What distinguishes successful students?

Will Landau

April 25, 2015

Contents

1 Introduction **2**

2 A first cleanup: getting ready to explore **2**

2.1 Variables for predicting success 2

2.2 Measuring student success 2

3 The best general issues for predicting success **3**

3.1 Ranking individual predictor variables 3

4 Focusing on the key issues **4**

4.1 Issue-specific datasets 6

4.2 Imputation 7

4.3 So which issue is most important? 10

5 Other countries **11**

5.1 Finding the key issues 11

5.2 Classifying students 13

6 Conclusion **16**

7 Acknowledgements **16**

References **17**

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 an extensive background questionnaire about their study habits, attitudes towards school, circumstances at home, etc., all 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 try 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. And because overall pedagogy and the survey's implementation are different among different countries, only students from the United States will be examined until the "Other Countries" section.

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 are poorly documented or nearly redundant with test scores, 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 temporarily suspend my skepticism and treat this factor as the gold standard of student success.

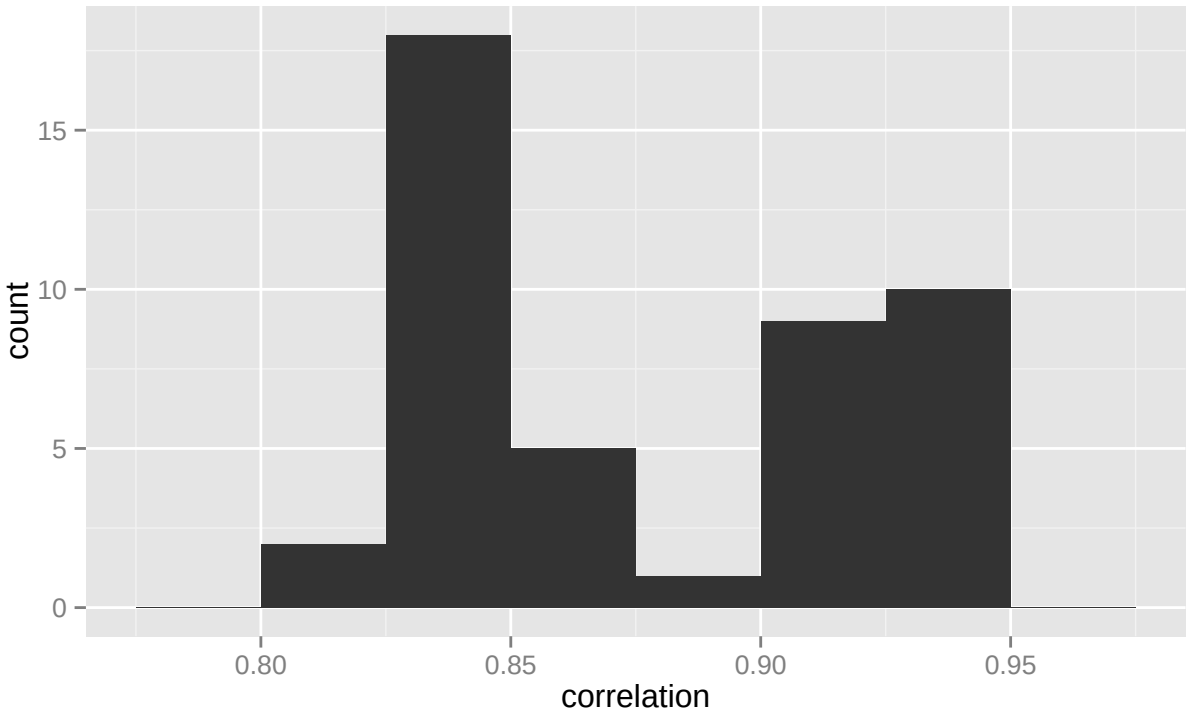

Figure 1: histogram of pairwise correlations among the original 5 reading and 5 math scores from the PISA tests. Correlations are high, so I do not lose much information in summing them up to produce a single total score for each student.

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 $x = (x_1, \ldots, x_n)$, 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,
   a. Create the binary vector $z = (z_1, \ldots, z_n)$, where $z_i = I(x_i \in s)$.
   b. Let

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

3. Let the matching score of x be $\max_s m_s$.

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 into a few levels) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives Late".

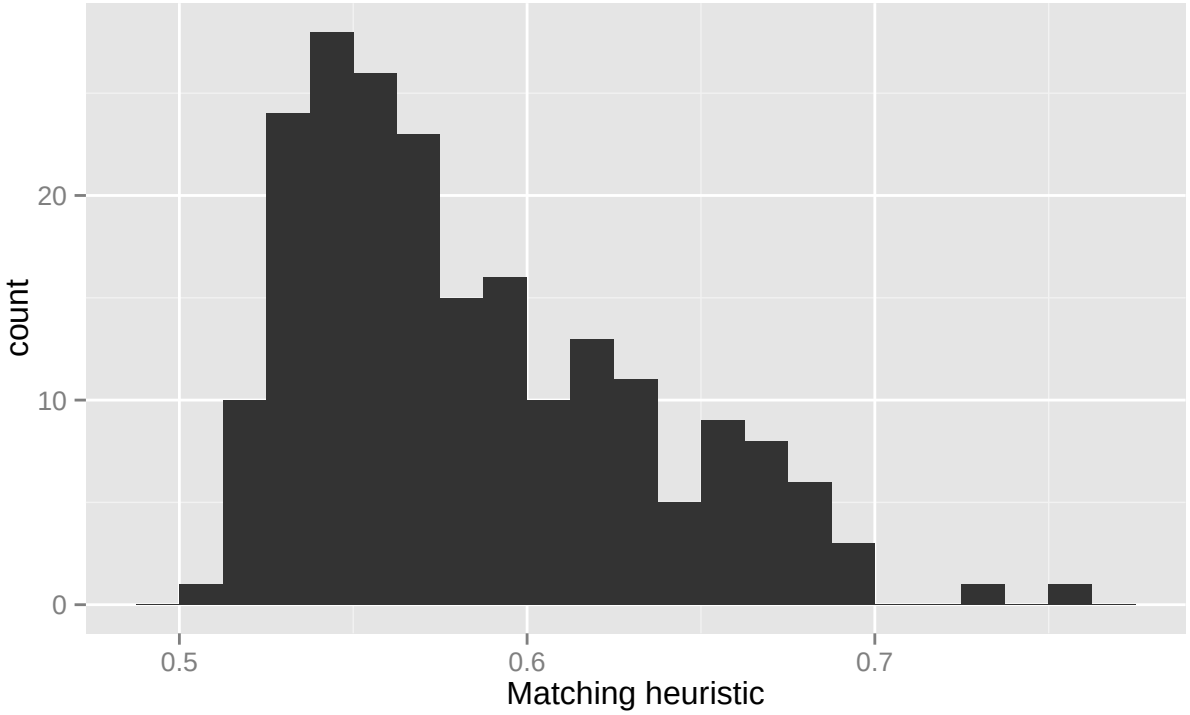

Figure 2: matching heuristics of all 210 predictor variables. Most individual variables predict better than chance. The two variables with matchings better than 0.7 are "How many books at home" (censored into a few levels) and "Vignette Classroom Management - Students Frequently Interrupt/Teacher Arrives 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. Notable 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, parental backgrounds, and school-related possessions could be important areas to investigate further. 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.

