Policy Research Working Paper 6536

An Expansion of a Global Data Set on Educational Quality

A Focus on Achievement in Developing Countries

Noam Angrist
Harry Anthony Patrinos
Martin Schlotter

The World Bank
Human Development Network
Education Unit
July 2013

Abstract

This paper assembles a panel data set that measures cognitive achievement for 128 countries around the world from 1965 to 2010 in 5-year intervals. The data set is constructed from international achievement tests, such as the Programme for International Student Assessment and the Trends in International Mathematics and Science Study, which have become increasingly available since the late 1990s. These international assessments are linked to regional ones, such as the South and Eastern African Consortium for Monitoring of Educational Quality, the Programme d'Analyse des Systemes Educatifs de la Confemen, and the Laboratorio Latinoamericano de Evaluacion de la Calidad de la Educacion, in order to produce one of the first globally comparable data sets on student achievement. In particular, the data set is one of the first to include achievement in developing countries, including 29 African countries and 19 Latin American countries. The paper also provides a first attempt at using the data set to identify causal factors that boost achievement. The results show that key drivers of global achievement are civil rights and economic freedom across all countries, and democracy and economic freedom in a subset of African and Latin American countries.

This paper is a product of the Education Unit, Human Development Network. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at hpatrinos@worldbank.org.

An Expansion of a Global Data Set on Educational Quality: A Focus on Achievement in Developing Countries

Noam Angrist, Harry Anthony Patrinos and Martin Schlotter[*]

JEL Classifications: I2, J24, O15

Keywords: Quality, Education, International, Achievement, PISA, TIMSS

Sector Board: Education

[*] The authors are associated with the World Bank, the University of Munich, and the Massachusetts Institute of Technology. The views expressed here are those of the authors and should not be attributed to the World Bank Group or to the author's institutions.

1. Introduction

A country's education level is of huge importance to its economic success. Indeed, the economic literature suggests that differences in human capital endowment among countries are largely responsible for huge economic development gaps observed between industrialized nations and developing countries. For a long time, most authors explained growth differences using quantitative indicators such as years of schooling or enrollment rates in primary and secondary schools (for example, Barro 1991; Mankiw et al. 1992). However, recent evidence has shown a quite different pattern: It is not the time spent in school that matters most, but rather what is effectively learned. Thus, qualitative skills acquired during schooling play a decisive role in influencing a country's growth (see Hanushek and Woessmann 2008 for an overview).

This new insight comes at the same time as a large increase in availability of international student achievement tests. These tests, carried out by institutions such as the OECD and the International Association for the Evaluation of Educational Achievement (IEA), measure student cognitive skills around the world. Several econometric studies show that the qualitative indicators measured by these international achievement tests explain growth patterns significantly more than quantitative indicators, such as school enrollment (see Hanushek and Woessmann 2008 for an overview). Moreover, recent analyses reveal a direct and persistent association between cognitive skills and economic growth even controlling for unobserved country differences, which might otherwise be the driving factor for both. Indeed, Hanushek and Woessmann (2009a) use an instrumental variables approach and a difference-in-differences methodology to demonstrate a causal chain between a nation's stock of cognitive skills and its economic growth.

This evidence motivates the identification of factors that enhance the stock of cognitive skills, which in turn drive country growth. The most common tool in such analysis is the estimation of education production functions which include a host of input factors, such as individual characteristics, family background, school inputs (e.g. class size), and systemic elements (e.g. accountability). These input factors drive an output, for example, educational success. In our case, educational success is measured by the stock of cognitive skills (see Hanushek 1979 for an overview).

Some of the more recent economic literature makes an attempt to examine the effect of these input factors on the educational outcomes. While results from these studies vary, systemic effects seem to matter hugely: Several studies, mostly using data from PISA and TIMSS, reveal large and positive effects of system elements on cognitive skills. Some of these key system elements include increased school autonomy (see Fuchs and Woessmann 2007), effective accountability systems (see Juerges et al. 2005), less stratified school systems (see Hanushek and Woessmann 2006) and competition between privately and publicly operated schools (see West and Woessmann 2010). These insights provide a first hint at successful education policies that could improve the precarious economic situations of many developing countries. Yet, the existing evidence has several shortcomings, calling these policy implications into question.

The biggest shortcoming is a lack of consistent and comparable data on education quality across countries, tests and over time. In particular, many studies have relied on only cross-country comparisons, which ignore how educational systems vary over time. Many studies have also relied on the fact that international achievement tests are highly correlated (Rindermann and Ceci 2009). While it is true that international achievement tests such as PISA and TIMSS produce similar results, it is important to adjust for differences in rigor and scaling among various different international tests. Finally, much of the current literature relies only on international achievement tests, which often do not include developing countries, and thus the implications of these studies are limited for the countries that demand the most educational reform.

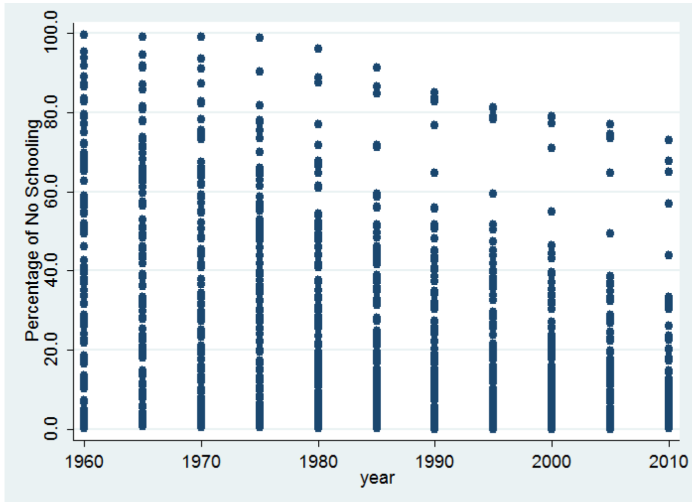
In this paper, we build on an approach taken by Altinok and Murseli (2007) that addresses many of these limitations in two ways. First, we use a methodology that allows us to include developing countries by making regional assessments comparable to international ones. Indeed, while many developing countries do not participate in international tests such as PISA and TIMSS, they do participate in regional assessments which if made comparable to other assessments, would provide insight into achievement in developing regions. For example, many Latin American countries participate in the UNESCO Laboratorio Latinoamericano para la Evaluacion de la Calidad de la Educacion (LLECE) and many African countries participate in the South and Eastern African Consortium for Monitoring Educational Quality (SACMEQ). Second, we link different tests by fixing them to the cognitive performance of United States on achievement tests, since the United States has participated in almost all international assessments since they began and is thus a good reference point. This allows us to incorporate a time dimension into our analysis of educational quality as well as create a uniform database of international achievement.

In particular, a massive database of 128 countries is developed that includes over 40 countries from the developing world. Our dataset includes test scores from 1965-2010 in five-year steps. Our main approach is to extend a data set created by Altinok and Murseli (2007) that makes test scores comparable across various international and regional achievement tests. To this end, we link regional tests to international ones by using countries that participated in both as reference points. Next, we similarly link different international tests by using the United States, which has participated in each for the past half century, as an anchor. Finally, we use the United States National Assessment of Education Progress (NAEP) to conduct a standardized comparison of test scores over time. The database we ultimately produce is an extension of the Altinok and Murseli (2007) database using data from the 2009 PISA survey, and employs pieces from methodologies developed by Altinok and Murseli (2007) and Hanushek and Kimko (2000).

As a next step, we use our database to confirm the insight which first motivated this paper: Although we know that education (Hanushek and Woessmann 2008) leads to country growth, increased school enrollments have not necessarily produced greater learning outcomes. Since ultimately one thing we care about is country growth, and cognitive attainment boosts growth, the lack of impact of increased enrollment rates on learning is concerning.

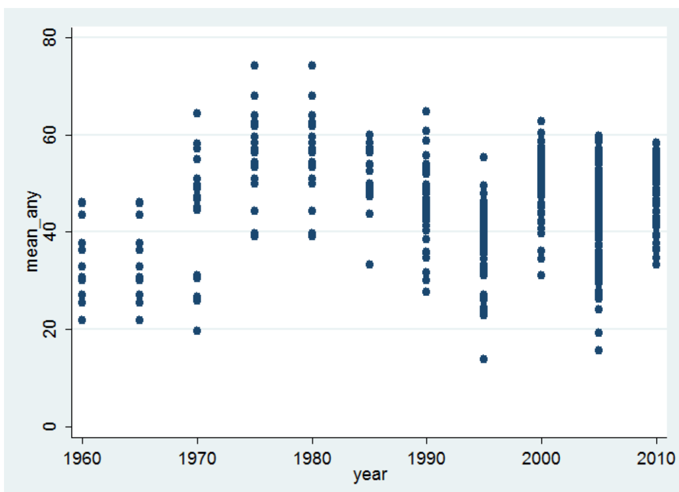
Figures 1.0 and 1.1 uses our internationally comparable data set extended from 1965-2010 shows that even though *overall* trends display intuitive trends – higher rates of schooling align with higher tests score - in recent years, as the “no schooling” rate has continued to plummet, test scores have in fact barely changed at all, and in some countries even dropped.

Figure 1.0: Average no schooling rate scatter plot (1965-2010)



Note: Data comes from Barro lee 2001

Figure 1.1: Average adjusted test scores scatter plot (1965-2010)



Thus, our data confirm that increased schooling is not synonymous with increased educational achievement, prompting an exploration of what does actually produce better learning outcomes. To this end, in this paper we both extend the Altinok and Murseli (2007) dataset as well as make a first attempt using such an internationally comparable dataset to answer this question. We include a host of potential explanatory variables, namely governance, to draw inferences about educational inputs that result in the most effective educational systems.

The paper is structured as follows: In section 2 we explain in detail the methods we use to build our test score database, focusing on advantages and possible shortcomings. Section 3 provides descriptive results of our database and overall trends. Section 4 describes the robustness of our adjusted test score database. Section 5 presents an application of this data set and describes the different econometric methods we use in order to estimate the association between tests scores and explanatory factors. Section 6 includes results from our casual analysis and application of this data set. Section 7 concludes.

2. Methodological Considerations

While far from perfect, outcomes of international student achievement tests are useful measures of educational quality. Among several advantages, international achievement tests allow us to compare achievement gains across countries and thus identify key factors that might be associated with country-by-country variation (see Hanushek and Woessmann 2010). Several earlier studies exploited this unique feature in order to study determinants of achievement such as school autonomy, accountability systems, tracking or the privately operated share of the education system (Hanushek and Kimko, 2000; Barro and Lee, 2001; Hanushek and Woessmann, 2006). As an example, high levels of school autonomy and competition between publicly and privately operated education systems characterized the highest ranked countries on international assessments. Thus, policy reforms favoring these beneficial systemic features seem to boost achievement. Yet, it might be premature to draw conclusions from simple cross-sectional comparisons of countries for two main reasons. First, it is likely that time-varying factors bias these regressions. Second, it is possible that these factors are subject to omitted variable bias and are therefore not linearly connected or causal. For example, a third factor, such as governance indicators, might drive both school autonomy as well as achievement. If we exclude this factor, then it seems as though school autonomy is driving higher test scores where this might in fact not be the case.

