# Comparing PISA and TIMSS over time - STAT 4701 Final Project

Barbara Welsh

May 15, 2014

## Overview

For this project, I chose to compare two international assessment tests - PISA and TIMSS. In 2003 both tests were given at the same time, which led to several studies comparing the assessments. I chose to focus on one in particular, which found that in mathematics western countries tended to do better on the PISA assessment and eastern countries tended to do better on TIMSS. I used this as a starting point to compare scores on the two assessments overall, as well as by subject and gender on the 2003 and subsequent assessments.

## Background

The PISA assessment is given to 15 year olds internationally on a three year interval, most recently in 2012. It generally consists of a reading, mathematics and science component, although other subjects have been included in various years. Each year the focus changes, in 2003 it was mathematics, in 2006 science, in 2009 reading, and in 2012 mathematics again. Significantly more data is available for the focus subject in each round of testing. The stated purpose of the PISA assessment is to assess student's ability to apply their knowledge to real world situations[1].

The TIMSS assessment is given to 8th graders (and 4th graders as well, although only 8th grade results are included in this analysis) internationally on a four year cycle, most recently in 2011. It consists of mathematics and science components. The stated purpose of the TIMSS assessment is to measure trends in math and science achievement[2].

The assessment done on the 2003 testing[3], found three factors that contributed to the difference seen between relative performance between western and eastern countries on PISA and TIMSS -

- Age vs Grade: the differing methodology on how select students for assessment
- Content Differences: the PISA focus on applied knowledge vs the TIMSS focus on school learning
- Amount of reading: PISA tests had significantly more reading involved

Content differences were found to be the biggest factor, and that was chosen to be the focus of the analysis.

## Data Collection

To collect data for this analysis, raw scores were obtained for each student who underwent the assessment. In reading the assessment methodology for both studies, it was discovered that in order to be able to have more questions in the assessment, without making the assessment unduly long, questions were broken up into booklets and each student received a small selection of all questions asked. The results were then calculated

[1] OECD. *OECD PISA website*. http://www.oecd.org/pisa/aboutpisa. 2014.

[2] BC. *TIMSS and PIRLS website*. http://timssandpirls.bc.edu. 2014.

[3] M. Wu. "Comparing the Similarties and Differences of PISA 2003 and TIMSS". In: *OECD Education Working Papers* 32 (2010). http://dx.doi.org/10.1787/5km4psnm13nx-en.

using an IRT model and scaled according to a complex methodology. The individual item responses, then, were really only of use for looking at individual questions, not for obtaining country-wide results.

The country-wide results and breakdowns by gender and content area were available, although sometimes only in PDF form. In a few cases an excel file was available with the data in it, or a PDF could be obtained where the table was not an image at least. Because the results from each assessment each round were available in a different format, no script could be written to compile the data. Instead, all of the results were put together on a csv, which could then be read into R for the analysis phase. Because the primary interest was in relative performance changes over time, but the assessments were not given on the same cycle, the assessment were broken into three rounds - the 2003 round, the 2006/2007 round, and the 2011/2012 round. This meant leaving out the PISA 2009 assessment, but as the focus that year was on reading, less data was available in the mathematics and science domains for that year.

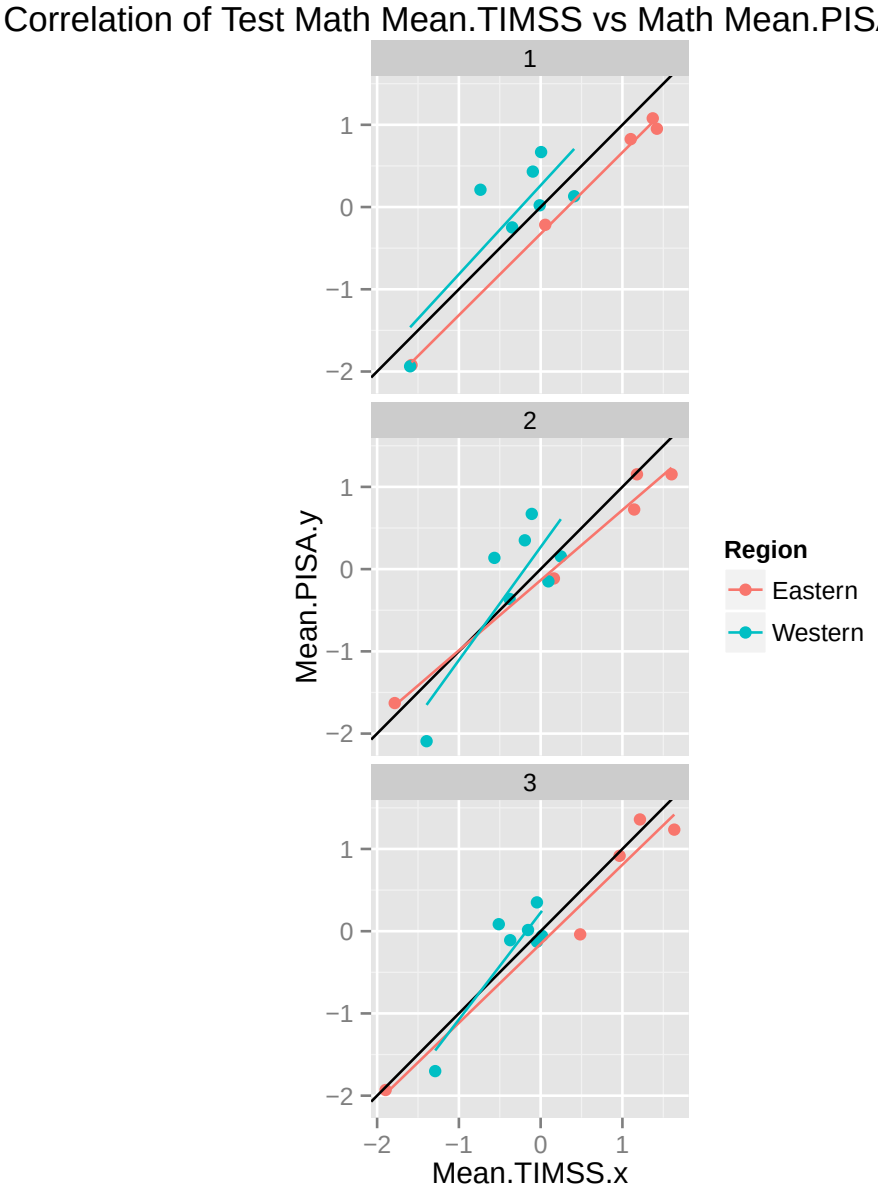
# Exploratory Analysis

R was chosen for the analysis phase of this project. There are several questions we can look to answer:

## Did the pattern of the relative performance between western and eastern countries seen in 2003 hold for later rounds of testing? Did the gap widen or narrow?