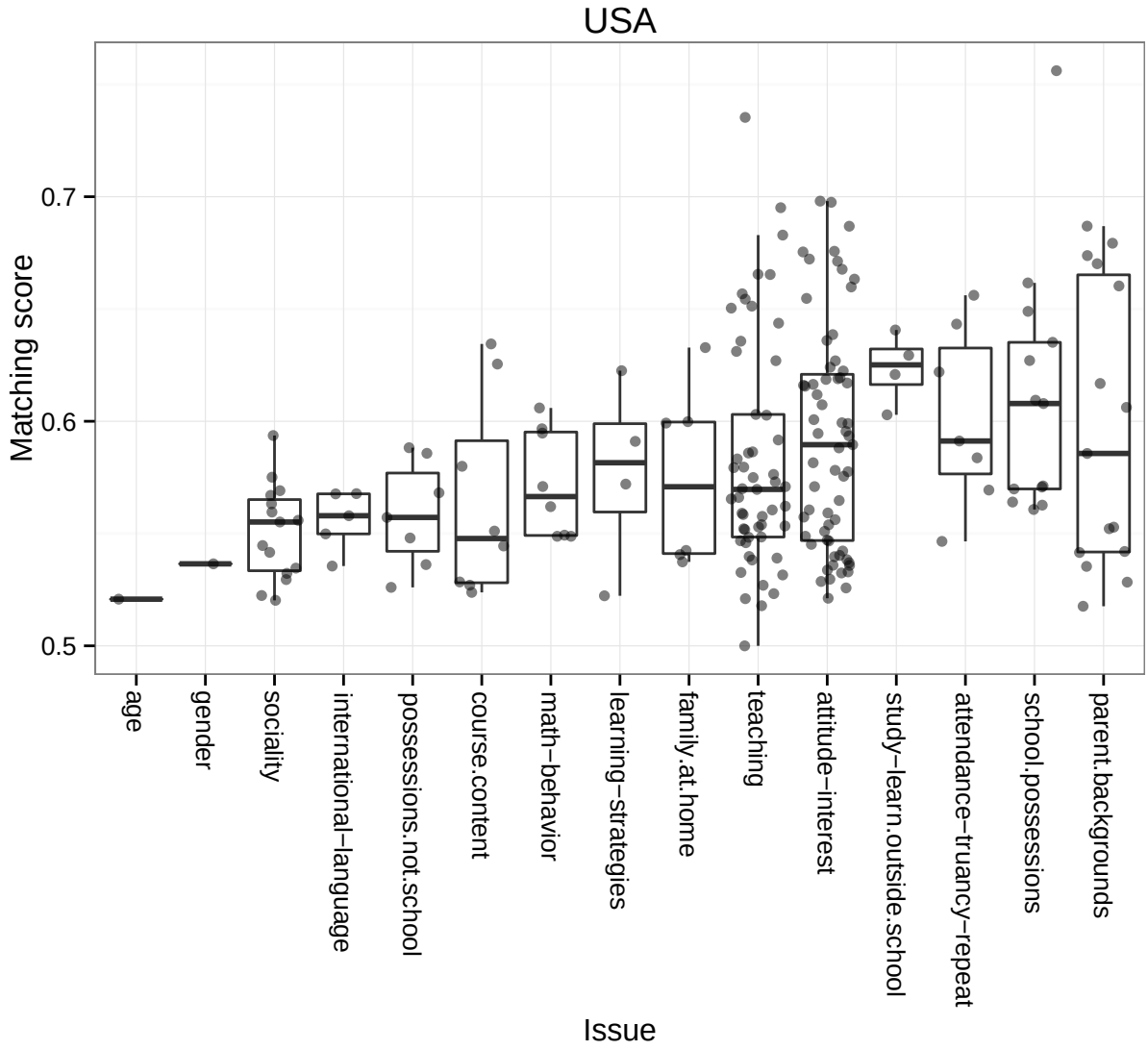

Figure 3: matching scores of the 210 variables, where the variables are grouped by the general issues they cover, such as possessions, attitudes, teaching, etc. The results are not definitive because the matching scores only apply to separate variables individually. However, we can start to identify potentially useful key issues in education. Notable 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.

4.1 Issue-specific datasets

4.1.1 Teaching

The USA teaching variables measure many different aspects of teaching style and quality as experienced by the students, such as the frequency of homework, the 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.