Beyond these econometric and methodological shortcomings, even if associations between the systemic features and the cognitive skill measures were causal, results would only be valid for the countries included in the specific samples. Since mostly industrialized nations participate in international achievement tests, these findings are less relevant for developing countries. This is an issue since these poorly performing countries demand the most rigorous and effective interventions. In particular, there exist many unanswered research questions pertaining to education quality in developing countries. For example, while a large gap in economic growth between the industrialized world and developing countries is evident, it is not *a priori* clear whether this is due to differences in human capital endowments or policies and institutions. This underlying difference in a country's educational performance has important implications. One might think that education systems are tremendously underdeveloped in such countries, and thus require fundamental support in the form of basic resources and infrastructure instead of improving specific factors such as school autonomy, accountability or tracking.

In order to address these issues, we build on studies conducted by Hanushek and Woessmann (2009b) as well as Altinok and Murseli (2007) in order to link regional assessments to international assessments. Indeed, while many developing countries do not participate in international tests, Latin America and Africa have at least participated in regional achievement tests carried out during the 1990s and recent years. These tests include the UNESCO Laboratorio Latinoamericano de Evaluacion de la Calidad de la Educacion (LLECE) and the Segundo Estudio Regional Comparativo y Explicativo (SERCE), which test students in third, fourth and sixth grades in a set of Latin American and Caribbean countries. Two tests with a focus on Africa include the South and Eastern African Consortium for Monitoring of Educational Quality (SACMEQ) and the Programme d'Analyse des Systemes Educatifs de la Confemen (PASEC). Specifically, SACMEQ conducted two surveys for South and Eastern African countries for third and fourth grade, and PASEC carried out two waves of testing for second and fifth graders in Francophone Africa.

We effectively utilize all of these tests by making the achievement scores comparable. To this end, we link the results of regional tests – LLECE, SERCE, PASEC and SACMEQ – to international tests such as PISA, TIMSS and PIRLS. As mentioned earlier, in order to normalize achievement test results across tests and time we mainly build on the previous work of Altinok and Murseli (2007) and their attempt to build an international database on human capital quality. In particular we extend their results from 2003 until 2010. In addition we refer to Hanushek and Kimko (2000) who tried to construct a database that manages to gather results of several student achievement tests of different countries in different years on a common scale.

Our approach first builds on Hanushek and Kimko (2000). We exploit the availability of a United States test score in all international achievement surveys conducted since the early 1960s. Therefore, we can express each country's performance in relation to the US result in a given test in a given year. Thus, US tests scores are a reference point, making country achievement comparable across tests. Furthermore, the national testing regime of the US allows for a comparison of test results over time: The almost biannually conducted National Assessment of Educational Progress (NAEP) yields comparable results of US-student achievement (in different subjects and grades) over time. Connecting these results to the most adjacent US score in the international achievement tests delivers comparable US results over time. This adjusted score can then be related to the results of all other countries that have participated in international achievement tests.

While this is a valid methodology, such an approach has limitations. One particular limitation is that this approach ignores all surveys without any United States test score availability, including those regional tests mentioned above. To deal with this, Altinok and Murseli (2007) use a new approach that exploits the appearance of a few countries in both international and regional achievement tests. These so-called *doubloon countries* help to relate regional tests to international tests (Altinok and Murseli, 2007). In a first step they compute the average result of a group of *doubloon countries* in a specific grade in a specific subject in a regional test. The following expression models this first step:

$$\overline{X}^{g}_{s,r,y,c_n} = \frac{X^{g}_{s,r,y,c_1} + X^{g}_{s,r,y,c_2} + \ldots + X^{g}_{s,r,y,c_n}}{n} \qquad (1)$$

where g is the grade level, s is the subject (math, reading or science), r is the specific regional test in which the US did not participate (for example from LLECE or SERCE), y is the year in which the test was taken, and c_n is the specific country which participated in a specific test.

We also compute the average performance of these *doubloon countries* in the same subject in a given test i, in which US performance is available (for example TIMSS).

$$\overline{X}^{g}_{s,i,y,c_n} = \frac{X^{g}_{s,i,y,c_1} + X^{g}_{s,i,y,c_2} + \ldots + X^{g}_{s,i,y,c_n}}{n} \qquad (2)$$

Next, we build a quotient of these two values to yield an index for the relation between the regional test *r* (without US participation) and the international test *i* (with US participation):

$$Index_s = \frac{\bar{X}^{g}_{s,i,y,c_n}}{\bar{X}^{g}_{s,r,y,c_n}} \qquad (3)$$

This index adjusts for two factors: First, this index will allow us to account for the varying scales of the tests; second, this index accounts for varying difficulty among different tests. Therefore this index reliably enables us to compare different tests across various countries.

It is however important to note that the regional test might measure a different grade and be administered in a different year than an international test. For example, the regional SERCE test is specific to grade 6, while the international TIMSS test might be specific to grade 8. Furthermore the SERCE test was conducted in 2006 while the TIMSS test was conducted in 2007. Therefore, while the mean score calculated for all countries that took a regional test such as SERCE in 2006 (equation 2) is unbiased, when we divide the SERCE 2006 mean by the TIMSS 2007 mean, we might be concerned about the integrity of the index. This potential bias, however, does not seriously affect the outcome of our methodology for two important reasons. First, we use the index to translate all of the original scores and since the same index is used for all of the original scores, then all scores are transformed equally. Second, it is unlikely that tests changed between years in a way that differentially affected certain countries, thus eliminating the concern of a potential bias in our index. For example, even if TIMSS 2007 was made more challenging as a result of 2006 SERCE test scores, which is highly unlikely to begin with, this change should not impact Colombia more than Bolivia. Thus, the index we produce can be a powerful and unbiased tool to link international achievement tests with regional tests.

Finally, we use this index to generate meaningful and comparable test scores. To this end we multiply our index by the regional test scores for those countries who did not participate in any test with a US comparison:

$$\hat{X}_{s,i,c_n} = \bar{X}^{g}_{s,r,c_n} \times Index_s \qquad (4)$$

Thus the test score from a regional achievement test has been converted to X^{g}_{s,r,y,c_n}, a score that is comparable to $\hat{X}^{g}_{s,i,c_n}$, an international test result with US participation. These test scores allow the inclusion of developing countries, which participate only in regional assessments, to be included in our international achievement data set.

Next, to effectively compare various types of international assessments across countries we adjust all achievement test scores in relation to the US. To this end, we construct a similar index to the one above where we create a ratio between US scores on the NAEP and US scores on international achievement tests for a given subject in the most adjacent year.

$$Index_{s,y,US} = \frac{X^{g}_{s,NAEP,y,USA}}{X^{g}_{s,i,y,USA}} \times 10 \qquad (5)$$

We then multiply our raw tests scores and doubloon country tests scores from equation (4) by our new index to obtain a dataset of test scores that is linked to the US and can be compared over time.

$$z = \hat{X}_{s,i,c_n} \times Index_{s,y,US} \qquad (6)$$

where z is an internationally comparable test score over time for a group of 128 countries, including developing countries.

While this methodology generates comparable scores, like all other adjustments methods, this method has its limitations. First, our transformation of regional scores into an internationally comparable value is more accurate the more *doubloon countries* are available. If the Index just relies on the relation of one country`s international score to its score in any regional study (just because it is the only country participating in both surveys) it is quite ambitious to convert all other regional scores using this quotient. As we know that *doubloon countries* are small in number, this problem is relevant for our approach.

Second, this approach refrains from adjusting a joint standard deviation over all tests. So, although anchoring test scores allows us to match different test results over different surveys and over time, we cannot say exactly by how much each country improved. For example, we might know that *country a* outperforms *country b* by 20 (adjusted) points in year x, and that it has increased its average test score level by about 40 (adjusted) points in year x+3. Now it outperforms *country b by 30 points*. So, *country a* has done better in both years than *country b* and it has actually improved over time. But we cannot definitely specify the scale of the improvement. For example, a 20 point difference can mean to have a larger knowledge gap in year x than 30 points in year x+3. This can depend on the countries that participated in the survey or the contents of the tests.

Our final database consists in the end of 128 countries where we have at least any test score at the country level within the period 1965-2010.

3. The Database of Adjusted Test Scores

Our database, which aggregates test scores across regions and tests over time, is constructed as a quasi-panel in five-year steps. While it would be ideal to have a test score for all countries for every year since 1960, test frequency is too low. Following Altinok and Murseli (2007), we provide a subject (Math, Reading and Science) and grade level-specific (Primary or Secondary) test score for every five-year period. If countries participated in several comparable tests in or around a specific year, we built the average over the results in the respective surveys. For example, a country's math score in secondary school in the year 2000 follows from its (adjusted) PISA 2000 and (adjusted) TIMSS 1999 eighth grader result if the country took part in both surveys. If just one adjusted test score is available for the country (either from TIMSS 1999 or PISA 2000), this single result is used as the country`s secondary math score in the year 2000. Or, the countries' 1995 math score on primary school level is a combination of LLECE scores (adjusted as described in section 2), Measurement of Learning Achievement (MLA) and TIMSS 1995 results.

We group test scores into five-year steps for a few reasons. First, we often have test scores that are comparable by subject and grade level, yet were administered one or two years apart. Therefore, unless we align our scores by year, we will not be able to linearly regress our explanatory variables on our outcome variables. So, we must focus on years that can be included in our analysis and group adjacent years into them. Second, there exist unequal distributions of time where tests might not have been administered, and so five-year steps of data allows us to maximize continuity of test scores from 1965-2010. Third, we need equal steps since if we have a seven-year jump between test scores followed by a three year jump, then our explanatory variables might explain a time gap in learning instead of specific determinants of achievement. Indeed, we assume that four more years of schooling will boost achievement. One particular transformation to note occurs during our extension of the Altinok and Murseli data from 2003 to 2010. Since we have data from 2003 and TIMSS/SERCE/PISA data from 2006 and 2007 we average these results and group them into the year 2005. Further, we group our adjusted PISA 2009 scores into the year 2010 in order to be compatible with the remaining adjusted test scores which occur in five-year steps.

Below, we highlight a few descriptive results on primary test scores in mathematics to showcase our database. In particular, we stratify our results by region and income level in order to present a coherent picture of overall achievement trends. Figure 2.0 and 2.1 describes test score availability by for primary math scores from 1985-2010 by region and income level, respectively.