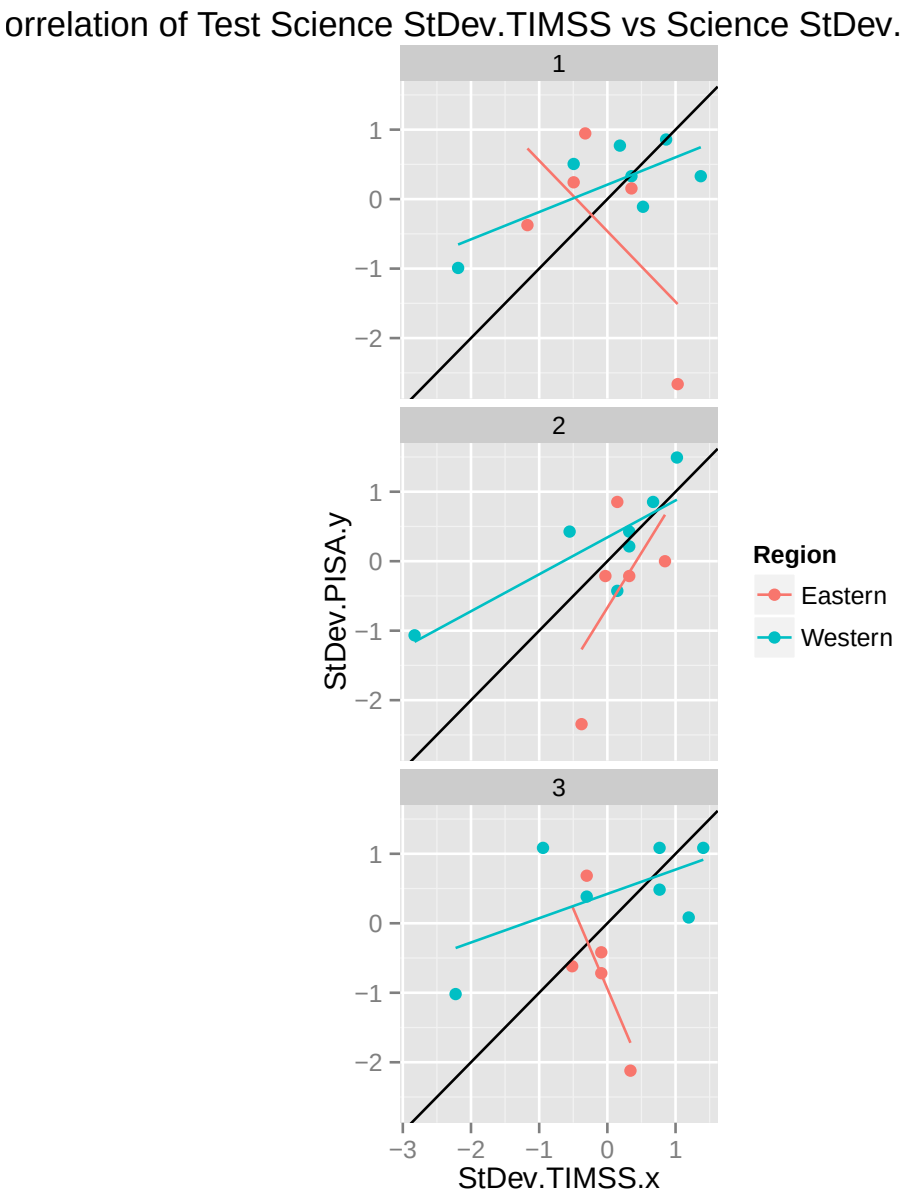
In order to explore this question, the data was read into R, and a function was written to create a series of plots, one for each round, to see how the relative performance of western and eastern countries changes over time. A discussion of the code to create all of the plots, using ggplot, is included at the end of this section. The parameters for this function are the raw data file, a vector of the rounds to include in the comparison, the dimension to plot by (Test in this case), the subjects (Math or Science) to compare (one input each for the x and y axis), and the statistic (Mean.[Test] or StDev.[Test]) to compare, again one input each for the x and y axis. The default is to compare the means for the math scores. This then allows for the comparison of either subject, either statistic, for the countries that participated in both tests for all rounds included. For example, to compare the mean scores in the math subject for all three rounds, we would run the following code:

```
source("countryData.R")
plotCorrelations("country_data_3.csv", c(1, 2, 3), "Test", "Math", "Math", "Mean.TIMSS",
    "Mean.PISA")
```



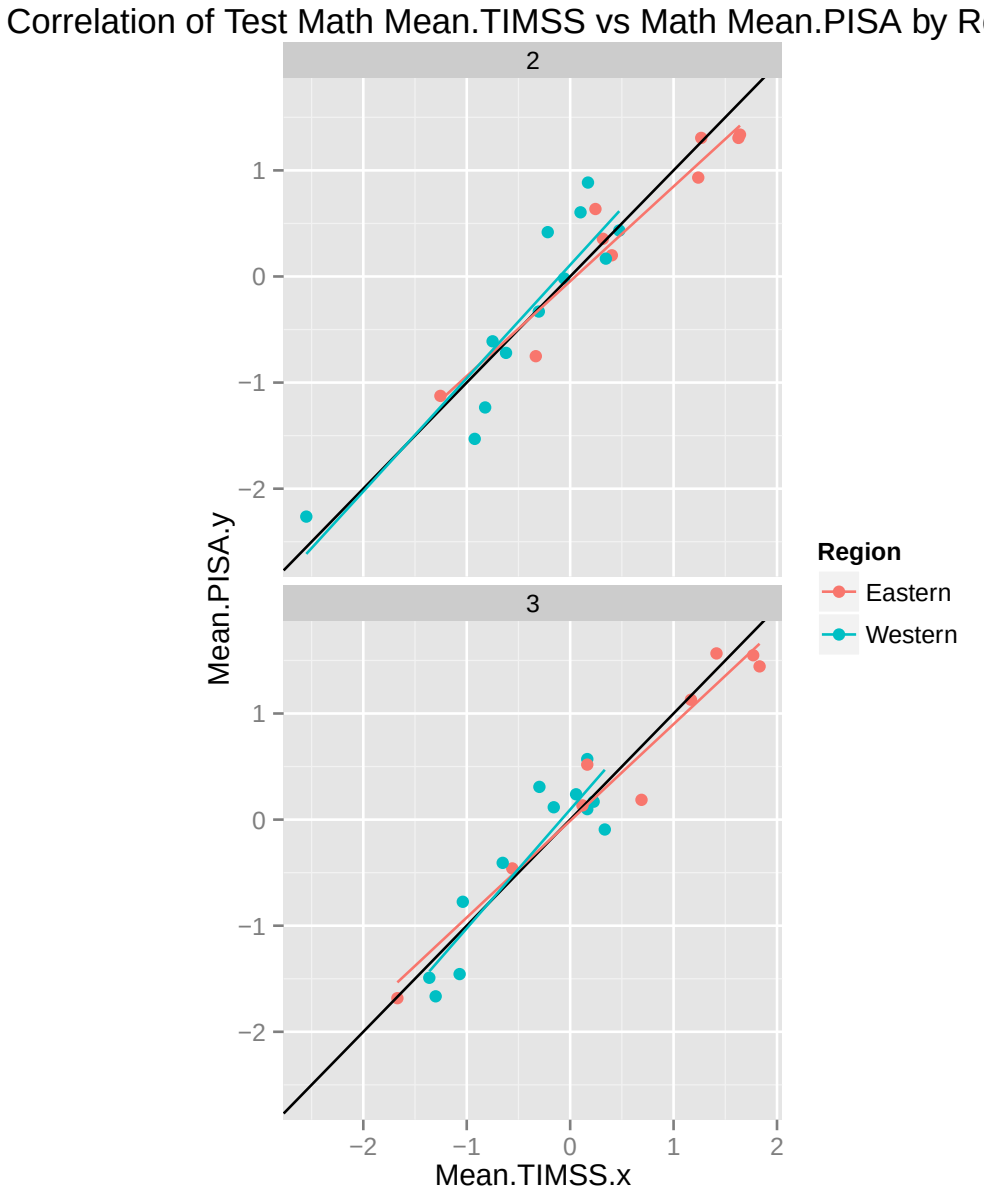

We can create a similar plot for science deviations:

```
source("countryData.R")
plotCorrelations("country_data_3.csv", c(1, 2, 3), "Test", "Science", "Science",
    "StDev.TIMSS", "StDev.PISA")
```




Only a handful of countries participated in all rounds of testing, so in order to see if a similar pattern held for more countries, we could restrict our view to just the last two rounds of testing, producing the following plot for math mean scores:

