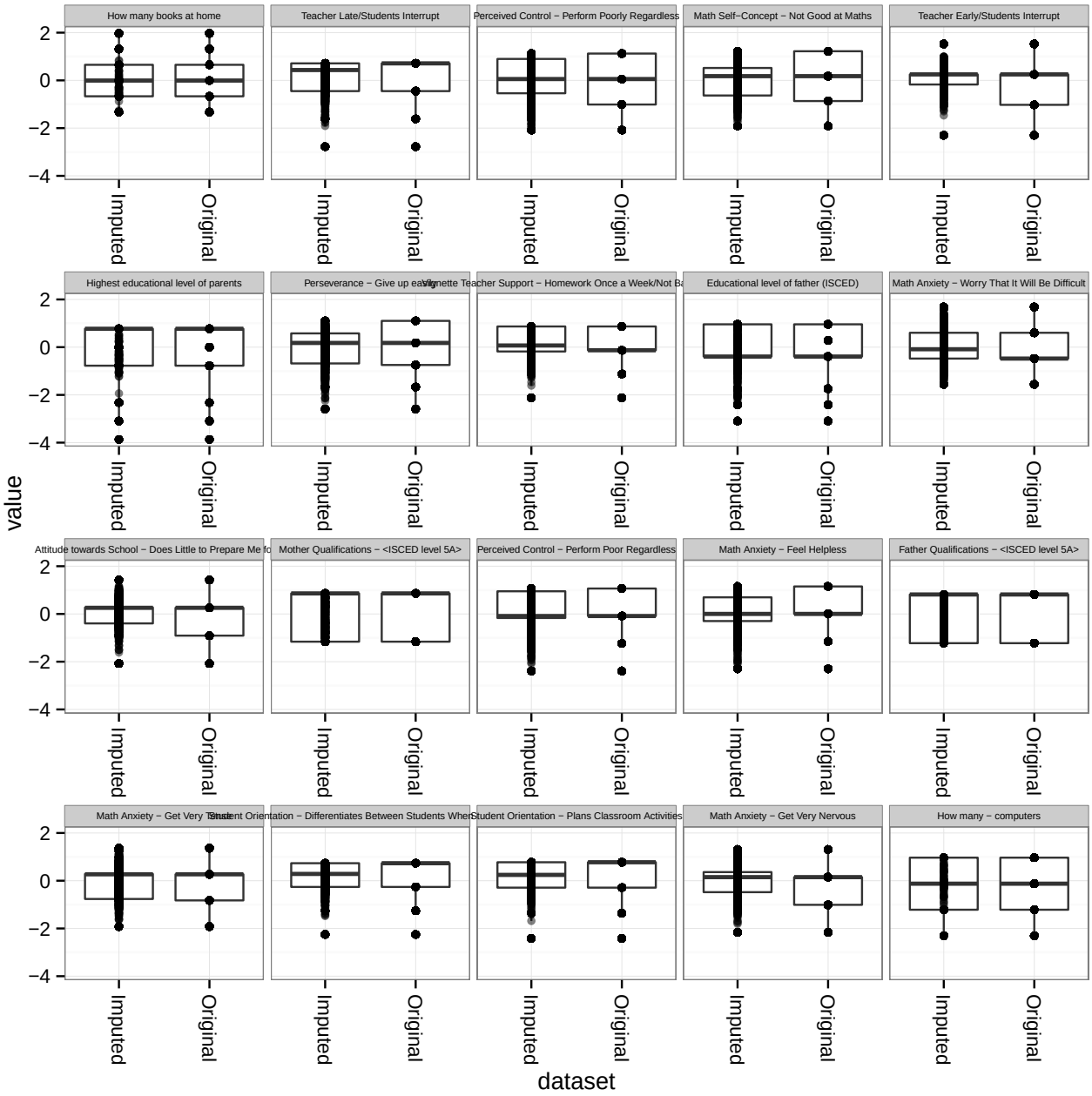

Figure 11: Here, I check that the summary information for the imputed data is the same as the summary information of the original data ("Top 20 variables" USA data). The ranges are the same between the imputed and non-imputed versions, and most of the quartiles are the same. The 10-nearest-neighbors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

- quadratic discriminant analysis (`MASS` package in R (Venables and Ripley 2002b)).
- K nearest neighbors with $K = 5$, 10, and 25 (`class` package in R (Venables and Ripley 2002c)).