Figure 2.0: Presence of primary math test scores by region (1985-2010)

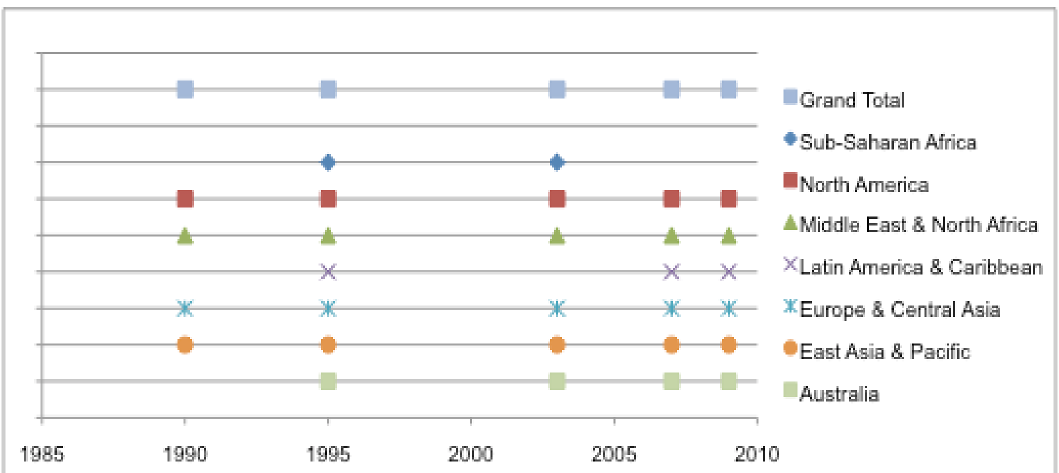
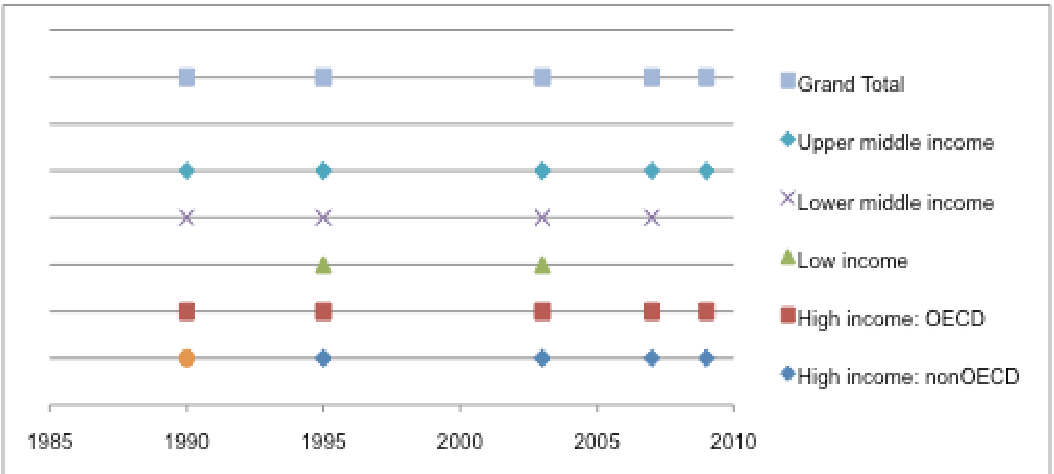


Figure 2.1: Presence of primary math test scores by income level (1985-2010)



The results from Figures 2.0 and 2.1 demonstrate that by creating an internationally comparable test score database, we have managed to obtain coverage even for developing and low-income countries, although data on these countries still remain scarcer than the more developed countries.

Next we use primary math scores to highlight achievement trends using our adjusted test score database. Figures 2.2 and 2.3 showcase the average adjusted test scores by region and income level, respectively. We further include a metric for the overall average adjusted test score in each year so as to determine which countries are performing well by world standards.

Figure 2.2: Average primary math tests core by region (1965-2010)

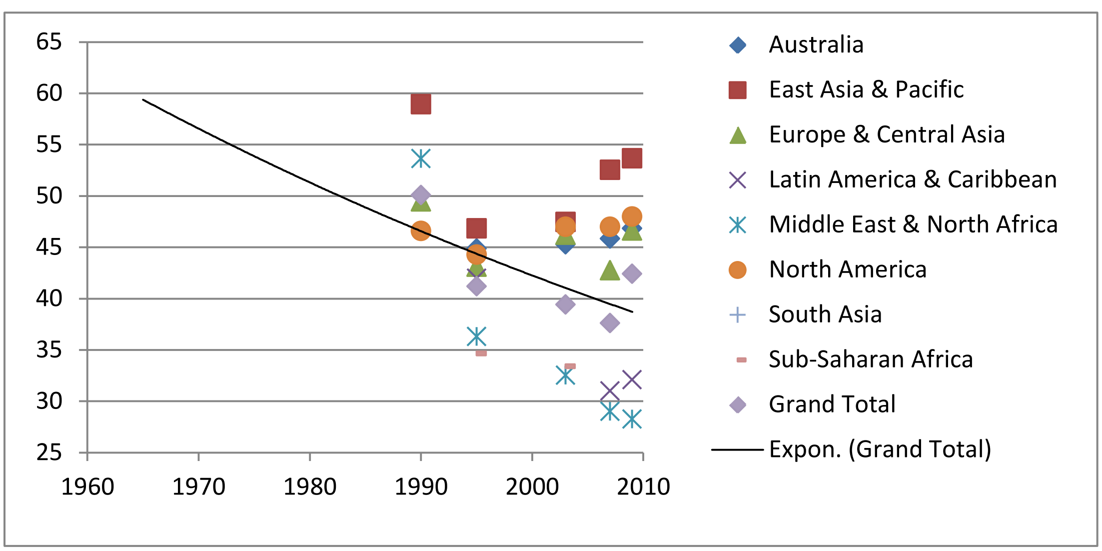
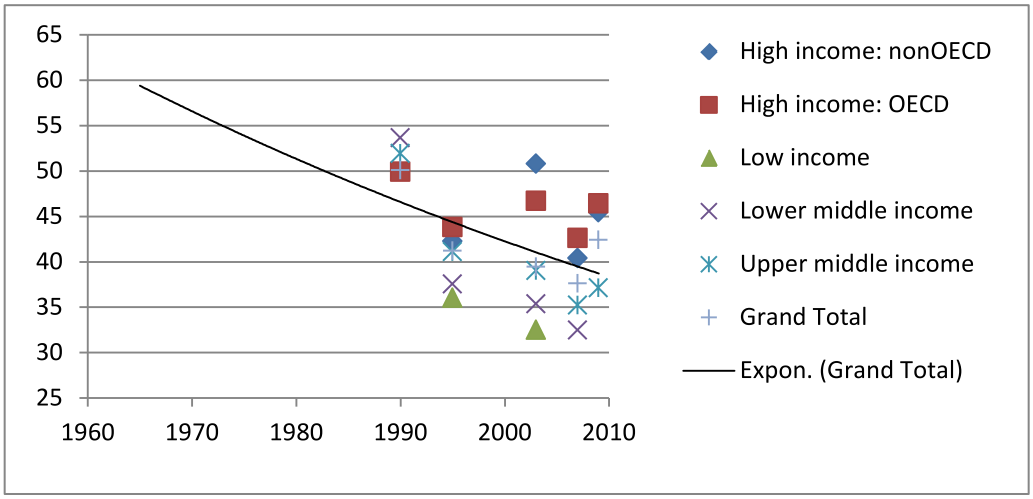


Figure 2.3: Average primary math tests core by income level (1965-2010)



Figures 2.2 and 2.3 reveal that just as developing countries lack data on test scores, they also perform significantly worse by world standards. Two obvious test score champions include the East Asia and Pacific region, as well as High income OECD countries.

This breakdown of results based on our database showcases the ability of an internationally comparable dataset which extends from 1965-2010 to uncover important learning trends. Although this dataset enables us to tackle questions related to global achievement, there are limitations to this dataset based on test score availability and assumptions we use. For example, in our final analysis we average test scores over subjects and even over grades in order to get better coverage of countries and time. While several previous studies pool scores over subjects and grades (see Hanushek and Kimko 2000 or Hanushek and Woessmann 2009), we are aware of the limitations and specific assumptions related to such an approach.

Some general patterns can be observed:

1. There is no full coverage over the whole period. While the first test scores are available for the year 1965 and the last ones for 2010, there is no test score for any country in 1975. This reflects both low testing during the 1970s and also specific merging of tests into five-year steps. For example, tests carried out until 1972 are assigned to the 1970 score; tests in later years of that decade are part of the 1980 score.

2. Coverage differs by subjects: While math test scores are already available for a set of countries in the mid-1960s (by the IEA assessment First International Math Study (FIMS), reading and science results are not available until the 1970s (First International Science Study (SISS) and First International Reading Study (FIRS).
3. We have also different coverage by grade: surveys that have assessed students in primary school are much scarcer than assessments carried out in secondary school. There is, for example, no primary math score for any country before the year 1995 (from TIMSS). Similarly, the first reading score for primary school students is available in 1990 (from the Second International Reading Study (SIRS).
4. There exist many gaps by subject. While the first reading assessment took place during the early 1970s (FIRS), there is a 20-year vacancy until 1990 when the Second international Reading Study (SIRS) was conducted. In math, there is also a fifteen-year gap between the 1965 scores and the 1980 results.
5. Coverage by country and, even more apparent, by whole world regions, differs considerably. In fact, African and Latin American countries, have not widely participated in any surveys before the 1990s and the setup of their regional tests such as SACMEQ, PASEC and LLECE. There are some results for single countries of these regions before the 1990s (for example FIRS scores for Chile and Malawi from 1970 or SIMS scores for Swaziland and Nigeria from 1980), but no broader coverage that facilitates carrying out intra-regional comparisons or averaging scores over these regions.

The facts described above can be studied in detail when looking at the graphs provided in the annex. We provide coverage by grade level (for primary school, see Figures A1-A5; for secondary school see Figures B1-B5) for every country that has participated in any test from 1965 until 2010[1]. We also show coverage if scores are averaged over different grade levels (see Figures C1-C5).

The database and its coverage corroborate the need for combining test scores over subjects and perhaps even over grades. While analysis over time by subject and by grade is, if at all, feasible

[1] Coverage by subject is available on request. The coverage between 2003 and 2010 (stemming from countries' participation in PISA 2009, PISA 2006, TIMSS 2007, PIRLS 2006 and SERCE 2006) is reported in the graphs, the adjusted test scores of the Altinok and Murseli (2007) database, however, only covers the period until 2003 (PISA and TIMSS 2003 are the most recent assessment integrated in their overview).

for a set of OECD countries, the inclusion of African, Latin American and Asia countries requires us to average test scores. The limitations of such an approach are obvious: A test score for a country in a specific year can consist of a single secondary school reading assessment (for example Chile`s 1970 test score in Science) which then has to be compared with later results from the same country perhaps resulting from a completely different subject or grade level. Yet, we put up with this drawback in order to get a broader coverage over the years and countries. We run several robustness test using only primary scores or secondary score and do separate analyses by subjects.