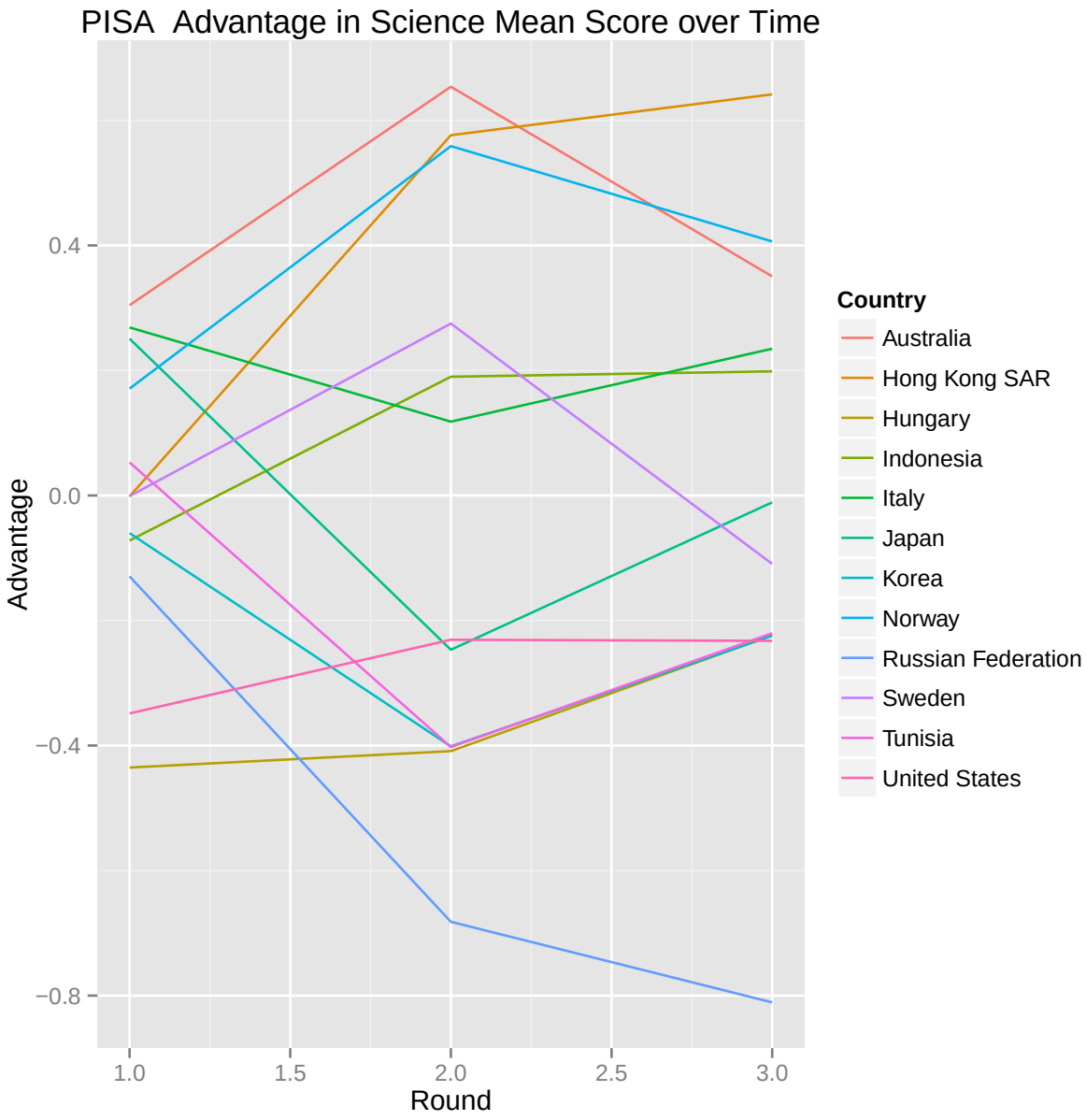
```
source("countryData.R")
plotCorrelations("country_data_3.csv", c(2, 3), "Test", "Math", "Math", "Mean.TIMSS",
    "Mean.PISA")
```




These plots tell us some interesting things. From the first plot, we can definitely see the delineation between western and eastern countries for the first round. The regression lines are almost parallel, but is above the line for western countries (indicating better performance on PISA) and below the line for eastern countries (indicating better performance on TIMSS). The pattern hold for subsequent rounds, although not to the same extent. This indicates the the gap is narrowing over time. The second graph shows standard deviations, instead of means. This is interesting as we can see that for all rounds the standard deviation regression line for western countries stays relatively constant, while is widely different for eastern countries. This indicates that over time, there's not a correlation between standard deviations on the different tests, at least for the eastern countries. Adding additional countries to the analysis in the third plot confirms the pattern seen in the initial plot.

Since this plot does not tell us about the specific countries, only relative performance by region, an additional plot was created to view relative performance by country over the rounds. The parameters for this function are the input file, a vector of the rounds, the dimension to plot by (Test, in this case), the subject (Math or Science), and the statistic (Mean or StDev). For example, we can look at the relative mean scores over all three rounds for science with the code below:

```
source("countryData.R")
plotAdvantage("country_data_3.csv", c(1, 2, 3), "Test", "Science", "Mean")
```



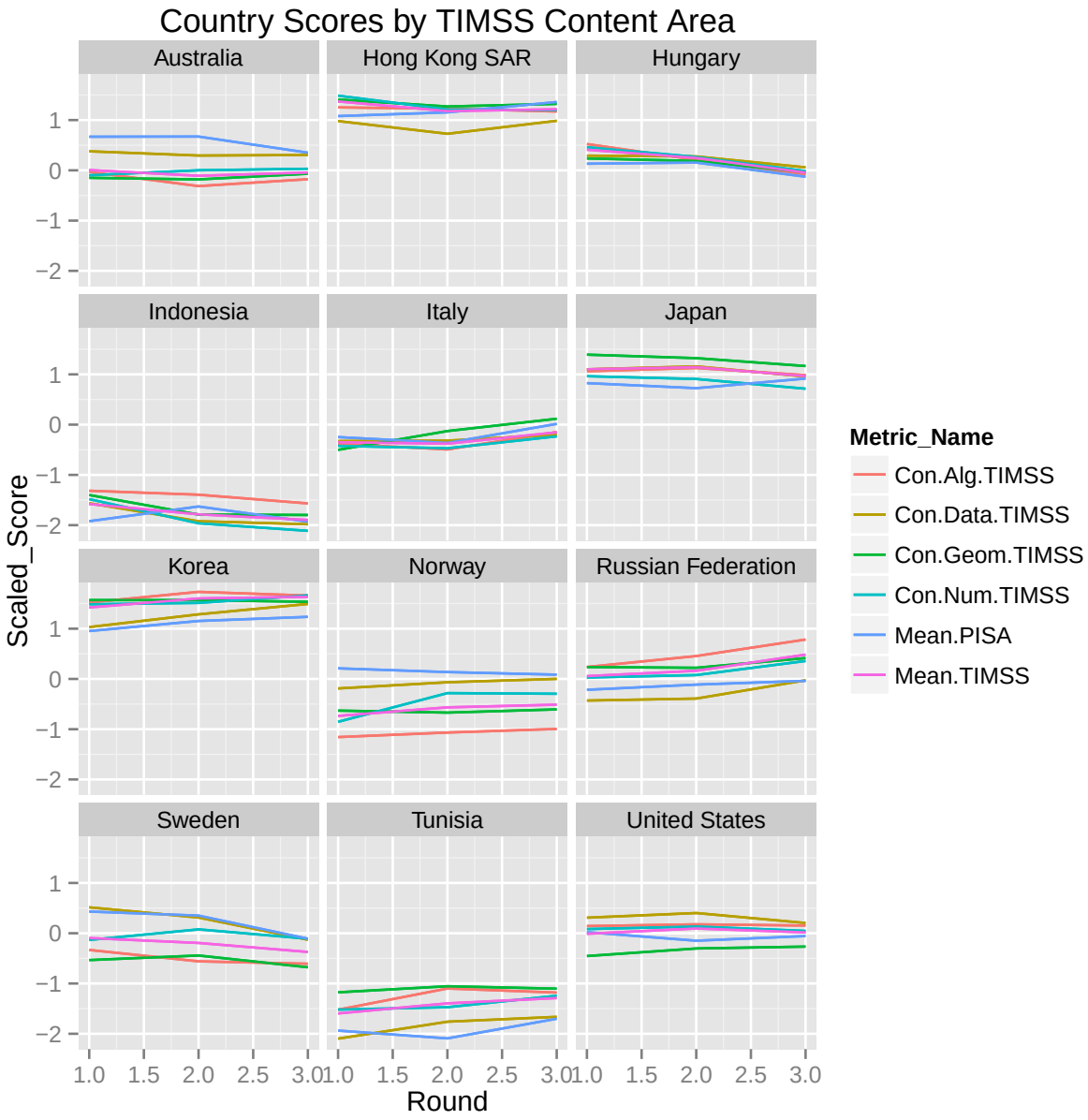

From this plot, it looks as though the gap in scores is widening over time, but upon closer inspection, it is clear that this is due to changes in two countries, Russia and Hong Kong. Russia has over time done comparatively worse in PISA and better in TIMSS, with Hong Kong going the opposite way. This graph displays science mean scores, while the initial graph displays math mean scores, indicating that a difference in performance change may exist between the two subjects.