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

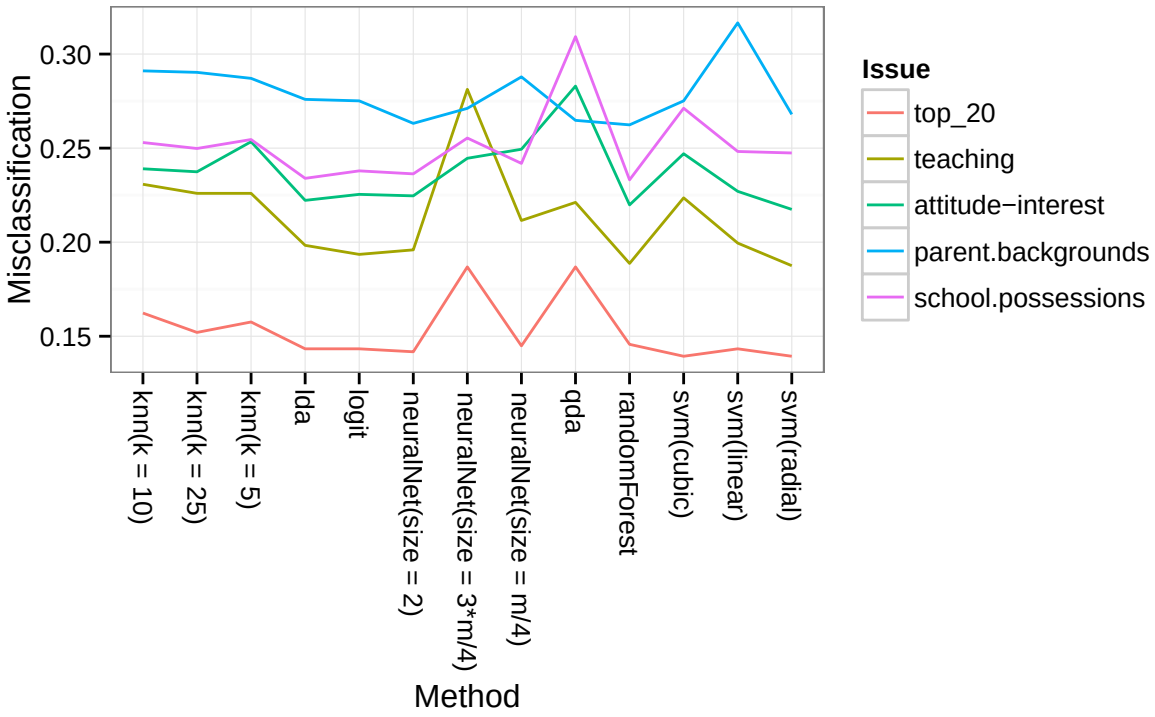

Figure 12: for the USA, test error rates of these classifiers, grouped by key issue, after training on the training sets. As expected, all the classifiers performed best on the dataset with the variables with the top 20 matching scores. After that, the teaching dataset was the most useful, then the attitude/interest/motivation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

It is tempting to conclude that the quality and style of teaching is more important to academic success than student attitudes and possessions, and that parental influences matter even less. However, these findings could easily be artifacts of the PISA survey design, both in the number and quality of the questions targeting each issue. For example, Dana Goldstein argues in The Teacher Wars: A History of America's Most Embattled Profession that in the United States, the public has historically placed unreasonably high levels of expectation and blame on teachers, making the issue of teacher quality even more central than otherwise in education policy debates (Dana Goldstein 2014). This backdrop may have influenced OECD employees to write more abundant, more thoughtful, and more research-backed questions about teaching than most of the other categories.

5 Other countries

How does the USA compare to other countries? Do different factors determine success in other parts of the world? For comparison, I look at the PISA results from Japan, Germany, and Peru. Japan and the USA are

frequently compared in American education news stories, Germany has a conveniently high response rate in the PISA survey, and Peru's education system famously lags behind those of other countries (with the possible exception of private schools in Lima).

5.1 Finding the key issues

Figure 13 is the analogue of Figure 2, but for the USA, Japan, Germany, and Peru all together. Matching scores of individual variables are grouped by key issue and plotted. In general, teaching and attitude/interest/motivation seem to lose some of their relative importance in the other three countries, but parental backgrounds remain key. For Japan, study habits and extracurricular academic activities ("study-learn.outside.school") seem are relatively more decisive than elsewhere. Interestingly, for Germany, the "international-language" category (languages spoken at home, home countries of the parents, etc.) is extremely important, as well as attendance and truancy. For Peru, school possessions have particular importance, along with attendance and truancy. And although I do not include it for further analysis, the sociality category (sense of belonging at school, social norms about success in math, etc.) is more important for Peru than for other countries relative to the other issues.