Founded on that database, we graph test score trends over time. The scores are adjusted to have a mean of 50 points and a standard deviation of 10 points. Figures D1 – D3 show the results over time for the world regions. D1 provides trends averaging over all subject results on primary level, D2 for secondary level, and D3 for averaging over grade levels and subjects, respectively. Results by subjects are available upon request. These three graphs reveal some of the problems described above. Regional averages (especially for African and Latin America and the Caribbean) are composed of very few countries so the trends are hardly interpretable. However, the level differences between the regions are quite obvious with the developed countries outperforming the rest of the world. Even the catch-up process of Asian countries during the last decades becomes rudimentary visible.

The first insights on a regional level are supplemented by graphs for every single country. Figures E1 – E5 show results for test scores of countries in every world region, averaged both over subjects and grade levels. A clear pattern can't be observed (many countries improved over time, others got worse). One peculiarity is the general increase in performance observed from 1965 to 1970, partly continuing until 1980, for all (mostly industrialized) countries that provide information within this time span. On the one hand, this might reflect the educational expansion over the industrialized world during the 1960s. On the other hand, It could also be due to the fact that all 1965 scores just consist of a single Math test from the First International Math Study on secondary level (FIMS), whereas the test score of 1970 is exclusively averaged over Science and Reading scores (from FISS and SISS in 1970), even including scores from primary school Reading. Longer lasting trends, especially for the Latin American and Caribbean countries requires a future inclusion of the assessments carried out after 2003 (PISA 2009, TIMSS 2007, PISA 2006, PIRLS 2006 and especially SERCE).

It is important to note that given the scarcity of previous data on test scores, our extension of the Altinok and Murseli (2007) data set is significant, especially by allowing for the inclusion of more developing countries.

In addition, since there was an error in the 2006 PISA survey in the United States, and the United States is our reference point for all countries, no reading scores since 2003 from any country were internationally comparable until our inclusion of 2009 United States PISA reading scores in this dataset.

Another key contribution of this data set is the inclusion of more *doubloon countries*, since two new Latin American countries (Panama and Peru) participated in the 2009 PISA Survey. The inclusion of these two new countries expands our sample of doubloon countries by 33 percent up to 8 counties. Since we use the average test scores of all *doubloon countries* within a region to calculate our test score adjustment index for these developing countries (described in section 2), this addition improves the accuracy of our adjusted test scores for developing countries.

Finally, whereas Altinok and Murseli (2007) include data from 2003 and in a recent updated paper (Altinok and De Meuleester 2010) for 2007, we intentionally group our most recent test score data into 5-year steps. We average 2003 and 2007 results into 2005 test scores, and grouping 2009 data into the year 2010. This approach allows us to align our recent adjusted test scores to previous test score intervals in the data set. As discussed in section 2, this generates the most accurate dependent variable of educational outcomes since if our test score steps are uneven we might pick up differences in years of schooling instead of determinants of achievement in our final analysis.

4. Robustness of the Database

In order to get an idea of how accurate our adjusted test score is, we first outline some examples from the adjustment of SERCE data to PISA 2009 and TIMSS 2007 used in our adjustment methodology. While the Altinok/Murseli database just provides data until 2003, we extend the series until 2009 including SERCE 2006, PISA 2006, TIMSS 2007 and PISA 2009. We use the adjusting method described in Section 2 in order to predict PISA 2006 and TIMSS 2007 values for those developing countries that have not participated in the international achievement tests. In a further step these adjusted values are put in relation to the United States values from TIMSS and PISA. For predicting TIMSS 2007 values, we use the two countries El Salvador and Colombia as *Doubloon Countries* because they participated in both surveys (SERCE and TIMSS 2007). For PISA 2006 we have four *Doubloon Counties* (Argentina, Brazil, Chile and Colombia). For PISA 2009 we have eight *Doubloon Countries* (Argentina, Brazil, Chile, Colombia, Mexico, Panama Peru, Uruguay). For TIMSS 2007 8^{th} grade values in math and science we take SERCE values in 6^{th} grade in math and science in order to compute the adjustment Index. For PISA values in math, reading and science we apply the respective SERCE values from 6^{th} grade. For Science 4^{th} grade we do not have an adequate SERCE value as Science is only tested in 6^{th} grade in this study. We also do not have predicted science values for Brazil and Chile in PISA 2006 and 2009 as those countries did not participate in the SERCE science test.

For all these *doubloon countries* we can conduct a robustness check by comparing the predicted values that come out from using the method by Altinok and Murseli with their original TIMSS or PISA value. Such a robustness check ensures that our methodology for standardizing test scores is valid. If our predicted values align with the original PISA or TIMSS score for *doubloon countries* we can be more confident that our predicted scores effectively predict standardized projections of regional test scores for all countries.

Table 1.0 gives an overview across countries on the comparison between the original values from the respective surveys and predicted ones, computed by the Altinok/Murseli method outlined above. We see that predicted scores in both reading and science are within 10 points from their original TIMSS or PISA value. These differences account for less than one tenth of a standard deviation and therefore indicate that we generated relatively accurate predicted scores. We have just one case where the difference between the original value and the predicted one is higher than 20 points. This difference exists for math scores in Colombia.

The difference from our predicted and original scores ranges from 1-23 points for math and from 0-9 for reading and science. This shows a clear pattern that our index more accurately predicts reading and science scores over math scores.

However, even a difference of 23 points for Colombia in math - our largest discrepancy - constitutes just a fifth of a standard deviation taken for all countries adjusted test scores. This indicates that even our larger discrepancies in math are not significant.

In fact, most differences are much lower than our ranges indicate, with some differences netting zero, which is the most accurate result. This holds true even for TIMSS where the adjustment Index is based on just two countries (Colombia and El Salvador) which, as outlined before, usually aggravates the computation of a precise exchange factor. This alleviates our concern that we can only rely on indexes that use a large number of doubloon countries to produce reliable predicted scores. Even our estimates, which use limited doubloon countries, produce reliable results. However, we do benefit from having more *doubloon countries* overall in our dataset than the original Altinok and Murseli dataset, which improves the general accuracy of our index.

Furthermore, studies show that achievement on international tests is highly correlated (Rindermann and Ceci 2009). Thus, by comparing projections for *doubloon countries* that participate in PISA to predictions using TIMSS scores we can verify that our estimates are consistent across achievement tests to bolster our confidence in these predictions.

The results from our robustness check are detailed in Table 1.0 and verify the reliability of our predicted scores.

Table 1.0: Original Test Scores and Adjusted Test Scores from Altinok/Murseli Method

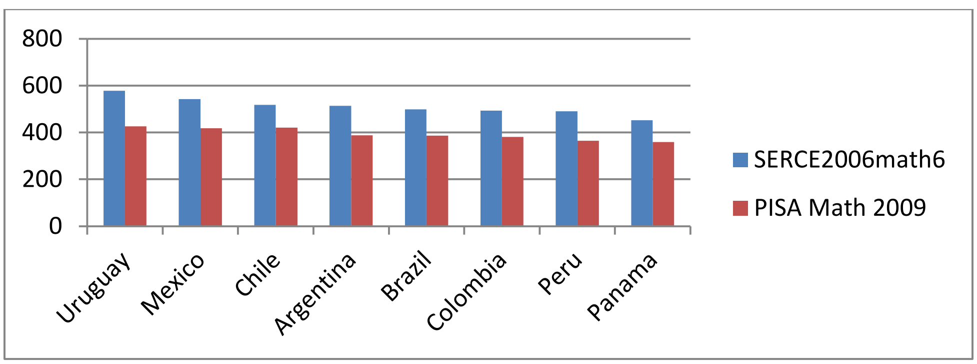
				PISA 2009					
	Math			Reading			Science		
	Original	Predicted	Diff.	Original	Predicted	Diff.	Original	Predicted	Diff.
Argentina	388	**395**	**7**	398	**403**	**5**	401	**392**	**9**
Brazil	386	**384**	**2**	412	**409**	**3**	405		
Chile	421	**398**	**23**	449	**449**	**0**	447		
Colombia	381	**380**	**1**	413	**408**	**5**	402	**404**	**2**
Mexico	419	**417**	**1**	425	**418**	**7**	416		
Panama	360	**348**	**12**	371	**375**	**4**	376	**379**	**3**
Peru	365	**377**	**12**	370	**379**	**9**	369	**373**	**3**
Uruguay	427	**445**	**19**	426	**424**	**2**	427		
				TIMSS 2007					
	Math 8th grade			Math 4th grade			Science 8th grade		
	Original	Predicted	Diff.	Original	Predicted	Diff.	Original	Predicted	Diff.
Colombia	380	**368**	**12**	355	**348**	**7**	417	**412**	**5**
El Salvador	340	**352**	**12**	330	**337**	**7**	387	**392**	**5**

Notes: *Original* shows the effective scores of the countries in PISA 2006, PISA 2009 and/or TIMSS 2007. *Predicted* is the score that we would yield for the countries if we applied the Altinok/Murseli method.

Additionally, we include descriptive data on country achievement as measured by raw scores on PISA 2009 as a robustness check. Our goal is to juxtapose raw scores from PISA in 2009 with our average mean score calculated from the Altinok/Murseli method. We see that our adjusted test scores similarly rank those countries that perform best on the raw PISA scale, validating our conversion method. Figure 3.0 details the relative rank of each country based on their adjusted test scores and raw test scores. The average rank differential is around 3-5, indicating that our adjusted test scores generally simulate the raw data.