## Do countries perform differently in different content areas?

In order to explore this question, a function was written to create a series of plots, one for each country of interest, with performance broken down by content area. Since content area was only available in every round for TIMSS, this plot displays scores for TIMSS only, and only for the mathematics content areas. Inclusion of the science content areas could be an enhancement to this analysis. The parameters for this

plotting function are the input file, a vector of the desired rounds, and the dimension to plot by (Overall Content, in this case). A plot of the math content scores for countries participating in all rounds is shown below:

```
source("countryData.R")
plotMathContent("country_data_3.csv", c(1, 2, 3), "Overall Content")
```

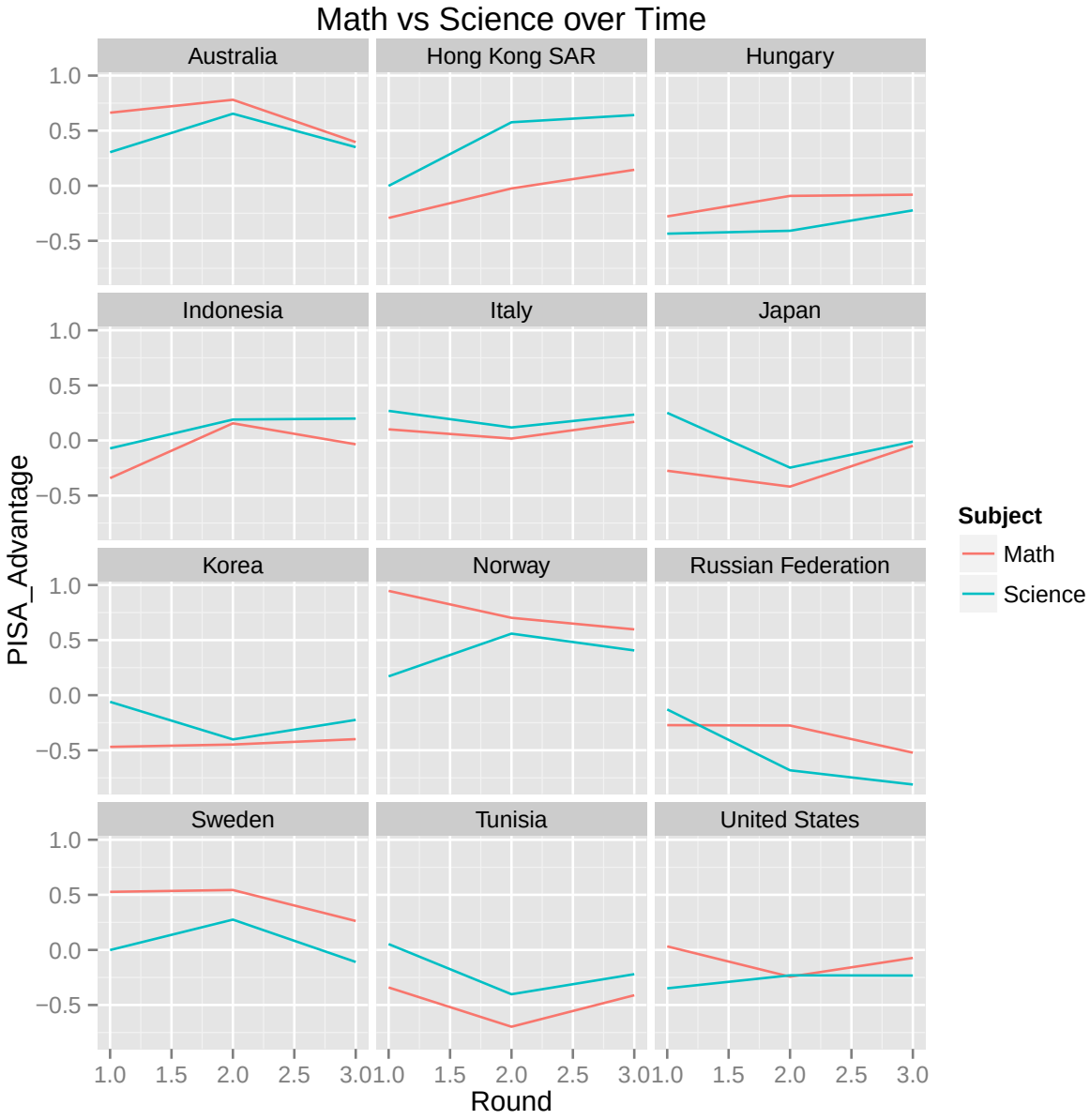



We can see a few things looking at this series of plots. Perhaps the most apparent is that countries vary widely in the spread of their performance by content area. Hungary, for example, does about the same in each content area while Norway's scores different by content area. For most countries, these patterns tend to stay relatively consistent over time. We can also see from here the general, Japan, Korea, and Hong Kong performed very well in every content area, while Indonesia and Tunisia lagged significantly. As these are scores from the TIMSS assessment only, we would expect to see higher scores for the eastern countries, based on our analysis so far.

### Do math and science scores correlate between the countries overall?

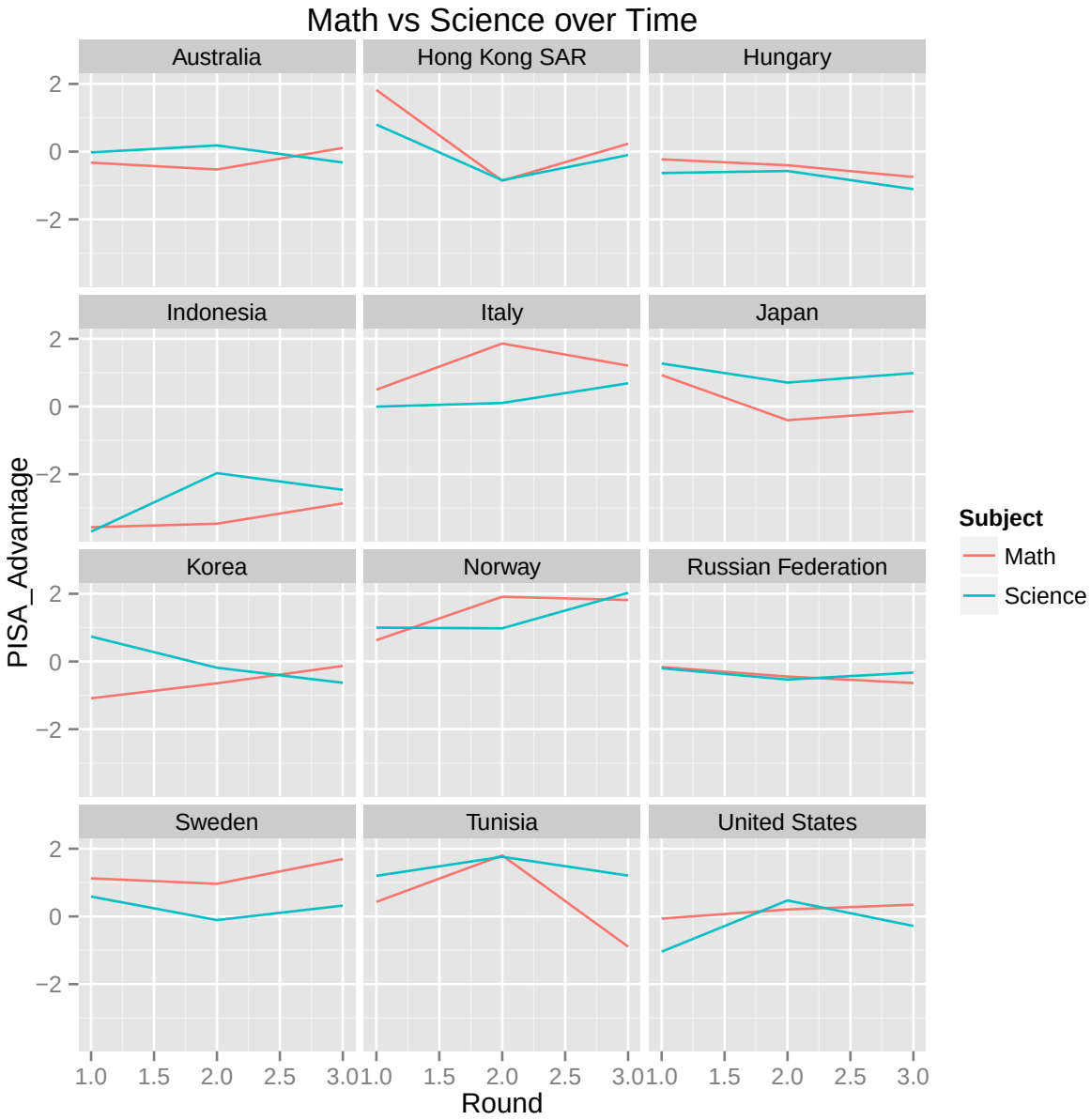
In order to explore this question, a function was written to create a series of plots, one for each country of interest, with performance broken down by subject area. The parameters for this plotting function are the input file, a vector of the desired rounds, and the statistic of interest (Mean or StDev). A plot of the mean scores for countries participating in all rounds is shown below:

```
source("countryData.R")
plotSubjects("country_data_3.csv", c(1, 2, 3), "Mean")
```