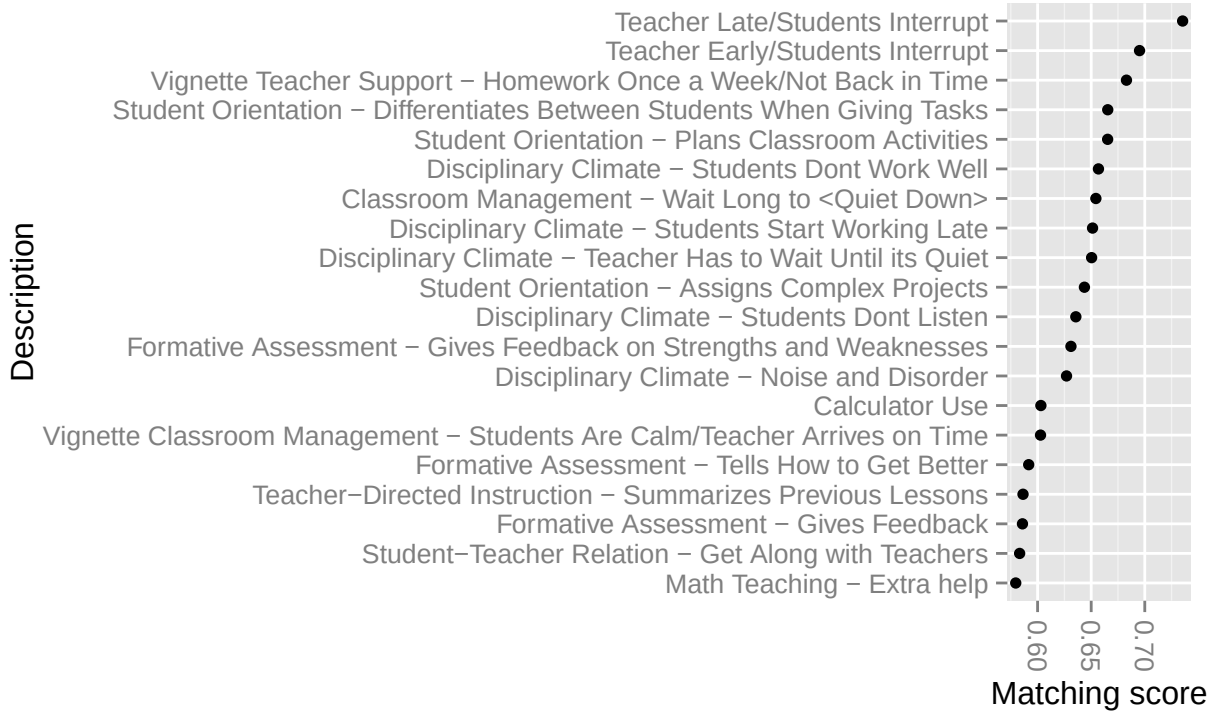

Figure 4: The USA teaching variables measure many different aspects of teaching style and quality as experienced by the students, such as the frequency of homework, the quality of feedback, classroom management, the disciplinary climate of the classroom, student-teacher rapport, assessments, and calculator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

4.1.2 Attitude/interest/motivation

The USA attitude/interest/motivation 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 USA 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. The nondescript "ISCED qualifications" variables are barely documented, and the only levels of these factors are the unhelpful "yes" and "no".) I use all 15 of these variables for the parental backgrounds dataset, shown in Figure 6.

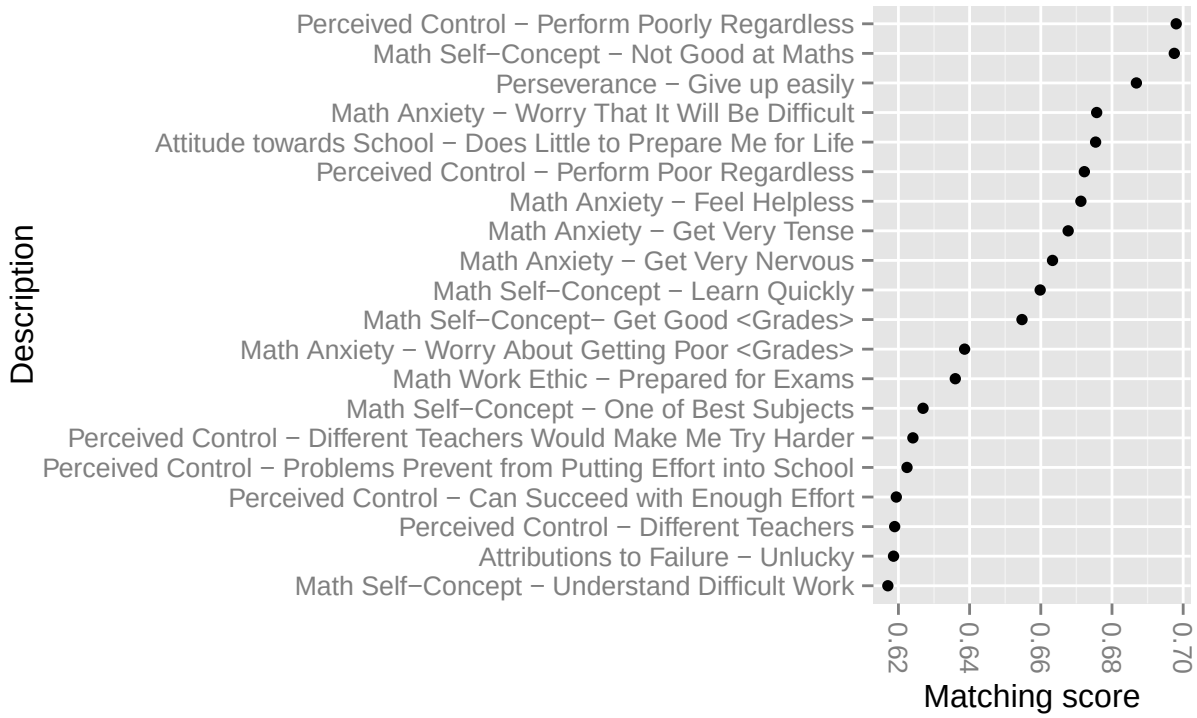

Figure 5: The USA attitude/interest/motivation variables are student self-reported measures of perceived control, work ethic, motivation, attitude towards school, anxiety, attributions to failure, and perseverance. For the attitude/interest/motivation dataset, I take the variables in this area with the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

4.1.4 School-related possessions

The USA 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, for the USA, 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 five datasets above has missing values, and before I can classify students, I need to impute them. In this subsection, I remove some students and then carry out a 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 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 portion of students (less than 4%) with over 15 missing values. Many students remain, possibly too many: imputing too many missing values could distort basic summary statistics on the data. Hence, I check that