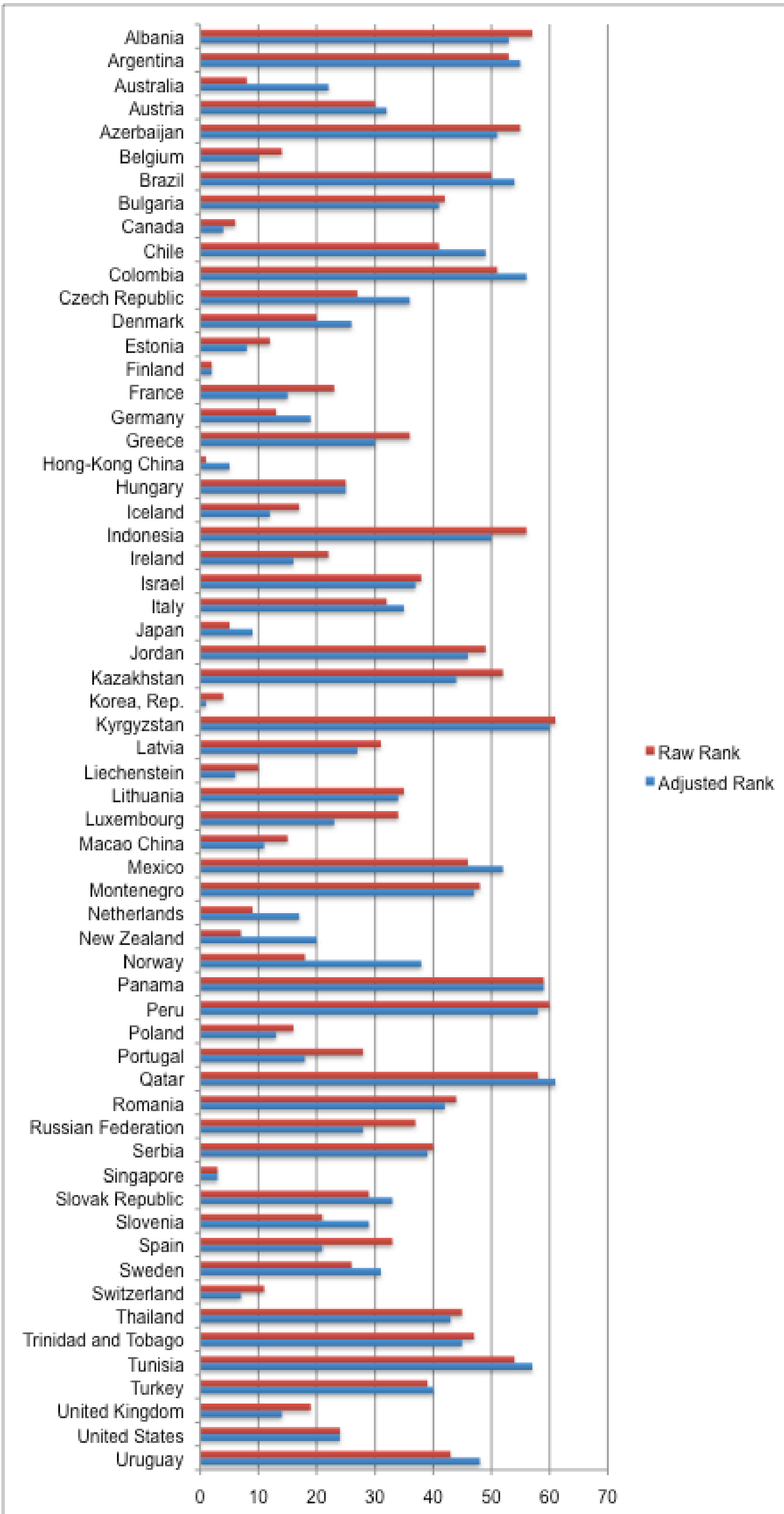
We further include descriptive graphs that focus on test scores in developing countries, in particular, Latin America. Figure 3.0 details the results of this comparison.

Figure 3.0: Average SERCE 2006 Math Scores vs. PISA 2009 Math Scores in Latin America



Notably, these graphs indicate that Latin American test scores on the SERCE and PISA place these countries in ranks that generally track each other – meaning that Latin American countries that perform best on the SERCE exam also perform best on PISA. This is true across years, since the SERCE test was take in 2006 while PISA was take in 2009, and across tests which we use to create our index described in section 2. Thus, the validity of our adjustment mechanism is strengthened.

Figure 3.1: Adjusted Mean Test Score Rank versus Mean PISA Test Score Rank, 2009



5. Descriptive Implications of the Data Set

The extension of our database to developing countries allows us to credibly include low-income countries in global improvement rankings over time. To this end, we conduct an exercise demonstrating the implications of this expanded data set.

First, we compare recent improvements in PISA test score gains between 2006 and 2009. We limit ourselves to this time period since there exist only sparse data on PISA test scores before 2006, making comparisons in other time periods challenging. This limitation motivates the expansion of our dataset.

Next, we compare test score gains using our expanded adjusted test score dataset. Since our adjusted test scores database is both standardized and comprehensive, linking regional test scores to international tests and pooling subjects and grade levels, we can accomplish two things we were unable to using only raw PISA scores. First, we can extend our comparison to a larger time period: 1995-2010. Second, we include additional countries, namely developing countries, in order to rank their learning progress on a global scale.

We start by examining raw PISA score improvements between 2006 and 2009. In Figures 3.2 and 3.3 we notice that the largest improvements in both math and reading are in Qatar, Bulgaria, Kyrgyz Republic and Romania. In math alone, Turkey, Italy, Portugal, Brazil, the United States and Mexico rank near the top. In reading, the top improvements in reading came from Serbia, Israel, Colombia, Argentina, Greece and Tunisia.

We compare these results to Figure 3.4, which details improvements using our adjusted test score database. Using our adjusted tests scores, which include more developing countries and cover a longer horizon, the top global performers include: Jordan; Iceland; Portugal; Canada; Hong Kong SAR, China; Greece; the United Kingdom; New Zealand; and Singapore. There are a few similarities in our comparison between raw PISA scores and our adjusted test score, such Greece and Portugal, our top performers list has largely changed. Thus, by expanding our dataset using standardized metrics, we gain new perspective on learning progress over time and across the globe.

Figure 3.2: PISA Math Improvement (06-09)

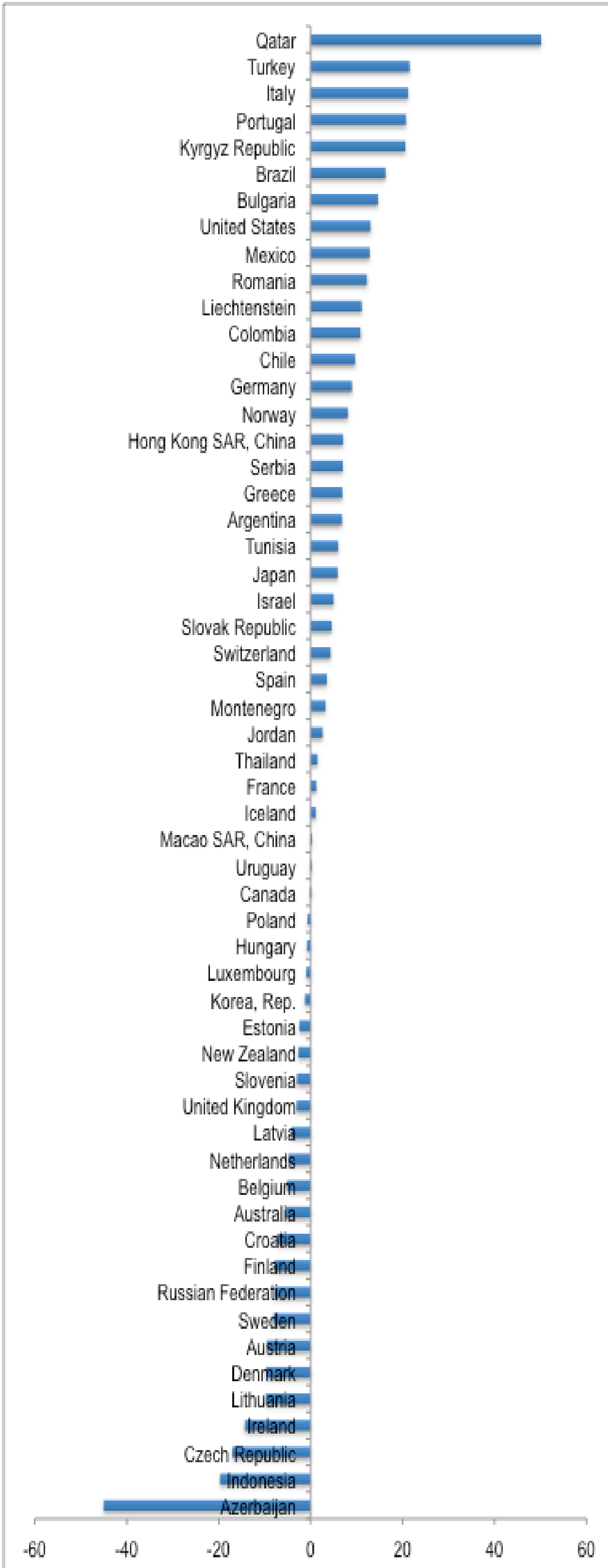


Figure 3.3: PISA Reading Improvement (06-09)