Similarly, we can produce a plot showing the standard deviation for each country for each subject:

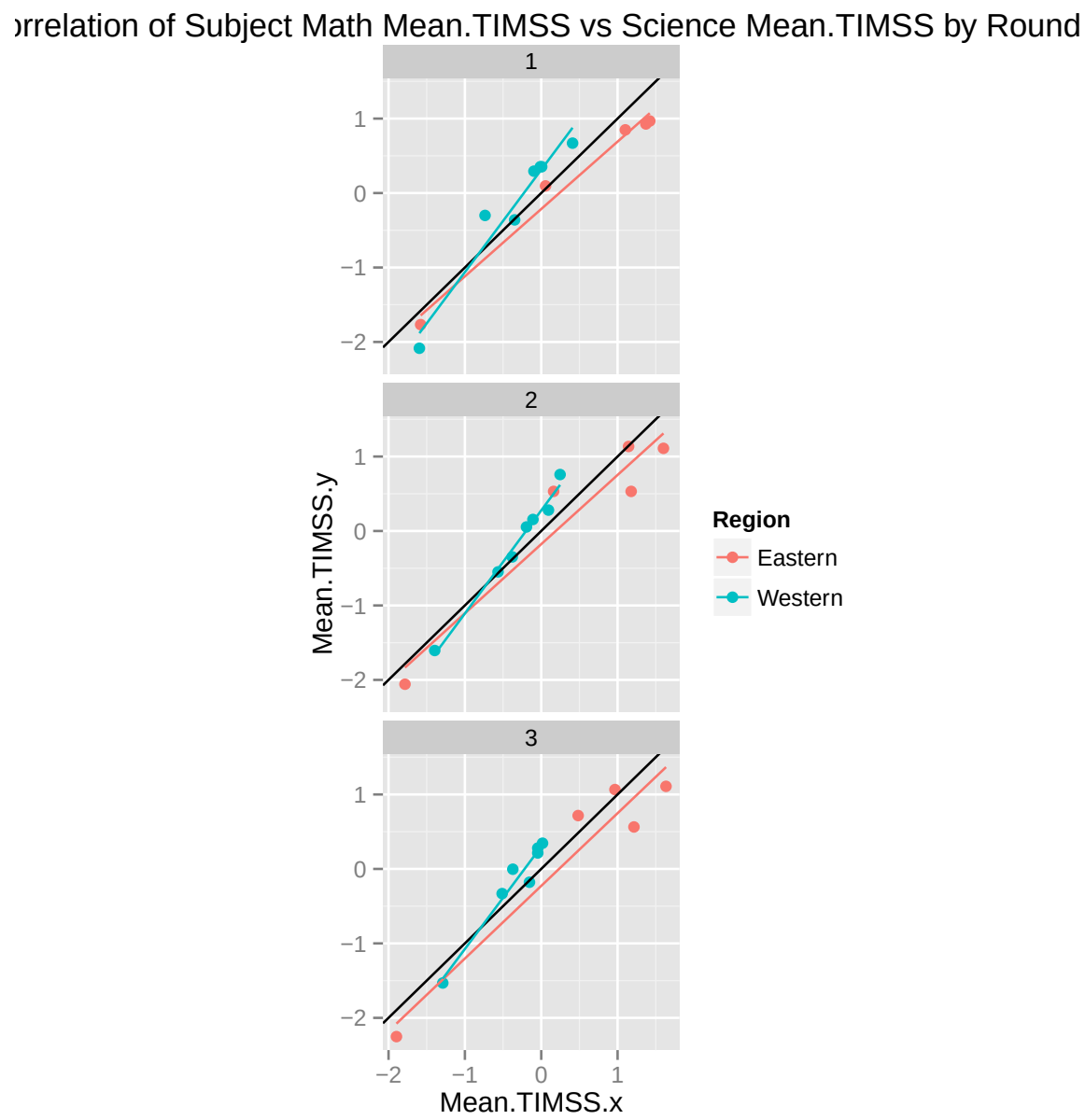
```
source("countryData.R")
plotSubjects("country_data_3.csv", c(1, 2, 3), "StDev")
```




In the first plot here, we notice that the difference in mean scores vary widely by country. For example, Hong Kong had a much better (relative) PISA score in science than in math, whereas there was not much difference for Italy. Another thing we notice here, is that Russia's drop in relative success on PISA was much greater for science then for math. Further analysis could investigate the reasons for these differences. The second graph, showing standard deviation, seems to show that the differences remain less consistent over time. For example, Korea, Norway and Australia all have math and science deviations trending differently over time.