5.2 Classifying students

Similarly to the USA data, I build a dataset for each key issue and country and use the datasets to attempt to predict student success. For each issue, I select the 20 variables with the top matching scores (or all of the variables if there are less than 20). I remove students with more missing values than 75% of the number of variables, and then I remove the variables that originally had more than 70% missing values. As before, I use 10-nearest-neighbors imputation to impute missing values, and although I do not show them here, I use figures similar to Figure 11 to verify that all imputations are successful. After imputation, I train multiple classifiers on 75% of the data and then test on the remaining 25%.

Figure 14 also show the misclassification rates for each country, faceted to compare the relative importance of key issues across countries. This plot may be full of artifacts of the inevitably different in-the-field implementations of the PISA survey, or it reveal key differences among the education systems of the different countries. Below, I profile each country according to the figure.

5.2.1 The United States

Teaching quality is more decisive here than in other countries. Attitude/interest/motivation, school-related possessions, and parental backgrounds are important, but less decisive relative to other countries. Study habits outside school, course content, attendance/truancy, and international backgrounds are not plotted because they were not decisive enough to be selected as potential classifiers of students. These findings may validate the US news media's obsessive attention to teaching quality. I do not claim that teaching is lacking overall. Rather, because they predict student success more accurately than other issues, it may be the most appropriate next issue to tackle to improve US education policy.

5.2.2 Japan

For each issue plotted, students were more difficult to classify in Japan than in other countries. It may be that teaching quality, possessions, enthusiasm, parental backgrounds, and truancy are less of a problem for Japan than in the other countries. At least, these factors are less decisive in determining student success, and the path to further innovating the Japanese education system does not lie in the usual places. The next breakthrough in Japan may be more surprising and interesting than in other countries.