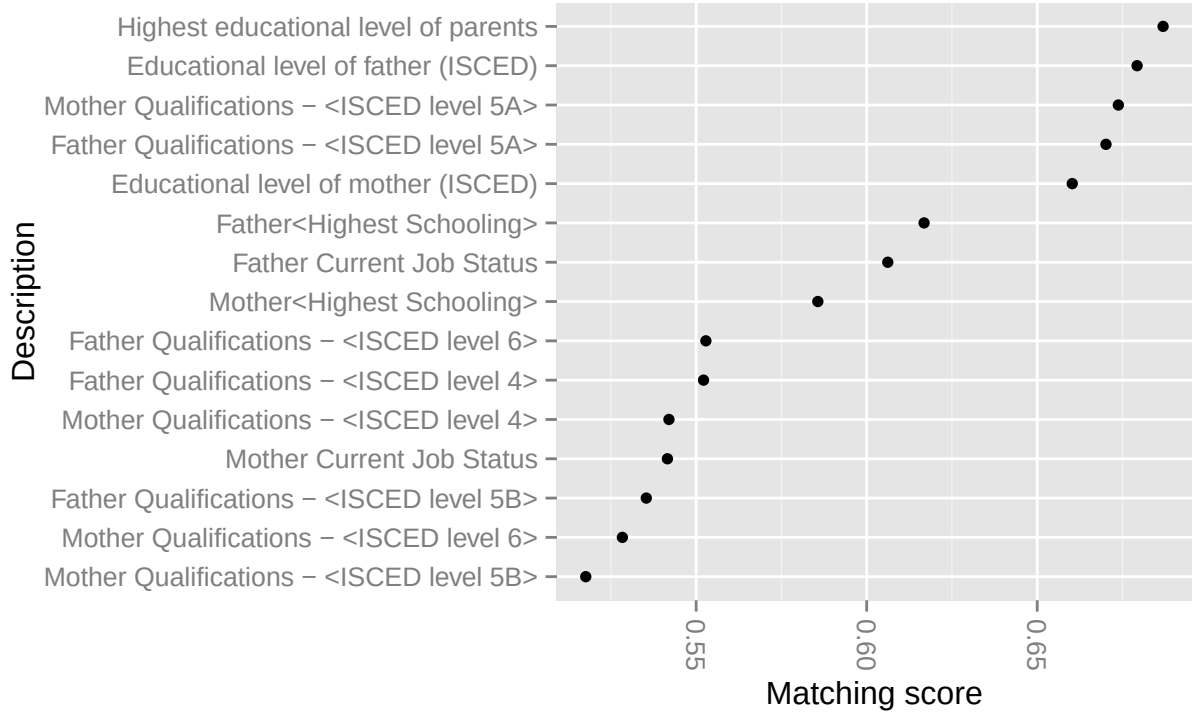

Figure 6: The USA parental background variables measure the educational levels, job statuses, and "ISCED qualifications" of the parents of each student. (It's a shame that PISA does not explain what these ISCED qualifications really mean. The nondescript "ISCED qualifications" variables are barely documented, and the only levels of these factors are the unhelpful "yes" and "no".) I use all 15 of these variables for the parental backgrounds dataset, shown here.

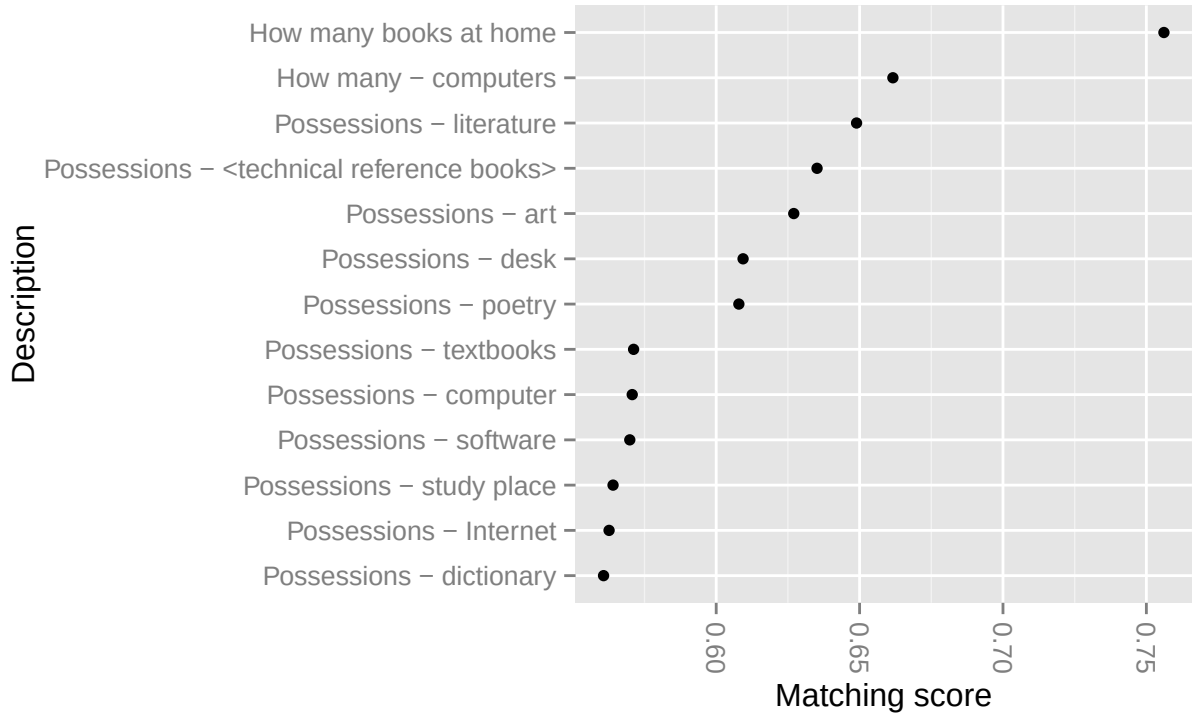

Figure 7: The USA school-related possessions variables measure things like the number of books at home (highest matching score by far), number of computers, number of textbooks, access to internet, and access to study space. I use all 13 of these variables, shown here.