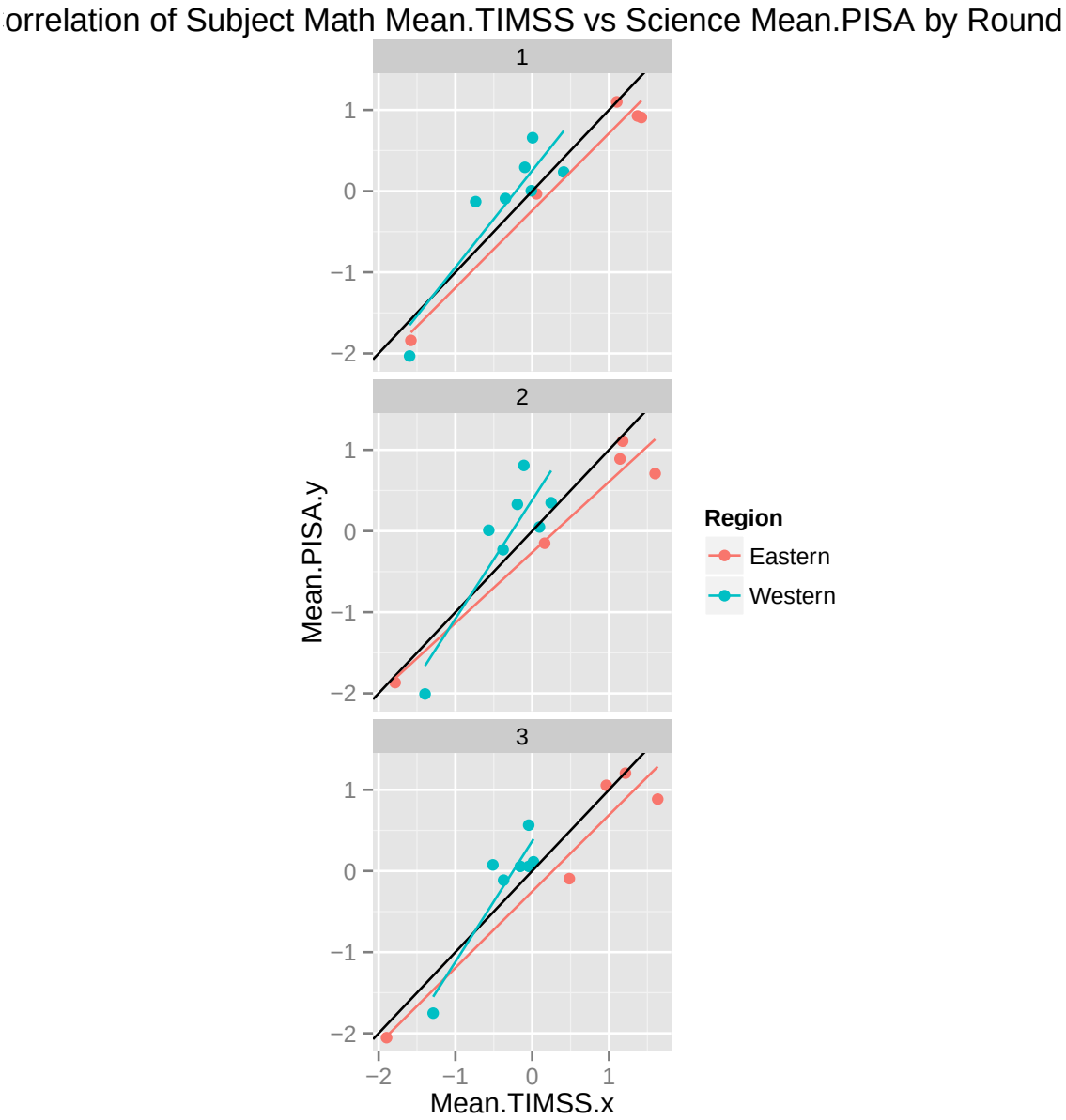
We may also be interested in the correlation in scores between the two subjects. The plotCorrelations function discussed above, can provide this with slightly different parameters. Here, the parameters are the input file, a vector of the rounds, the dimension to plot by (Subject, in this case), the subjects (in this case we would want one subject to be Math and one to be Science), and the statistic (Mean or StDev). A plot comparing mean scores in math and science over the three rounds for TIMSS only is shown below:

```
source("countryData.R")
plotCorrelations("country_data_3.csv", c(1, 2, 3), "Subject", "Math", "Science",
    "Mean.TIMSS", "Mean.TIMSS")
```




We also also create plots that compare one subject for one test with the other subject for the other test, for example, math scores in TIMSS with the Science scores in PISA:

```
source("countryData.R")
plotCorrelations("country_data_3.csv", c(1, 2, 3), "Subject", "Math", "Science",
    "Mean.TIMSS", "Mean.PISA")
```



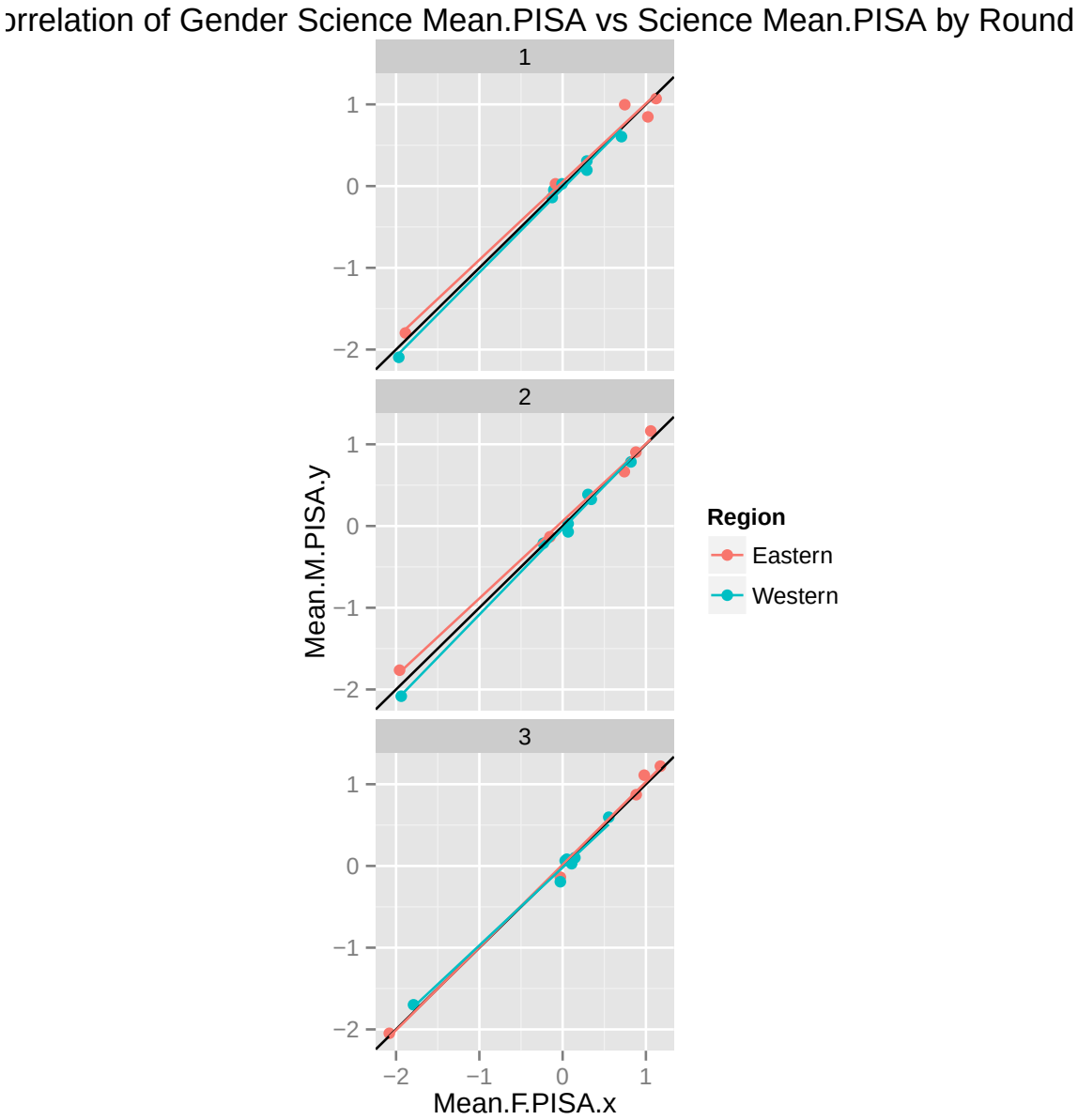

In these two plots, we see a consistent pattern for the eastern countries, relatively better math than science scores for all rounds. This can be seen by a fairly consistent regression line below the diagonal for eastern countries. For western countries, it looks as though for the countries that performed better overall, science scores were relatively better, while for countries that performed worse overall, math scores were relatively better. It would take further analysis to assign meaning to this finding. Also of note is the extremely consistent regression lines in both of these plots, over all three rounds.