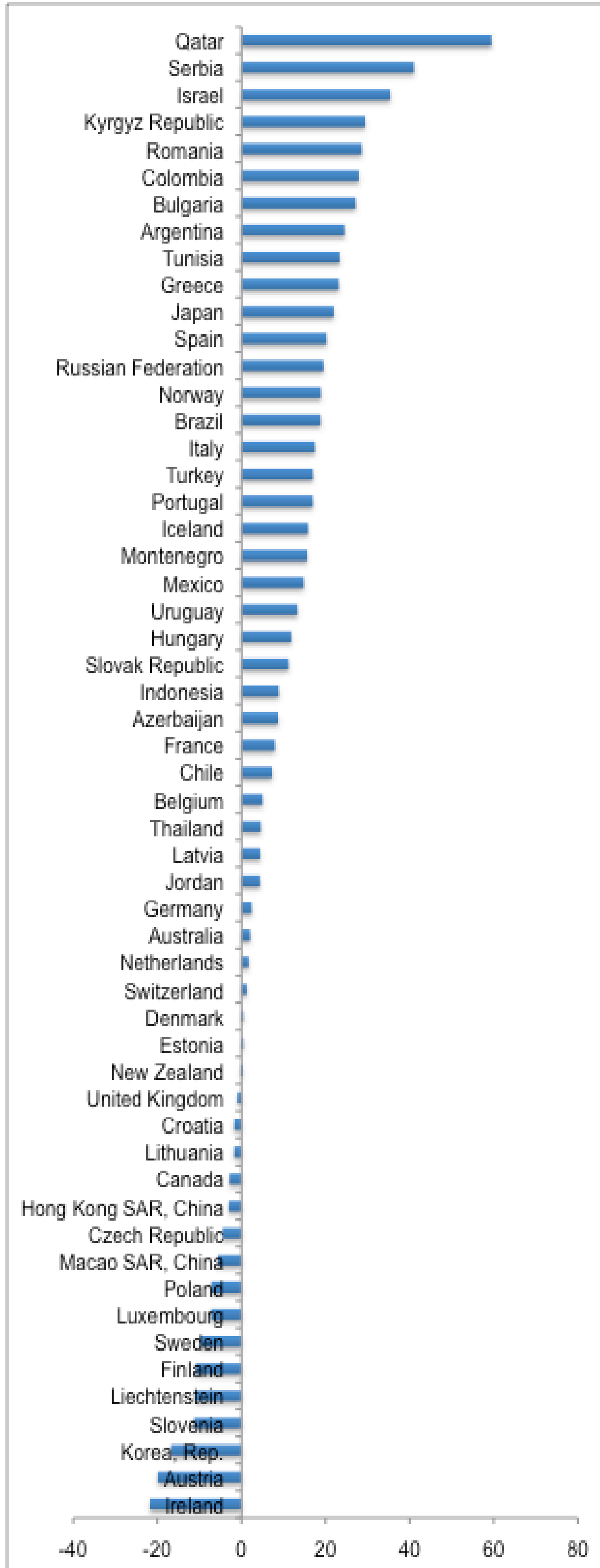
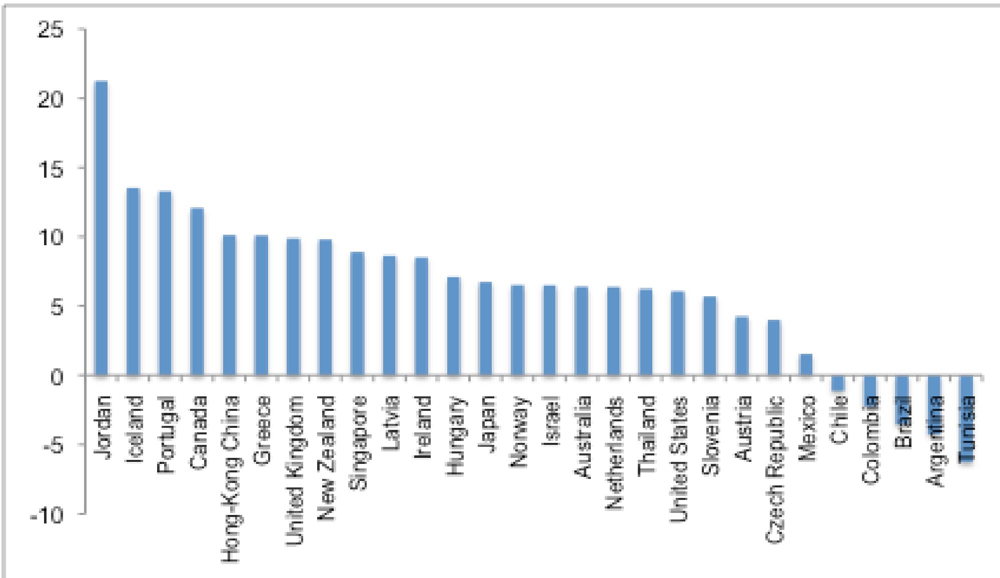


Figure 3.4: Average Improvement in Adjusted Test Scores (1995-2010)



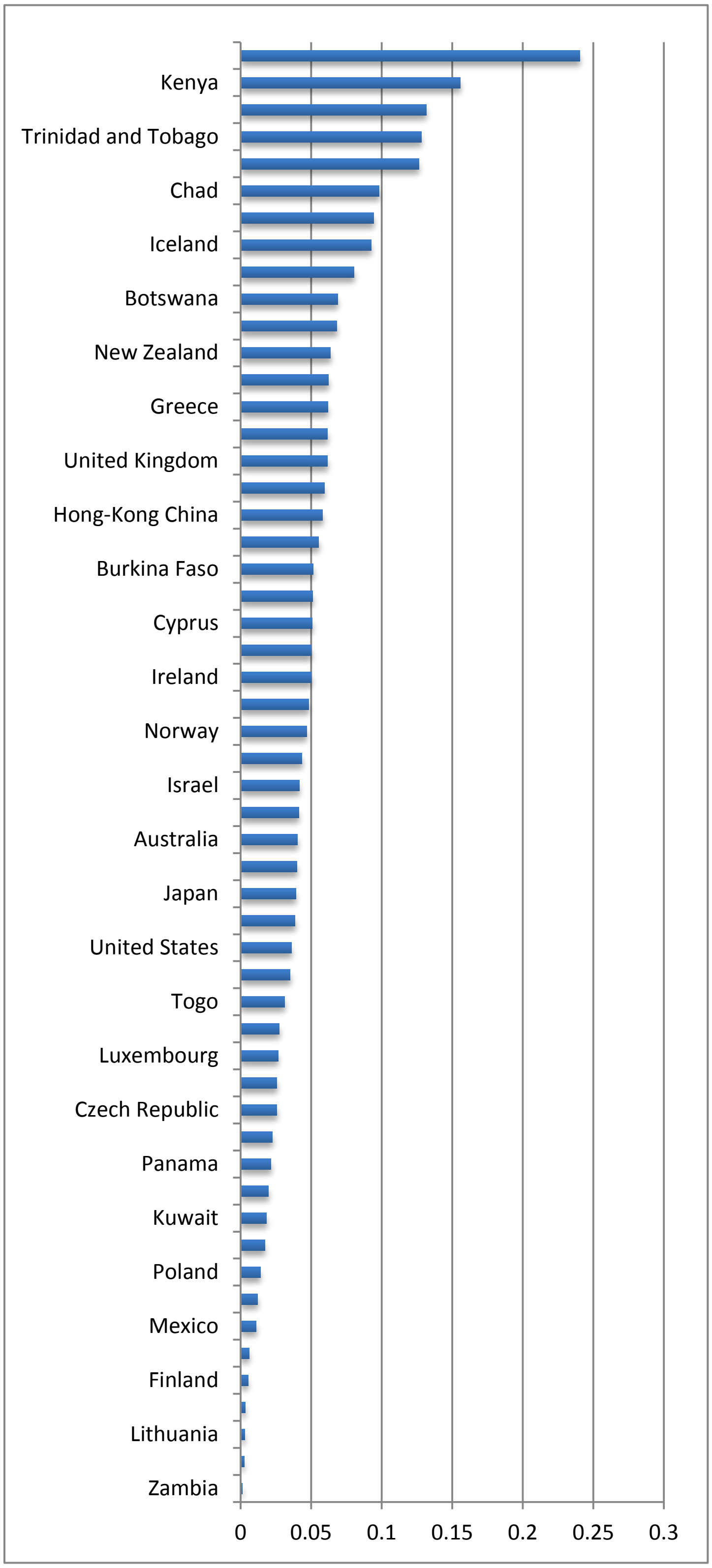
Note: countries with missing data points in either 1995 or 2010 are not included in this graph.

Next, a figure is introduced which aims to create an even more comprehensive ranking of learning progress. Given that developing countries often lack data, even Figure 3.4, which uses adjusted test scores in 2010 and 1995, is sparse. If there is no data point for one of these points, it is impossible to measure improvement. Therefore, an additional metric is constructed by which one can compare learning progress: average annual learning progresses. This metric averages improvements in adjusted test scores each year they are available between 1995 and 2010. This allows me to expand my sample of countries from 28 in Figure 3.4 to 93 countries in Figure 3.5. The results of this comparison is included in Figure 3.5 for only those 54 countries that showed net positive annual average learning improvements.

According to this newest ranking of countries over the last 15 years, we obtain a new list of top performers. The top improvements come from Jordan, Kenya, Madagascar, Trinidad and Tobago, Tanzania, Chad, Namibia, Iceland, Portugal, Botswana, Canada, and New Zealand. All but Portugal, Canada, and New Zealand are new to the list of highest learning progress.

By using our adjusted test score database, we can better inform policy on a standardized scale so that we can accurately determine and target learning trends and include developing countries on this scale in a credible manner.

Figure 3.5: Average % Annual Improvement in Adjusted Test Scores (1995-2010)



Note: countries with missing data points in either 1995 or 2010 are not included in this graph.

6. Application of the Data Set

Our database consists of internationally comparable test scores from 1965-2010 and provides a useful measure of education quality. This outcome measure can be used for an empirical analysis of the determinants of educational performance. Thus, we provide a first example of how to use our extended and updated data set to determine causal inputs in successful education systems.

One major motivation for this analysis stems from the concentration of specific types of countries on both ends of the achievement spectrum. Indeed, most of the countries that perform worse than the world adjusted test score average are concentrated in Africa, Latin America, and the Middle East, and are considered developing countries. This large discrepancy begs the question: Why do some countries achieve better learning outcomes when other countries do not?

To this end, we use our panel data set to demonstrate one possible causal analysis to explain differences in qualitative achievement on international assessments among countries and over time. In particular we focus on governance variables.

In this section, we briefly explain the econometric strategy applied to our data and the set of explanatory variables included in our dataset used to reduce econometric issues arising from omitted variables and other biases. The estimation strategies we want to apply are the following: First, we use a fixed effects approach, capitalizing on the variation in systemic elements over countries and time in order to establish a causal link between such elements and resulting cognitive skills. Second, we control for certain confounding factors such as macroeconomic indicators. Third, we include lagged variables as explanatory factors to see if our causal estimates persist. Our approach can be modeled as follows:

$$Y_{i,t} = \alpha + \beta * X_{i,t} + Z_{i,t} + u_{i,t} \qquad (7)$$

$$Y_{i,t} = \alpha + \beta * X_{i,t-1} + Z_{i,t-1} + u_{i,t-1} \qquad (8)$$

$Y_{i,t}$ is the outcome of interest from our international adjusted test score database, $X_{i,t}$ is the vector of explanatory variables, $Z_{i,t}$ is the vector of covariates, and $u_{i,t}$ is the error term. Our estimator, , provides an estimate of the effects of the different systemic elements of school systems, explanatory variables, and covariates respectively, on the adjusted test score. In addition, we include covariates so as to control for several other potentially confounding between our variables of interest and our outcome. In our case, these potentially confounding factors include several macroeconomic and demographic factors at the country level, such as GDP per capita. In a pure cross sectional approach, several studies have already applied this estimation strategy for sub-samples of countries using TIMSS or PISA data (see for example Woessmann 2003). While results from these types of studies are the starting point of our extended approach here, they have many drawbacks which also apply to possible results from equations (7) and (8). First, they probably all suffer from tremendous omitted variables biases, because country-specific institutional variables could be associated with many other unobserved factors that affect test scores at the same time. Thus, it is hard to draw causal conclusions that policy makers are after. Nonetheless, we will present results from our estimations using equations (7) and (8) as a baseline for our other estimations.

We further include country and time fixed effects. By this, we focus on the variation of our variables of interest over time *within* a single country as well as variation of characteristics *across* countries. This allows us to omit any potential bias to our association between systemic elements of educational systems that could stem from time-invariant factors at the country level and time varying factor. The equations for these fixed effects estimates can be expressed as follows:

$$Y_{i,t} = \alpha + \beta * X_{i,t} + Z_{i,t} + E_i + T_t + u_{i,t} \qquad (9)$$

$$Y_{i,t-1} = \alpha + \beta * X_{i,t-1} + Z_{i,t-1} + E_i + T_t + u_{i,t-1} \qquad (10)$$

E_i is an entity fixed effect at the country level and T_t is a time fixed effect at the year level. This approach allows us to eliminate further bias by controlling for *both* differences across countries as well as changing determinants within a country over time. Thus, our results can more likely be interpreted as causal. Still, we could still be confronted with unobserved heterogeneity as soon as the change of systemic elements of the education system coincides with other changes that drive test scores, especially since we have gaps in our test score data over time. That’s why we try to control for as many as possible other explanatory variables that vary over time.