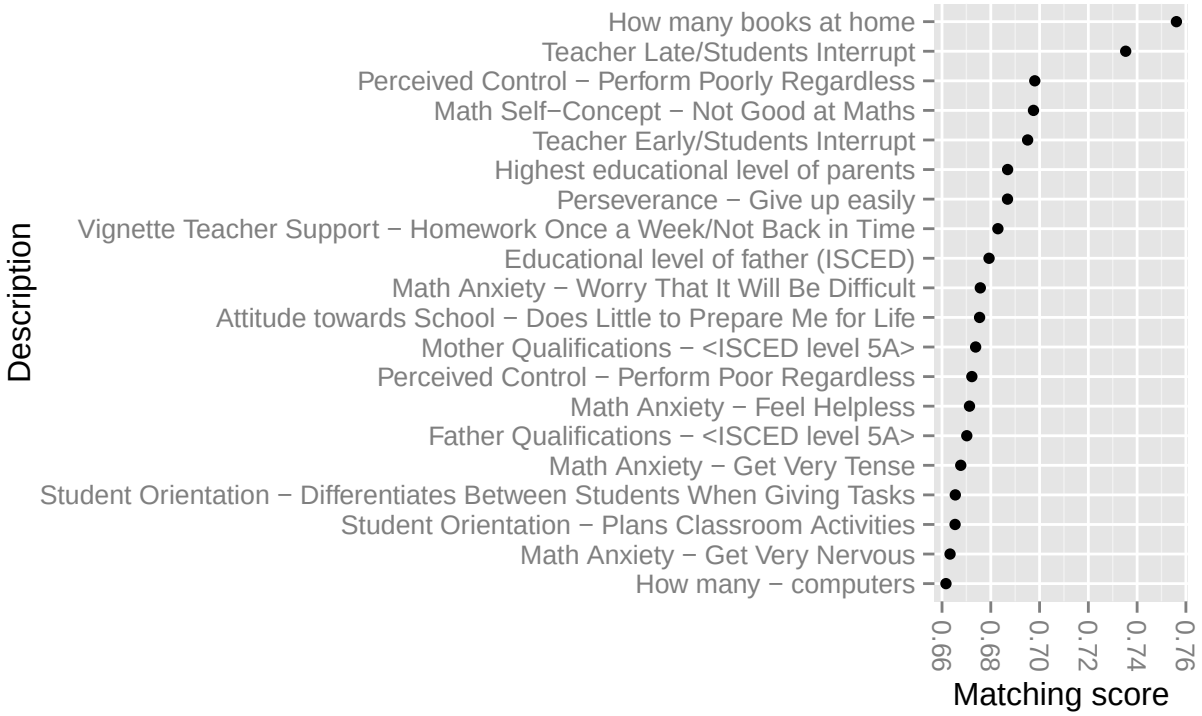

Figure 8: For the sake of comparison, for the USA, I collect the variables with the top 20 matching scores out of all the usable 210 factors from the USA PISA student dataset, shown here.

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.

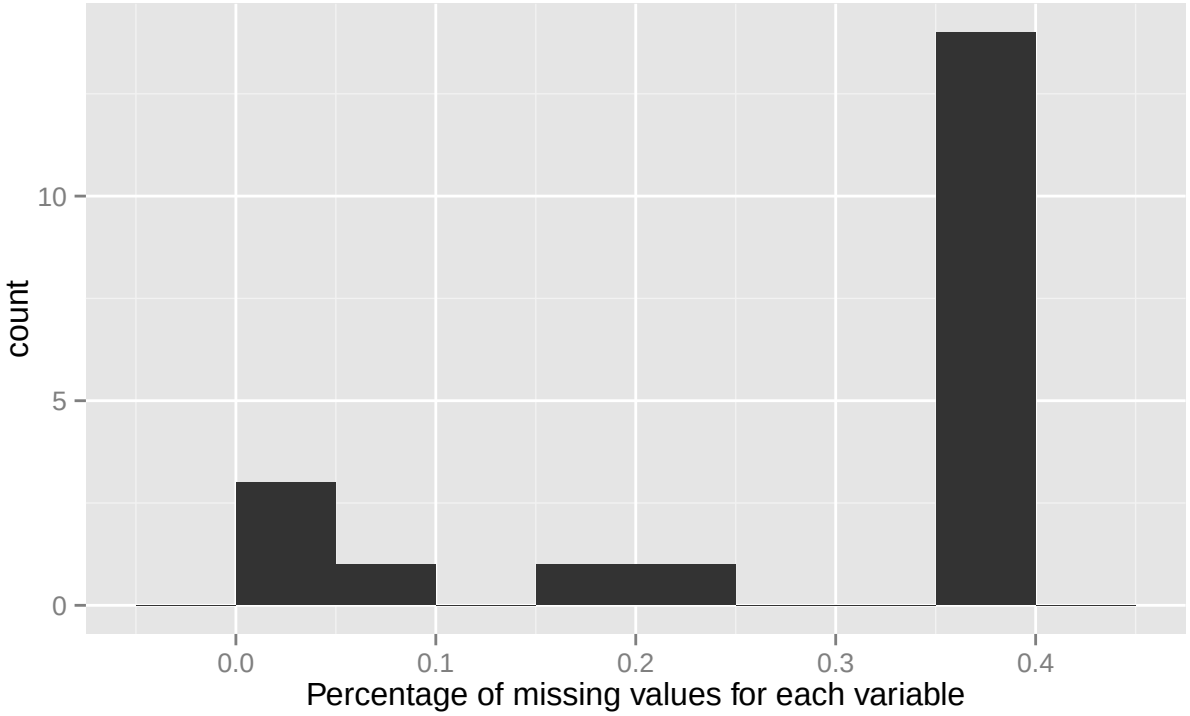

Figure 9: percentage of missing values for each variable in the "Top 20 variables" USA dataset. Many variables have an entire third of their values missing, so the imputation process is messy.

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 success on the PISA exam. 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)).
- 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 2002b)).

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 rates of correct classification on the test sets, grouped by key