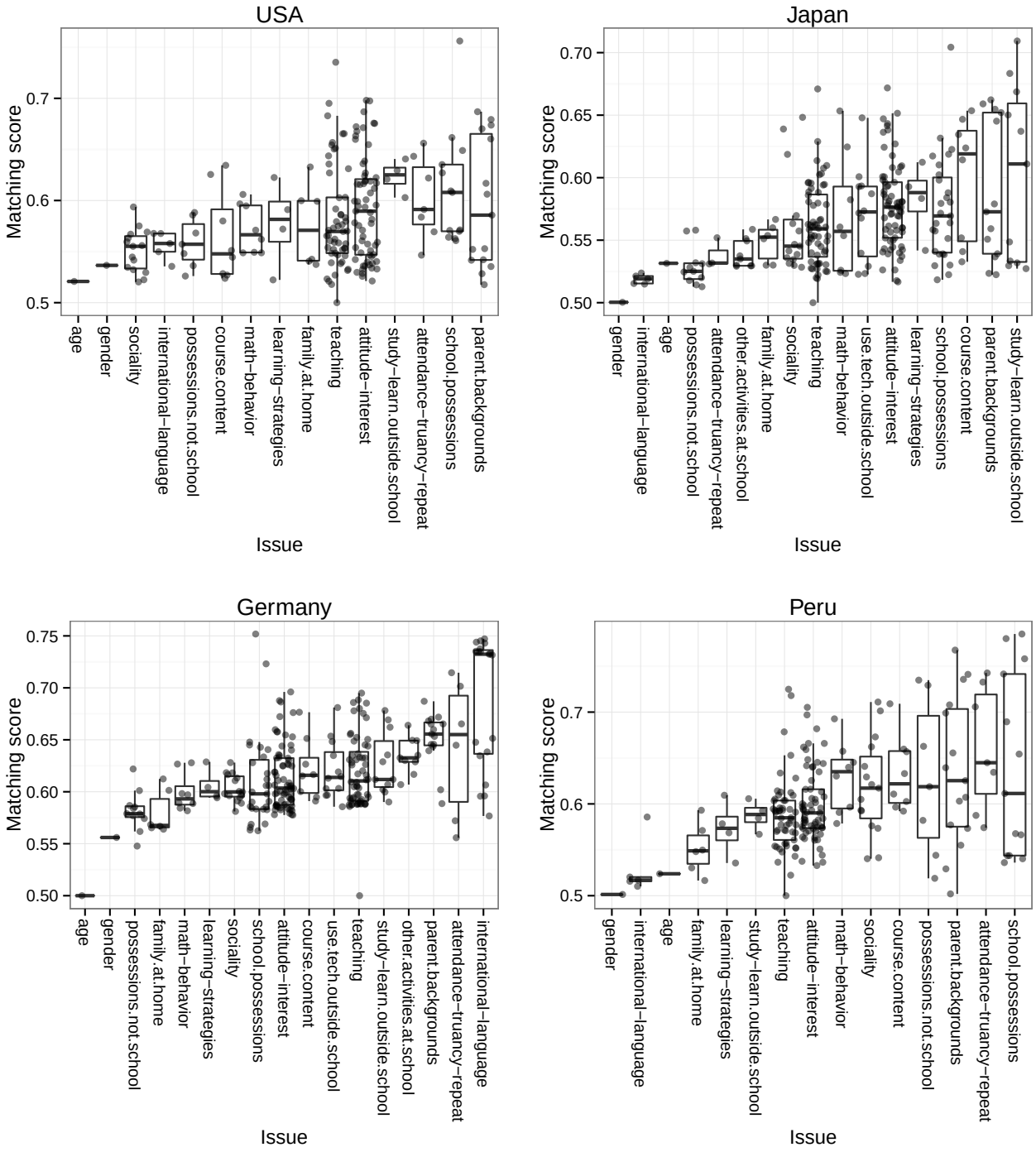

Figure 13: analogue of Figure 2, but for the USA, Japan, Germany, and Peru all together. Matching scores of individual variables are grouped by key issue and plotted.