Next we present a systematic overview on all the explanatory variables that enter in our analysis. We also discuss the difficulties that arise due to missing data in some of our core indicators of educational systems.

Our data set of adjusted test scores between 1965 and 2010 needs to be complemented by respective explanatory variables for the same period (or even longer if we also try to include lagged variables). We complemented the adjusted test score database with data on explanatory factors such as the overall governance of countries beyond educational systems. Governance variables tend to have large impacts on educational quality within a country. A recent study by King et al. (2010) suggests that several governance indicators have a particularly significant impact on the rate of return to education. This finding is based on T.W. Schultz' hypothesis (1975) that economic returns to schooling vary with the capacity to manage unforeseeable price, productivity or technology shocks (see King et al. 2010, p. 3). Thus, more freedom and rights allow individuals to reallocate their time and resources when unforeseeable shocks occur. In turn, investing in human capital becomes critical to ensure that not only is individual reallocation allowed, it is also efficient.

While this positive association between better governance indicators and higher returns to education is robust to the inclusion of several macroeconomic indicators, it could also be the case that better governance indicators just coincide with more promising institutional changes in the education system that affect skills and also returns to education. So, governance indicators are included as further explanatory variables for the adjusted test scores. Specifically, we include a measure for Economic Freedom from the Heritage Foundation's Index of Economic Freedom. This data exists since 1994 and provides an index consisting of several indicators such as the ease to open a business, openness to trade, taxes relative to income etc. We also include Globalization as a governance variable, which comes from an Index by Dreher (2006) for the years 1970-2006. A measure for Civil Rights comes from the Empowerment Rights Index (which is available since 1981, see Cingranelli and Richards 2005). The Empowerment Rights indicator is constructed from several sub-indicators such as freedom of speech, freedom to participate in politics and freedom of religion. We also include a ranking that rates countries by their democratic institutions (on a scale from 0 to 10), which comes from the Freedom House Imputed Polity measure (available since 1972).

Apart from the measure for countries' governance, we add several macroeconomic variables including the population of the country, the log of GDP per Capita and the openness of the country[2].

Next, we provide results on the association between the adjusted test scores and our explanatory governance variable discussed above. Table 2.0 begins with the association between adjusted test scores (in five year steps) and our governance indicators (also included in the respective years in which the test scores are reported). The dependent variable is the Average Score (over all domains), the Average Primary Score (over all domains) and the Average Secondary Score (over all domains), respectively. We additionally control for other macroeconomic factors such as GDP per Capita, Population and Trade Openness. Columns (1)-(3) provide cross-sectional evidence pooling data over time (without country or time fixed effects).

We find positive associations between our indicator of Globalization, Economic Freedom and Democracy and the respective test score measures in columns (1), (2) and (3), respectively. As already discussed, this association could be biased by unobserved omitted variables at the country-level and which vary over time. To address these biases, we include fixed effects in columns (4), (5), (6), and (7) and examine the impact of various governance indicators on average test scores. In column (4) we include country fixed effects and control for macroeconomic factors. Column (5) includes country and time fixed effects as well as macroeconomic controls. Column (6) and (7) include lagged governance indicators in order to explore whether changes in governance in period t-1 affect contemporaneous test scores.

Column (5) is of the most interesting, since these results can most credibly be interpreted as causal due to implicit controls for all potential effects at the country level and over time. Our results indicate that Economic Freedom is positively and statistically significantly associated with higher test scores. This effect persists from our country-fixed specification in column (4). This result might indicate that as economic freedom increases, then so do people's capacity to respond to shocks, as evidenced by King et al. (2010). Thus the returns to education rise and

[2] Population measures stem from the United Nations National accounts, GDP data and Openness from the PENN World Tables.

families and students internalize the benefit of going to school and learning (Bóo 2010). As a result, students invest more in their own human capital.

Another interesting result is the effect of Civil Rights on tests scores. Where Civil Rights have a positive and statistically significant impact in our country-fixed effects model, this effect disappears when we control for time-varying factors. One potential explanation for this is that more civil rights are an indicator for a merit-based society where education leads to better life outcomes and thus students and families invest in education. However, when you control for time-varying effects within a country, this impact disappears since it is truly the rise of merit-based opportunities over time not the act of, for example, speaking freely, that drives achievement. Indeed, when we include a lagged variable for Civil Rights, we see that a high baseline amount of civil rights has positive impacts on achievement, but an additional marginal increase in civil rights actually results in *less* achievement. This might be the case since too many civil rights might distract from education. For example, increased civil rights might result in more teacher strikes.

Table 2.0: Test Scores and Governance Indicators

	Without Country Fixed Effects			With Fixed Effects			
Dependent Variable	Average Score (1)	Average Primary Score (2)	Average Secondary Score (3)	Average Score (4)	Average Score (5)	Average Score (6)	Average Score (7)
Civil Rights	-0.012	0.23	-0.183	**0.698****	-0.022	**-0.508***	
	(0.26)	(0.29)	(0.3)	(0.32)	(0.34)	(0.29)	
Globalization	0.155***	-0.026	0.105	0.149*	0	0.002	
	(0.05)	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	
Democracy	0.478	0.201	0.760**	-0.192	-0.039	0.524	
	(0.31)	(0.36)	(0.38)	(0.78)	(0.68)	(1.07)	
Economic Freedom	0.053	0.196**	0.025	**0.177***	**0.168***	0.028	
	(0.07)	(0.08)	(0.08)	(0.1)	(0.1)	(0.12)	
Openness (in percent)	0.020*	0.013	0.031**	0.060**	0.038	-0.003	-0.001
	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)
Log Population	1.207***	0.701	0.507	4.798	-13.259*	57.878**	43.967**
	(0.37)	(0.46)	(0.38)	(6.82)	(7.44)	(24.64)	(20.38)
Log GDP per Capita	3.114***	2.623***	3.728***	-3.146	-4.375	-6.801**	-6.908**
	(0.64)	(0.72)	(0.78)	(3.05)	(3.93)	(2.62)	(2.71)
Lag of Civil Rights						**0.437***	**0.604****
						(0.22)	(0.23)
Lag of Globalization						-0.130*	-0.106*
						(0.08)	(0.06)
Lag of Democracy						-0.056	-0.014
						(0.35)	(0.32)
Lag of Economic Freedom						0.065	0.083
						(0.1)	(0.11)
Lag of Openness (in percent)						0.002	0.002
						(0.03)	(0.03)
Lag of Log Population						-65.913***	-50.916***
						(22.26)	(19.21)
Lag of Log GDP per Capita						-1.936	-1.17
						(4.41)	(4.19)
R-Squared	0.632	0.497	0.587	0.171	0.936	0.652	0.622
Observations	186	122	120	186	186	138	138
Number of Countries	95	84	72	95	95	91	91

Notes: Dependent Variable: Score averaged over all test score domains. Columns (1), (2) and (3) report OLS estimations. Columns (4), (5), (6) and (7) report Fixed Effects Estimations. Column (4) includes country fixed effects, Column (5) includes both country and time fixed effects, and Columns (6) and (7) include country fixed effects as well as lagged variables. All regressions are estimated with robust standard errors, clustered on the country level.

As we are especially interested in developing countries, we next estimate the same regressions only for a sub-sample of African and Latin American countries. We can include up to 24 different African countries and 17 Latin American countries in the fixed effects estimations for which we have at least two test scores over time as well as information on governance and macro indicators.

We focus on the results of column (5) in Table 2.1, since it includes time and fixed effects and is therefore most causal, we find a significant positive effect of all Democracy on the average test score in the sub-sample of Latin American countries. This result is the most reliable, as without controlling for unobserved time invariant heterogeneity at the country-level, several factors could cast doubt on the causality of this relation. Changes in governance over time could simply have initiated more concrete changes in the educational system, which are, in turn, responsible for the improvement in test scores. So, it is perhaps not the advancement in governance indicators that affects the test scores, but the systemic changes in the educational systems that come along with improvements in Economic Freedom, Democracy, Civil Rights or Globalization.

Like in Table 2.0, it seems that more favorable governance indicators boost educational attainment even in developing countries. It is interesting to note that Democracy is the key governance indicator that boosts tests scores in developing countries, while in more developed countries, Economic Freedom and basic civil rights matter most.

Table 2.1: Test Scores and Governance Indicators: African and Latin American Countries

	Without Country Fixed Effects			With Country Fixed Effects			
	Average Score	Average Score	Average Score	Average Score	Average Score	Average Score	Average Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Civil Rights	-1.057*	-0.569	-0.454	-0.613	0.028	-1.187	-0.844
	(0.59)	(0.73)	(0.4)	(0.93)	(1.0)	(0.77)	(0.75)
Globalization	-0.204*	-0.164	-0.162**	0.098	-0.055	-0.087	-0.065
	(0.12)	(0.11)	(0.08)	(0.18)	(0.12)	(0.1)	(0.11)
Democracy	0.66	1.5	0.387	**1.776*** **	**3.200*** *	**1.592*** *	**1.555*** *
	(0.63)	(1.04)	(0.47)	(0.55)	(1.4)	(0.62)	(0.69)
Economic Freedom	0.281	0.315**	0.298***	**0.432*** *	0.15	**0.231*** *	**0.205***
	(0.19)	(0.13)	(0.11)	(0.21)	(0.1)	(0.11)	(0.11)
Openness (in percent)	-0.008	0.005	-0.003	-0.001	-0.026	0.004	0
	(0.04)	(0.03)	(0.03)	(0.05)	(0.08)	(0.03)	(0.04)
Log Population	0.905	1.704	1.371	-12.679	-28.566	-2.115	14.821
	(1.34)	(0.91)	(0.8)	(8.03)	-29.28	(6.49)	(15.3)
Log GDP per Capita	2.123	-1.274	3.203**	-4.146	-7.948	-8.056	0.184
	(1.71)	(2.26)	(0.94)	(5.92)	(11.29)	(4.91)	(7.21)
R-Squared	0.35	0.416	0.46	0.486	0.802	0.455	0.95
Observations	40	28	68	40	28	68	68
Number of Countries	24	17	41	24	17	41	41

Notes: Dependent Variable: Score averaged over all test score domains. Columns (1), (2), and (3) report OLS estimations for Africa, Latin America, and both, respectively. Column (4), (5), (6) and (7) report Fixed Effects Estimations. In particular, column (4), (6), and (7) report country fixed effects for Africa, Latin America, and both, respectively. Column (5) reports country fixed and time fixed effects for Latin America. All regressions estimated with robust standard errors, clustered on the country level.