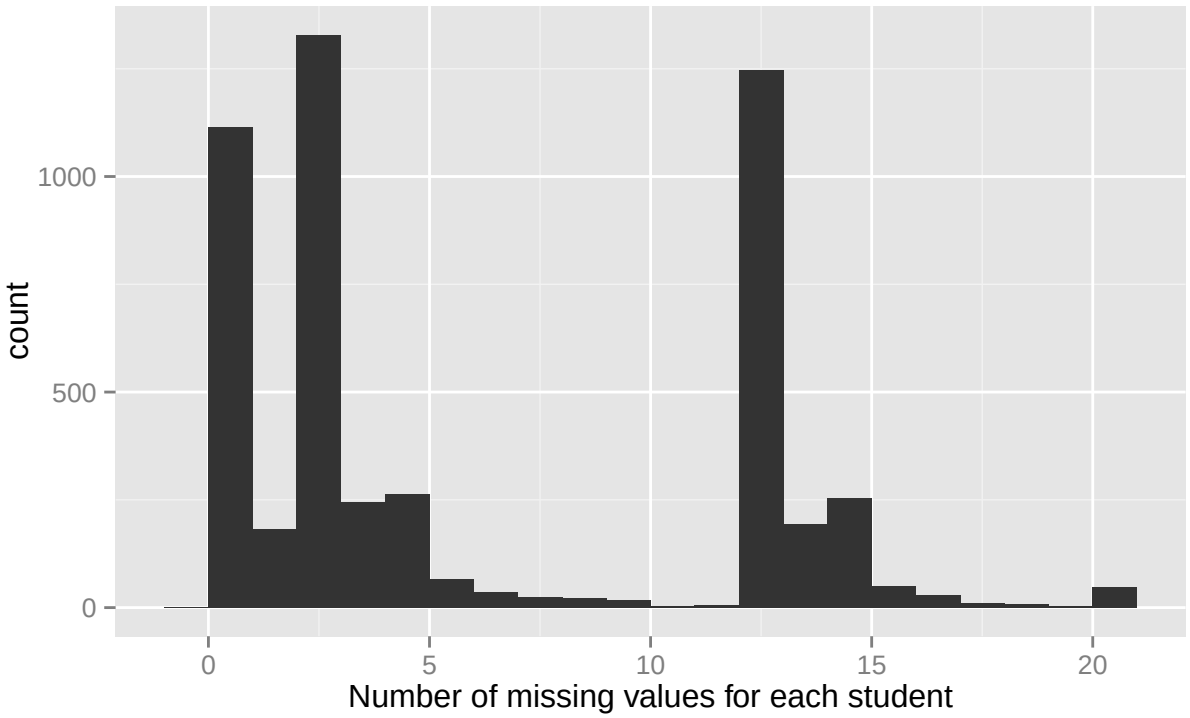

Figure 10: number of missing values per student in the "Top 20 variables" USA dataset. Some students have an unmanageably high number of missing values.

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, then the parental backgrounds dataset.

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 about most of the other categories.

5 Other countries

How does the USA compare to other countries? Are different factors linked with 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 atti-

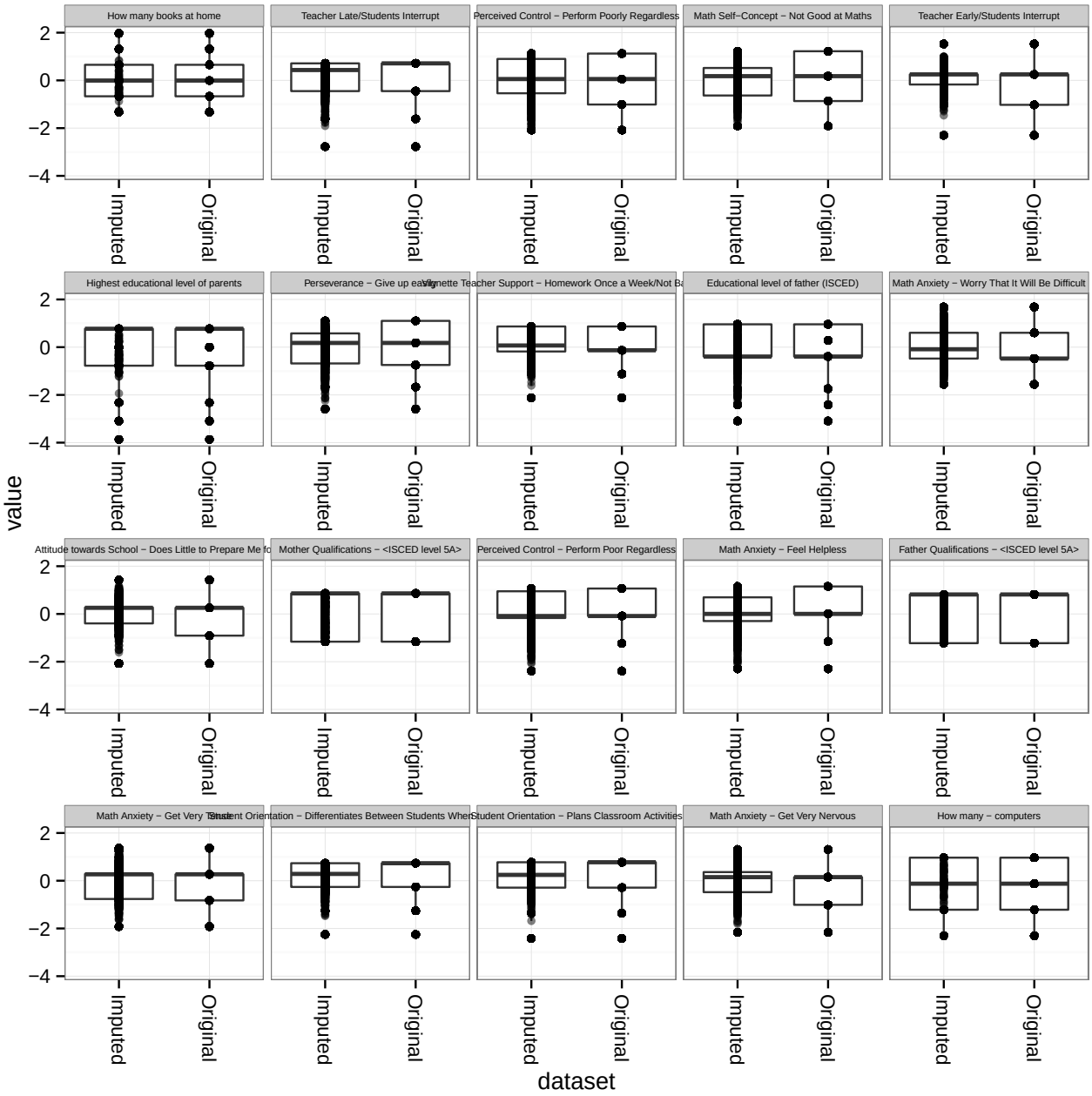

Figure 11: Here, I check that the summary information for the imputed data is the same as the summary information of the original data ("Top 20 variables" USA data). The ranges are the same between the imputed and non-imputed versions, and most of the quartiles are the same. The 10-nearest-neighbors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