### How about gender differences? Are these consistent by subject area? What about between countries?

Lastly, gender differences were looked at overall and by content area. First, the same function described above to plot correlations, could be used to plot the correlation between male and female scores for a given test and subject area. The parameters in this case are the input file, a vector of the rounds, the dimension to plot by (Gender, in this case), the subjects (Math or Science), and the statistic (Mean or StDev). For example, in order to see the mean scores for males and females for the PISA assessment in science, the

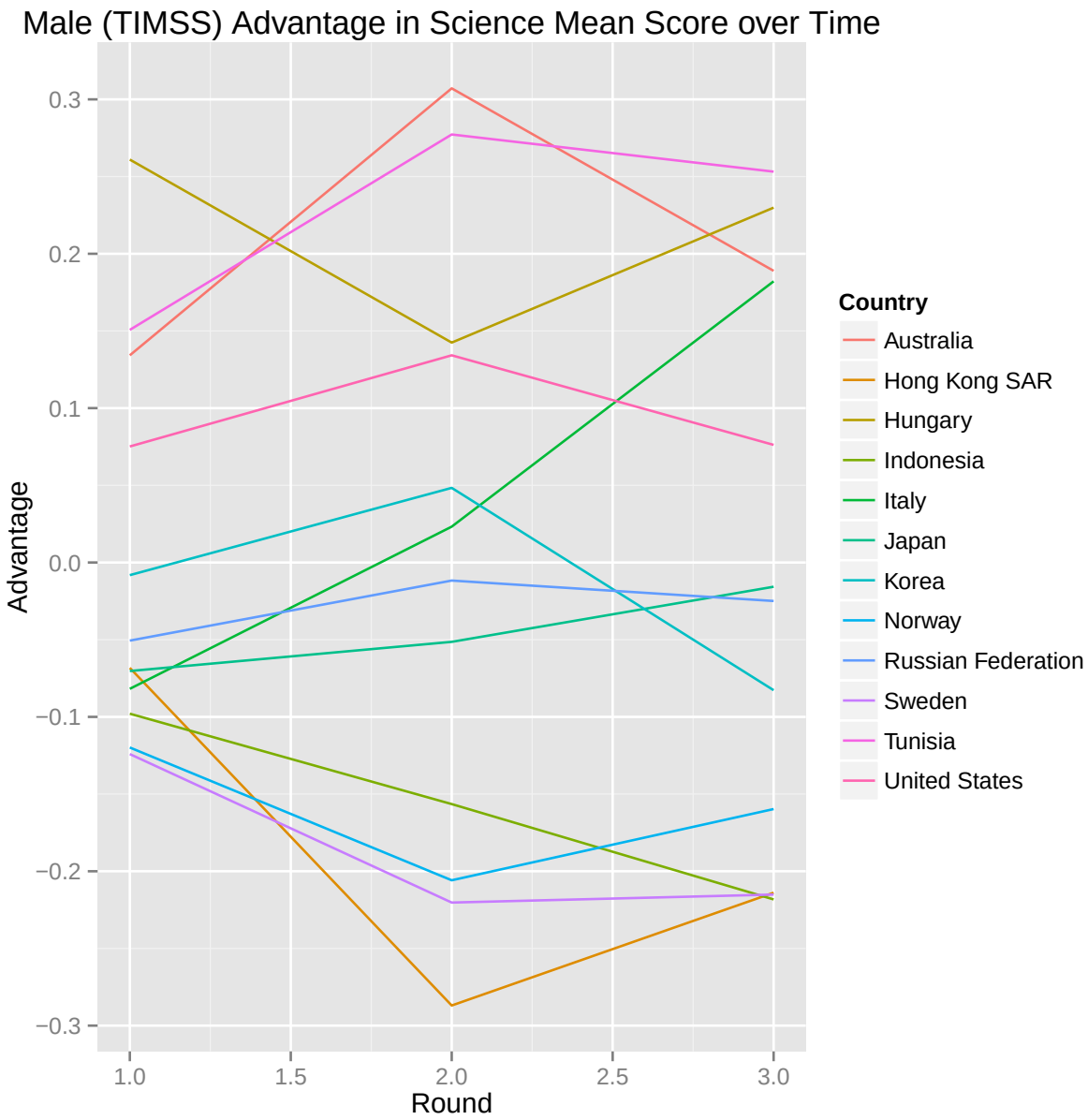
following could could be run:

```
source("countryData.R")
plotCorrelations("country_data_3.csv", c(1, 2, 3), "Gender", "Science", "Science",
    "Mean.PISA", "Mean.PISA")
```




This again does not give us detail on the countries, so we can run the plotAdvantage function again, this time using gender as the dimension to plot by. The parameters in this case would be the input file, a vector of the rounds, the dimension to plot by (Gender, in this case), the subject (Math or Science), and the statistic (Mean). There is currently no data for standard deviation by subject in this project, it could be added as an enhancement. Here again is mean differences for science by gender.

```
source("countryData.R")
plotAdvantage("country_data_3.csv", c(1, 2, 3), "Gender", "Science", "Mean")
```



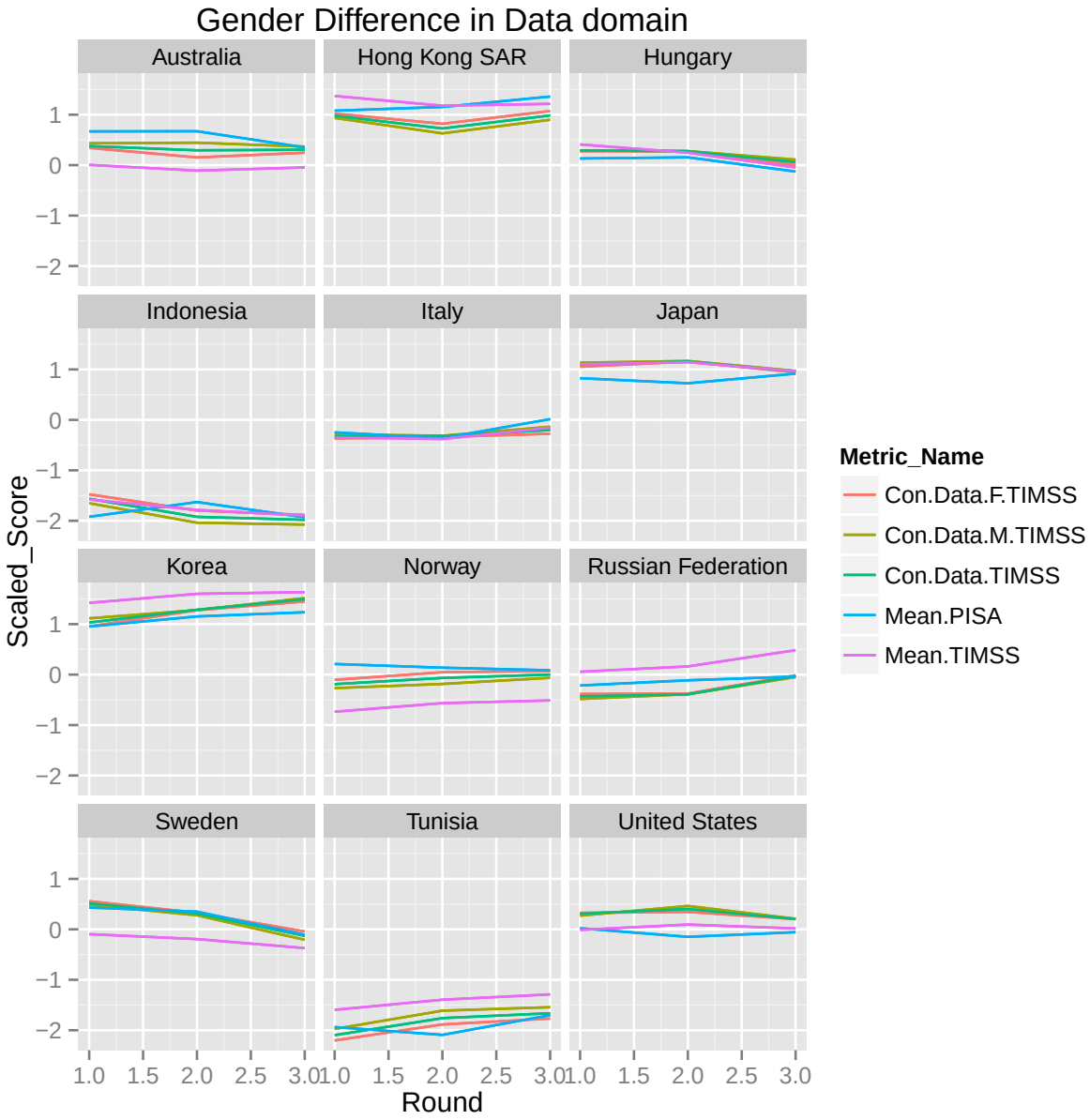

Looking at the first plot, we can see that in general, there is very little difference in scores between males and females for PISA. At the top, there seems to be more variation, but it is modest at best. We can also see no significant differences between eastern and western countries in terms on gender performance.