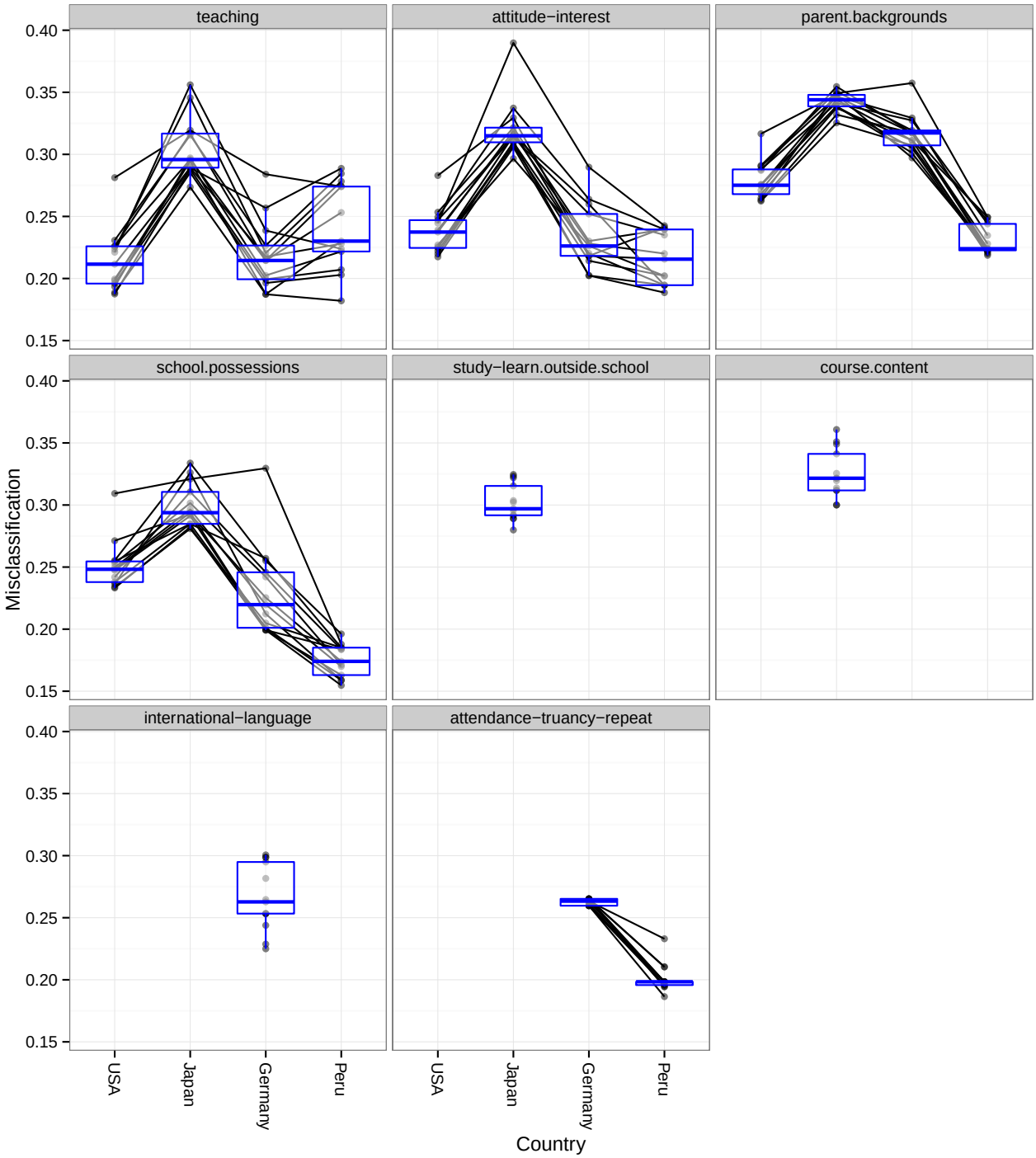

Figure 14: misclassification rates for each country, faceted to compare the relative importance of key issues across countries.

5.2.3 Germany

Germany is very much like the USA in that teaching quality is the most decisive factor plotted here relative to other countries, and that attitude/interest/motivation, school-related possessions, and parental backgrounds are important, but less decisive relative to other countries. However, a key difference between the USA and Germany is that attendance/truancy is decisive enough to be plotted. This could mean that truancy is more of a problem in Germany, or it could be that Germany executed the PISA program more thoroughly and had higher response rates.

5.2.4 Peru

Teaching quality was not as decisive in the USA, but still is important. Attitude, parental backgrounds, possessions, and attendance were all more decisive in Peru than in the other three countries. Wealth inequality is a limiting factor here, and part of the solution may be to supply schools and aid in curriculum design on a massive scale. Not much can be done about parental backgrounds except maybe adult education, which is far less common than the education of children and teens, but mentorship programs could help put students on different academic and professional tracks than their parents.

6 Conclusion

The PISA dataset is messy, but there's a story here. PISA surveyors focused most of their questions on the quality of teaching and the attitudes of students, both of which were at least somewhat important for the success of students in The United States, Japan, Germany, and Peru. In The United States, teaching quality was the most decisive factor in determining student success, more decisive in the USA than in the other three countries. In Germany, the deciding factors are similar, Peru's deciding factors are more material and familial than elsewhere, and Japan's students are difficult to classify based on the usual suspects. The key issues targeted above could very well be relative strengths or weaknesses as reflected by the misclassification rates above. The wealth inequality obstacle in Peru is obvious both in the data and in real life, and the teaching quality debate rages in America. Alternatively, the findings could all be artifacts of the PISA design, products of the biases and preferences of OECD employees, or reflections of the differences in response rates among different countries. In the most optimistic interpretation, these results begin to point to the next logical steps to improve education policy.

7 Acknowledgements

I would like to thank Dr. Cook for steering me in the right direction. The PISA data is cumbersome, and the guidance is very appreciated. In addition to core R (R Core Team 2014), I used the packages `class` (Venables and Ripley 2002c), `DMwR` (Torgo 2010), `e1071` (Meyer et al. 2014), `gdata` (Warnes et al. 2014), `ggplot2` (Wickham 2009), `gridExtra` (Auguie 2012), `knitr` (Yihui Xie 2014), `MASS` (Venables and Ripley 2002b), `nnet` (Venables and Ripley 2002a), `plyr` (Wickham 2011), `randomForest` (Liaw and Wiener 2002), and `reshape2` (Wickham 2007).

References

Auguie, Baptiste. 2012. *GridExtra: Functions in Grid Graphics*. http://CRAN.R-project.org/package=gridExtra.

Dana Goldstein. 2014. *The Teacher Wars: A History of America's Most Embattled Profession*. Doubleday.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by RandomForest." *R News* 2 (3): 18–22. http://CRAN.R-project.org/doc/Rnews/.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2014. *E1071: Misc Functions of the Department of Statistics (E1071), TU Wien*. http://CRAN.R-project.org/package=e1071.

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

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Torgo, L. 2010. *Data Mining with R, Learning with Case Studies*. Chapman; Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Venables, W. N., and B. D. Ripley. 2002a. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

———. 2002b. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

———. 2002c. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Warnes, Gregory R., Ben Bolker, Gregor Gorjanc, Gabor Grothendieck, Ales Korosec, Thomas Lumley, Don MacQueen, Arni Magnusson, Jim Rogers, and others. 2014. *Gdata: Various R Programming Tools for Data Manipulation*. http://CRAN.R-project.org/package=gdata.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York. http://had.co.nz/ggplot2/book.

———. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

Yihui Xie. 2014. *knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Friedrich Leisch Victoria Stodden and Roger D. Peng. Chapman and Hall/CRC.