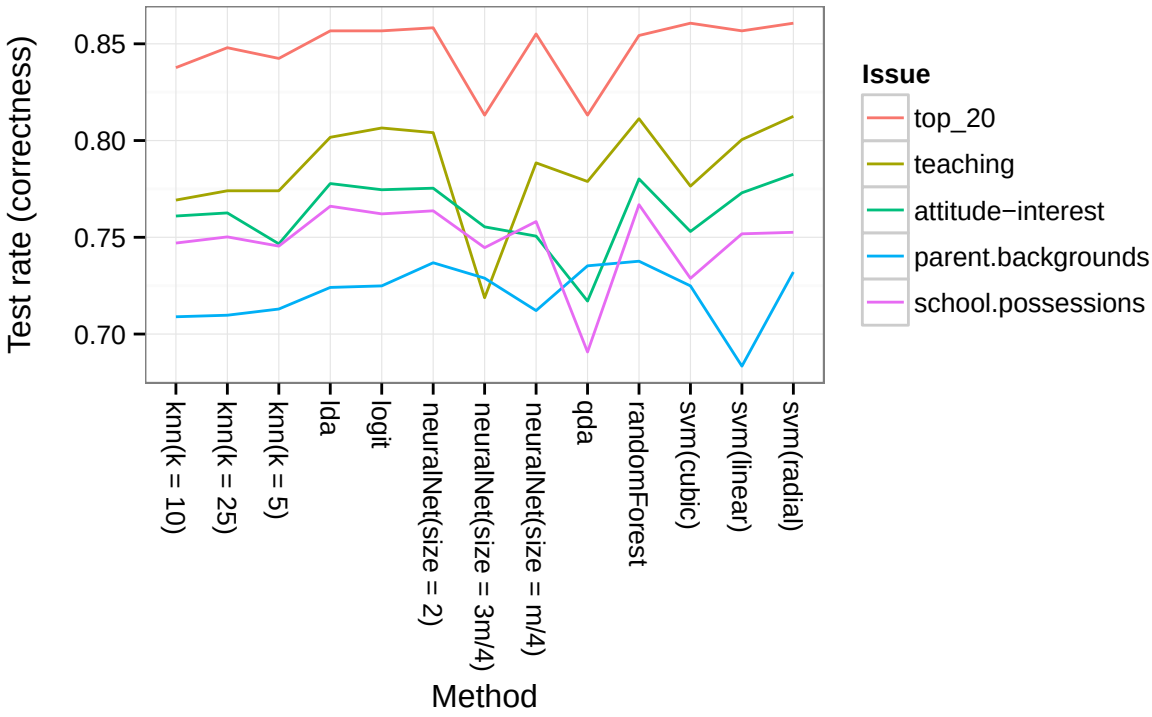

Figure 12: for the USA, rates of correct classification on the test sets, grouped by key issue, after training on the training sets. As expected, all the classifiers performed best on the dataset with the variables with the top 20 matching scores. After that, the teaching dataset was the most useful,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

tude/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 noteworthy 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 shows the rates of correct classification on the test sets for each country. (Here, I should say that for some cases, the k-nearest-neighbors classifiers failed because of too many ties. Those results were not plotted.) The different patterns in the plot may be consequences of different implementations of PISA in different countries, or they may reveal real differences among the education systems of the different countries. Below, I make observations about each country according to the figure.

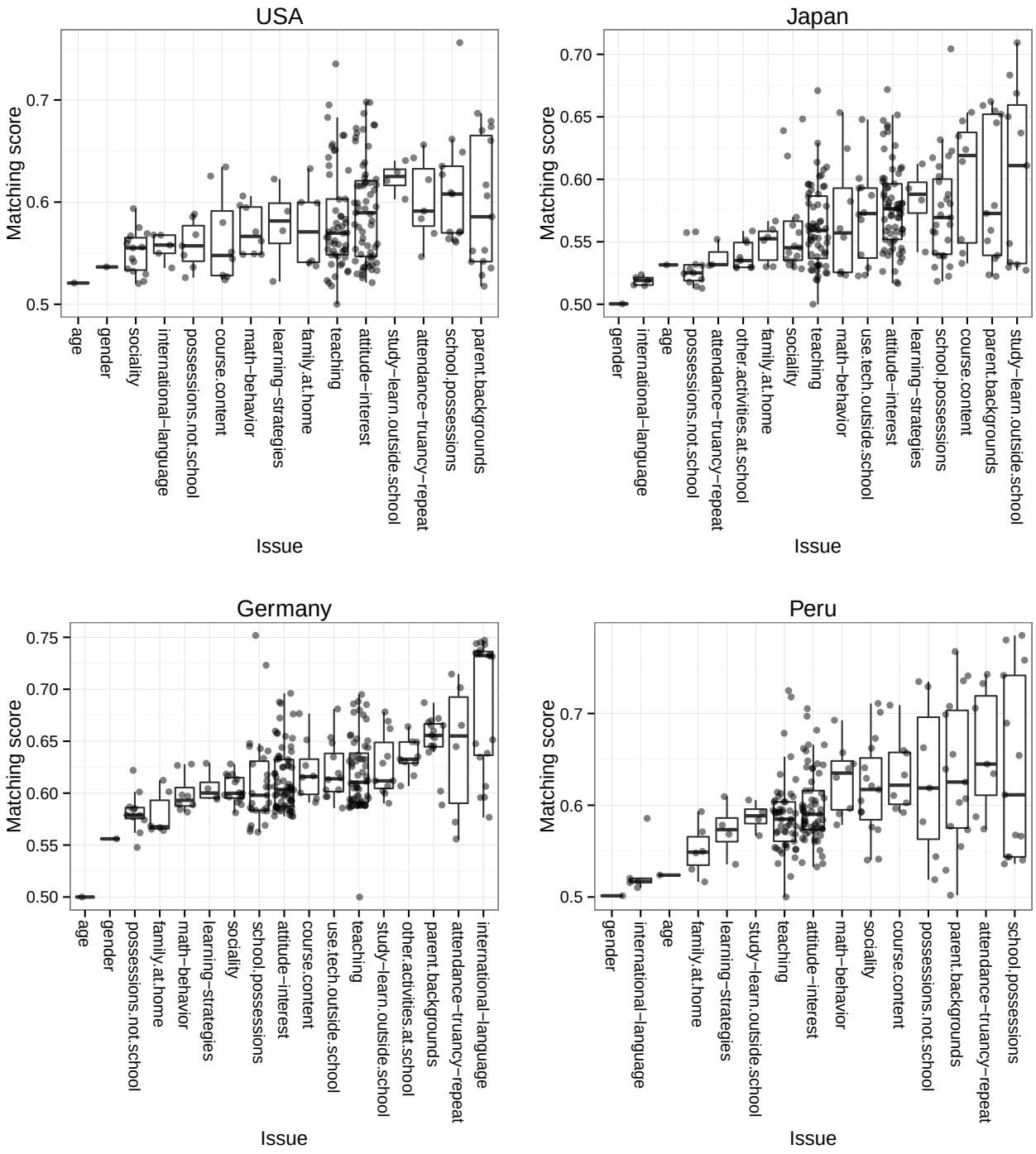

Figure 13: analogue of Figure 2, but for the USA, Japan, Germany, and Peru all together. Matching scores of individual variables are grouped by key issue and plotted. In general, teaching and attitude/interest/motivation seem to lose some of their relative importance in the other three countries, but 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.