Next, the second plot allows us to see countries in more detail. The gap for TIMSS seems to be widening over time. In this case, however, the gap does not seem to be due to just a few outliers, but rather a real trend. What this implies is that difference between male and female scores is widening, not necessarily in the same direction, as some countries have seen a relative improvement in female scores and others in male scores, but overall, in countries where males do better, they are doing even better recently, and in countries where females do better, they are also doing better recently then back in round 1. Comparing this to the first plot, it looks as though PISA has seen relatively steady gender performance over time, but the same could not be said for TIMSS.

Lastly, we may want to see if there is any difference in score for content areas by gender. We can use the plotMathContent function to do that, with gender set as the dimension to plot by. The parameters here would be the input file, a vector of the desired rounds, and the dimension to plot by (Gender Content, in this case), and the content area of interest (Num, Alg, Geom, or Data). For example, we can plot gender

differences for the data content area:

```
source("countryData.R")
plotMathContent("country_data_3.csv", c(1, 2, 3), "Gender Content", "Data")
```




For the Data content area, there do not seem to be many significant gender differences. Males in Tunisia and Australia tend to do a little better while females in Indonesia and Norway tend to do a little better. Some countries, like Italy and Hungary see no difference at all. Again, further analysis could yield insights into these effects. As they seem consistent over the three rounds, it seems unlikely that they are due simply to chance.

## Producing the plotting code

The plotting code has evolved over time. At the start, a different function was created for each plot, and different munging functions were written to get the data into exactly the right form for each plot. A

refactoring effort was undertaken to clean up the code and remove duplication. The result is that the plots are now created with a single munging function, and there are only 4 separate plotting functions that create all of the plot types seen above. The lines of code was also reduces by more than half, making it easier to maintain in the future. In addition, several helper functions were written. The first was the load in the data, another to generate column names as part of the munging process, and two more to reshape data and standardize the scores. The last, called getMatchedCountries, was written to return a list of countries that participated in all of the rounds passed in. This is probably the most important helper function as the scores need to be standardized only based on the countries participating in all exams of interest, in order for the comparison to be valid.

The plotting functions were refactored to allow plotting based on several dimensions, within the same function. Each is unique and the explanation of how to use each one, along with examples, is detailed above. One large gain in refactoring was the realization that the data was already in the appropriate form for using the ggplot facet capabilities, which allowed the production of small multiple plots, without the gridExtra package that was used intially. This change had two main benefits - significantly less code to maintain, along with the elimination of several loops, and a common axis, legend and plot size came for free, with no additional work required.

The code for all of these functions can be found in the file at this location: https://github.com/bwelsh/edav/blob/gh-pages/assets/project/r_ scripts/countryData.R.

# Visualization

Since one of the factors being explored was geographic, it seemed natural to create a map for visualization. The vision was to create a world map in D3 and color the countries based on their performance. The data was broken down by round, subject, statistic, gender, and in some cases content area, so it seemed possible to make all of these factors available, to allow the user to select a combination of them and then see the results. Since this would not allow someone to necessarily see country-specific details, an additional piece would be to add a feature to the map such that when a country is clicked, an additional graph or set of graphs would appear below the map, with country-specific data.

## Munging the Data

The data needed to be in a multi-dimensional object for data binding, doing this within R proved difficult, so an R script, countryMap.R, was written to export the data into a series of csvs. A Python script, project_ data_ munge.py, was then written to read in the csvs and then output a json file. The scripts, csvs and json file are located in this folder: https://github.com/bwelsh/edav/tree/gh-pages/assets/project/data_ munge. The scripts were rather hastily written, and so are not of great quality, but they do work. In order to match the country names in the csvs to the country codes that would be needed in the d3 visualization, a Python package, pycountry, was used and proved very helpful in the translation.

## Creating the Visualization

The final visualization is located here: http://bwelsh.github.io/edav/assets/project/d3worldmap.html. It is comprised of several components.

### Map Controls

The map controls are a series of lists of text that represent the different categories that the map can filter based on. For example, the default is to show a comparison of the two tests, filtered by countries that participated in all 3 rounds, for Round 1, in math, showing the mean for both genders and all content areas. Each text element has a class. When a text element is clicked on, a handler changes the class of that item to 'selected', and the classes of all of the other items in that list to 'unselected'. The handler then calls two additional functions, one to update the data on the map to show the data for the new filters, and another to update the detail charts below the map. Not all filters apply to all categories, so an improvement here

could be to dynamically update the list of options upon clicking on one, but for now, if the filter selections result in no data to display, a blank map is shown.

### The Map

The map was created using a json file found here: https://bitbucket.org/john2x/d3test/src/2ce4dd511244/d3/examples/data/countries.json. Antarctica was removed and then the file was read in using the d3 function for reading in json files. The map projection is based on this example: http://bl.ocks.org/mbostock/2869760. Next, the score data json file created from the Python script is read in, and all data, the coordinate data from the countries file and the score data from the Python script output, was then bound to the path elements for each country. This way, data does not need to be re-bound at any point. Additionally, events were attached to each path to show/hide a tooltip with the country name and score on it, as well as to display the country specific charts when a country is clicked on.

The colors were chosen to be diverging, so that darker colors are displayed the farther away the country is from the average. The colors are interpolated using the HCL color palette. The context of the map at any given time depends on the filters selected, and is messaged to the user below the map. For the default, the map is showing relative performance of the countries that participated in all rounds for mean scores in mathematics. The darker the purple, the better the country performed (relatively) on PISA, and the darker the orange, the better the country performed (relatively) on TIMSS. We can see that for Round 1, the pattern seen, that countries in the west perform relatively better on PISA, does seem to hold. By then clicking through the rounds, we see that for Round 2 and Round 3, the pattern also does seem to hold somewhat, although colors seem to get lighter overall, indicating that the difference is not as pronounced for later rounds. A different color palette is used when viewing performance on only one test, in order to visually cue the user that the context is different. Here, blue signifies a higher score and red a lower score. The legend displays this change as well.

### Legend and Messaging

In order to help give context to the colors, a legend was added. This legend is actually composed of two gradients, one for the negative colors and the other for the positive colors, placed next to each other. When the filters are changes, the colors and text in the legend are changed as well, to match the context of the new filters.

Additionally, because there are so many categories to filter on, a message was added in between the map and detail graphs in order to help the user understand what is being displayed. This was done by adding svg text elements with the appropriate text in them. A helper function was written to get the currently filtered items and create a string for the text to display.

### Detail Charts

Two detail charts were created and can be accessed by clicking on a country. The left chart shows the performance of the selected country over the rounds it participated in, for the filters selected. For the default filters, we can see by clicking on Australia, that over time, Australia's performance has actually gotten relatively better on PISA, although not by much. The line graph was adapted from this code: http://bl.ocks.org/benjchristensen/2579599. A function takes the data for the country, gets the filters selected, and returns the scores in an array for those filters. The graph then plots those values, as well as labels for the axis and chart. If no data is available, the chart will be blank.

The right chart is a bar chart to display the gender differences between the different subjects. Again, if we look at the default filters for Australia, we can see that Australia performed relatively better on PISA in both math and science, but the difference was more pronounced for math. In both subjects, females had a higher differential, meaning that their scores were relatively better on PISA and the males scores were relatively better on TIMSS (compared to female scores in Australia, not the scores as a whole). This chart was created using the columnChart library located here: http://bl.ocks.org/llad/3766585. Two minor modifications were made to the library to allow the color to be controlled more granularly, and to make the scales better fit the data.

## Conclusion

The map mirrors a lot of what was seen in the R graphs, but more importantly, allows for additional exploration along every dimension. It is not possible to explore every facet here, but this provides a starting point. In its current state, additional rounds could very easily be added to the map. In 2015, both exams will again be given in the same year, which will provide even more data to explore and learn from.