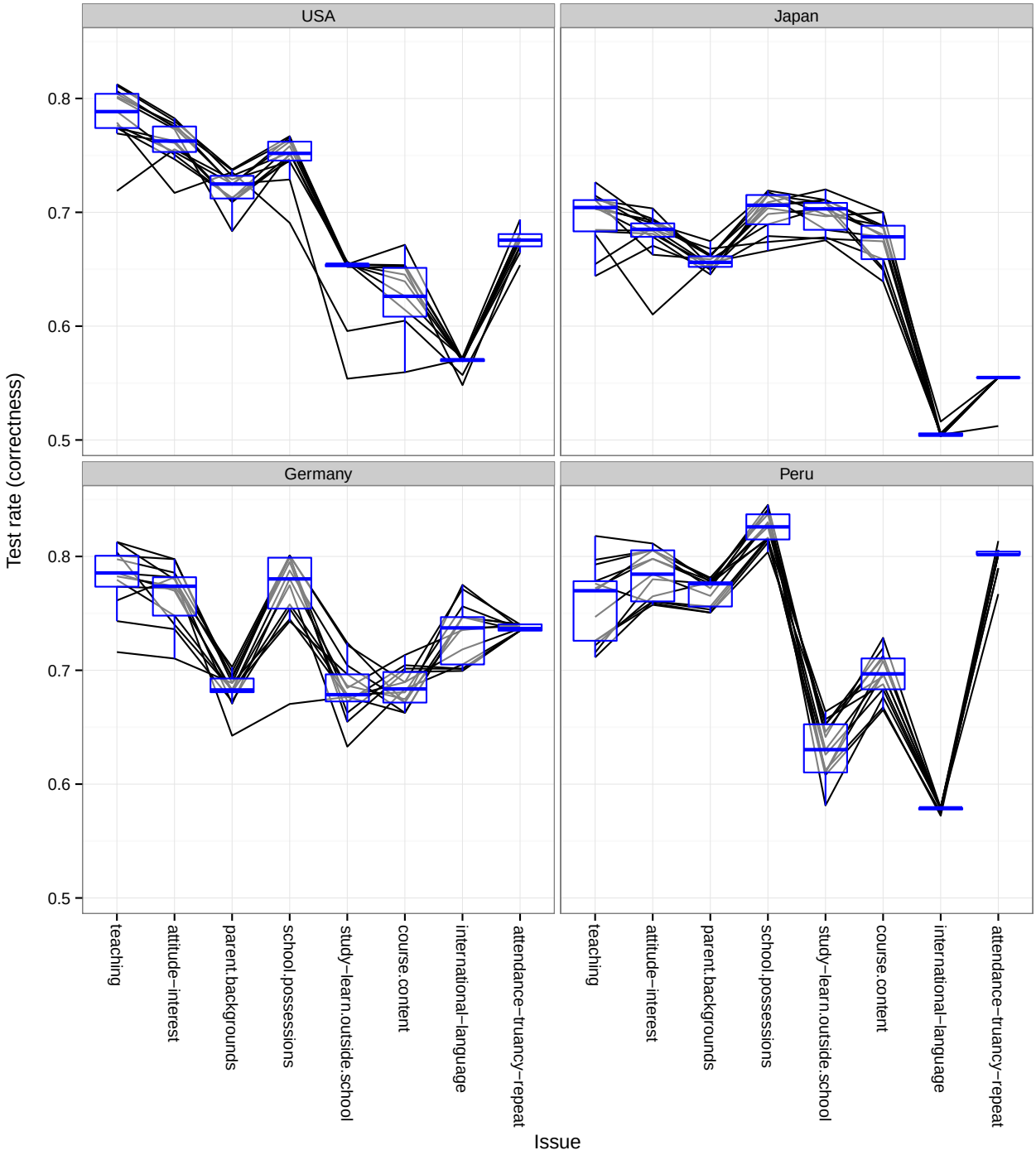

Figure 14: rates of correct classification on the test sets for each country. The different patterns in the plot may be consequences of different implementations of PISA in different countries, or they may reveal real differences among the education systems of the different countries. In the "Classifying students" subsection of the "Other countries" section, I make observations about each country according to this figure.

5.2.1 The United States

Teaching quality is relatively decisive, while study habits outside school, course content, attendance/truancy, and international backgrounds (whether the parents are from other countries, whether a foreign language is spoken at home, etc.) are less relatively useful in predicting success. 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 teaching quality predicts student success more accurately than other issues, it may be the most appropriate next issue to tackle to improve US education policy, hopefully in a way that honors and fully compensates teachers.

5.2.2 Japan

In general, students were more difficult to classify in Japan than in other countries. It may be that Japanese schools do not have the typical problems that other schools do, or it could be that 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.

5.2.3 Germany

Germany is very much like the USA in that teaching quality and attitude/interest/motivation are important. Unlike the USA, however, course content (mostly questions about the kinds of math problems on routine in-class exams), international backgrounds, and attendance/truancy are more decisive in determining student success. Granted, these factors could appear more decisive here than in the USA because response rates are higher for Germany. However, it's possible that curriculum design and school attendance policy are relatively easy opportunities to effect change.

5.2.4 Peru

Peru's school system has been notoriously under-resourced for generations, which explains why school-related possessions and attendance/truancy are so telling. In addition, there is not as much upward mobility as in the other countries here, which is why the parental backgrounds are of particular importance. Overall, Peru's students are easier to classify than in the other countries listed, so the road to reform may be clearer here than elsewhere. However, speaking outside the scope of PISA, Peru's education problems are very much tied up in inequality issues and corruption in government.

6 Conclusion

I combed through the 2012 PISA student dataset, attempted to extract the larger issues, and tried to offer some basic implications for education policy. The recommendations above are few and unsurprising, and the underlying findings could be tangled with the biases, preconceptions, and priorities of the PISA surveyors, as well as the different implementations and nonresponse patterns in different countries. However, this approach does profile a large and cumbersome dataset, and the analysis hopefully offers some intuition.

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 2002b), `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.

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.