7. Conclusion

In this paper, we presented an overview of the construction of an international database comparable over countries and over years. One special focus was on the inclusion of developing countries. In particular, using the methodology of Altinok and Murseli (2007) we built a data set of comparable test scores from 1965-2010 for a set of 128 countries.

To construct this data set we standardized international assessments, such as PISA and TIMSS, across types of exams by linking them to the United States as a reference point, since the Unites

States participates in all international assessments. We further standardized tests over time by linking our United States reference point to the National Assessment of Educational Progress (NAEP), which has been administered in the United States since 1969. Finally, we include developing countries that have participated in regional assessments such as LLECE, PASEC, and SACMEQ, by using scores from countries that participated in both a regional and international assessment as an index.

While our database allows the comparison of many countries over time, the development of this database still requires improvements and extensions. For example, our database should continually be updated with results from the most recent international and regional achievement tests. Additionally, the anchoring methodology for developing countries which makes use of *doubloon countries* could probably be made more accurate, as it surely will be as time goes on and more Latin American and African countries participate in PISA and TIMSS.

Additionally, in this paper, we have provided an application of our international database. Our ultimate goal was to use our adjusted test scores as dependent variables in regressions that could explain differences in human quality development over countries and over time. To that end, we used our extended and updated version of the Altinok and Murseli (2007) database to inform us about which causal inputs lead to better learning.

We are especially interested in governance indicators and macro variables. We can identify some insightful associations between our governance indicators and the test score development, both for the full sample of countries but also for the sub-sample of African and Latin American countries. Governance indicators involving Economic Freedom, Democracy and Civil Rights show positive association with test scores. The result is robust to the inclusion of several governance and macroeconomic indicators as well as lagged variables. The use of country and time fixed effects supports a causal interpretation of our results.

In conclusion, this is a first attempt to facilitate the benchmarking process of human capital quality and educational institutions all around the world; it is just a starting point. We have created one of the first databases on student achievement that is comparable across tests, countries and time. We have also included developing countries. Much more research has to be done in order to improve this approach. In particular, there should be a focus on the adjustment of test scores to make them comparable over different surveys and years. Finally, countries should be encouraged to participate in as many international surveys as possible. That would ease the interpretation and the reliability of all methods that seek to make test scores from different surveys comparable. Another point of research includes utilizing our dataset to identify additional inputs, beyond governance variables, which boost achievement.

References

Altinok, Nadir and Murseli, Hatidje (2007): International Database on Human Capital Quality. *Economics Letters* 96 (2): 237-244.

Barro, Robert J. (1991): Economic Growth in a Cross Section of Countries. *Quarterly Journal of Economics* 106 (2): 407-443.

Barro, Robert J. (1998): Determinants of Economic Growth: A Cross-Country Empirical Study," *MIT Press Books,* The MIT Press, edition 1, volume 1, number 0262522543.

Barro, Robert J. and Lee, Jong-Wha (2010): "A New Data Set of Educational Attainment in the World 1950-2010," *NBER Working Papers 15902,* National Bureau of Economic Research, Inc.

Beck, Thorsten, Feyen, Erik, Ize, Alain, Moizeszowicz, Florencia (2008): Benchmarking Financial Development. World Bank Policy Research Working Paper 4638.

Bóo, Florencia (2010): Returns to Education and Macroeconomic Shocks: Evidence from Argentina. IZA Discussion Paper.

Brunello, Giorgio and Checci, Daniele (2007): Does school tracking affect equality of opportunity? New international evidence. *Education Policy* 22 (October): 781-861.

Cingranelli, David L. and Richards, David L. (2005): Measuring the Level, Pattern, and Sequence of Government Respect for Physical Integrity Rights. *International Studies Quarterly* 43(2): 407–17.

Dreher, Axel (2006): Does Globalization Affect Growth? Evidence from a New Index of Globalization. *Applied Economics* 38(10):1091–110.

Fuchs, Thomas and Woessmann, Ludger (2007): What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data. *Empirical Economics* 32 (2-3): 433-464.

Hanushek, Eric A. (1979): Conceptual and Empirical Issues in the Estimation of Educational Production Functions, *Journal of Human Resources* 14 (3): 351-388.

Hanushek, Eric A. and Kimko, Denis D. (2000): Schooling Labor-Force Quality and the Growth of Nations. *American Economic Review* 90 (5): 1184-1208.

Hanushek, Eric A. and Woessmann, Ludger (2006): Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries. *Economic Journal* 116 (510): C63-C76.

Hanushek, Eric A. and Woessmann, Ludger (2008): The Role of Cognitive Skill in Economic Development. *Journal of Economic Literature* 46 (3): 607-668.

Hanushek, Eric A. and Woessmann, Ludger (2009a): Do Better Schools Lead to more Growth? Cognitive Skills, Economic Outcomes, And Causation. NBER Working Paper 14633.

Hanushek, Eric A. and Woessmann, Ludger (2009b): Schooling, Cognitive Skills, And the Latin American Growth Puzzle. NBER Working Paper 15066.

Hanushek, Eric A. and Woessmann, Ludger (2010): The Economics of International Differences in Educational Achievement. NBER Working Paper 15949.

Juerges, Hendrik, Schneider, Kerstin and Büchel, Felix (2005): The Effect Of Central Exit Examinations On Student Achievement: Quasi-Experimental Evidence From TIMSS Germany. *Journal of the European Economic Association* 3(5): 1134-1155.

King, Elizabeth M., Montenegro, Claudio E. and Orazem, Peter F. (2010): Economic Freedom, Human Rights, and the Returns to Human Capital – An Evaluation of the Schultz Hypothesis. World Bank Policy Research Working Paper 5405.

Lee, Jong-Wha and Barro, Robert J. (2001): Schooling Quality in a Cross section of Countries. *Economica* 68 (272): 465-488.

Mankiw, N. Gregory, Romer, David, and Weil, David N. (1992): A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics* 107 (2): 407-437.

Rindermann, H. and S.J. Ceci. (2009): Educational Policy and Country Outcomes in International Cognitive Competence Studies. *Perspectives on Psychological Science* 4 (6): 551-577.

Schultz, Theodore (1975): The Value of Ability to Deal with Disequilibria. *Journal of Economic Literature* 13 (3): 827-846.

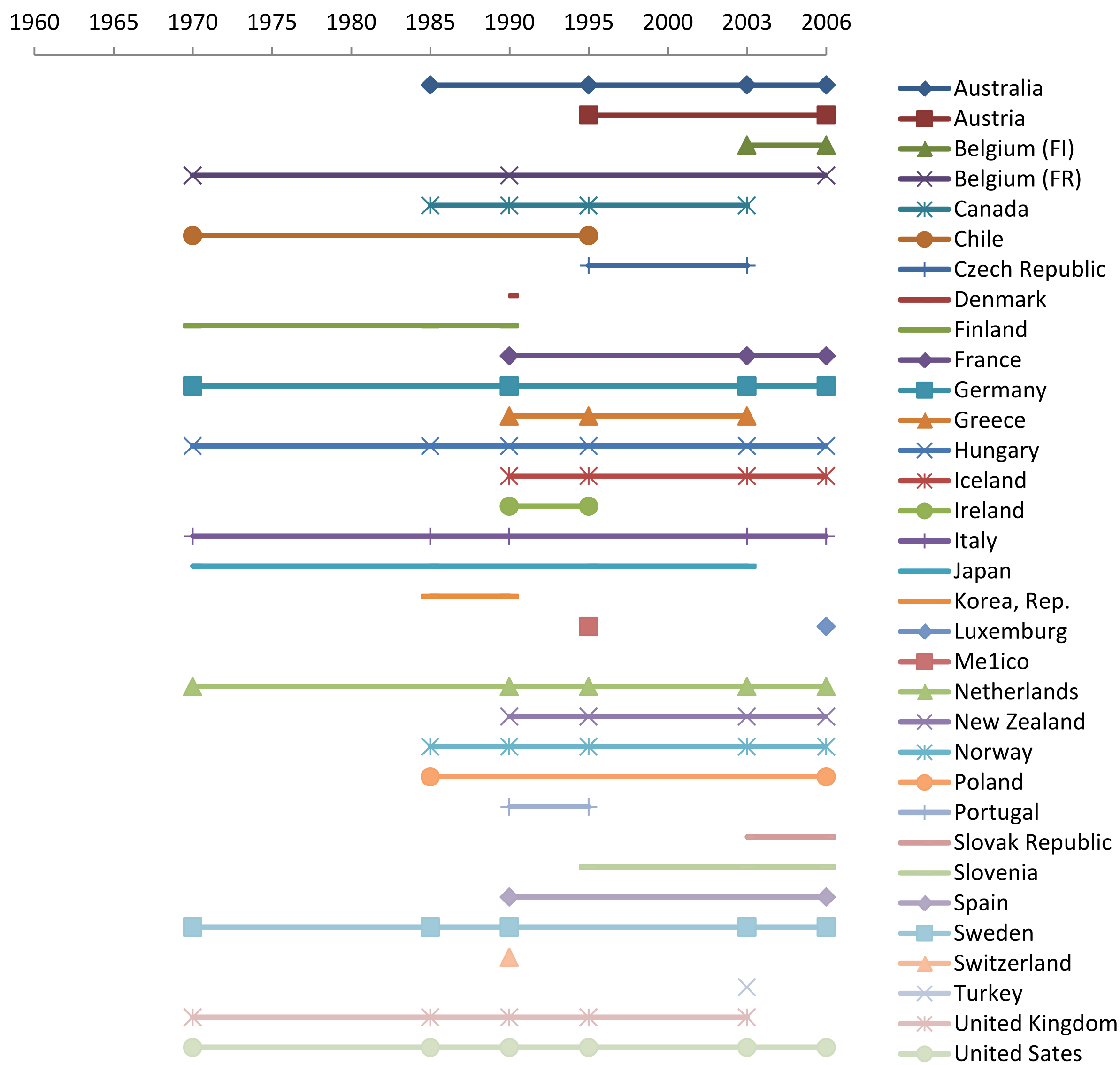
West, Martin R. and Woessmann, Ludger (2010): Every Catholic Child in a Catholic School: Historical Resistance to State Schooling, Contemporary School Competition, and Student Achievement across Countries. *Economic Journal* 120 (546): F299-F255.

Woessmann, Ludger (2003): Schooling Resources, Educational Institutions, and Student Performance: The International Evidence. *Oxford Bulletin of Economics and Statistics* 65 (2): 117-170.

ANNEX

Figure A1: Test Score Availability by Country – Primary Scores Averaged over Subjects

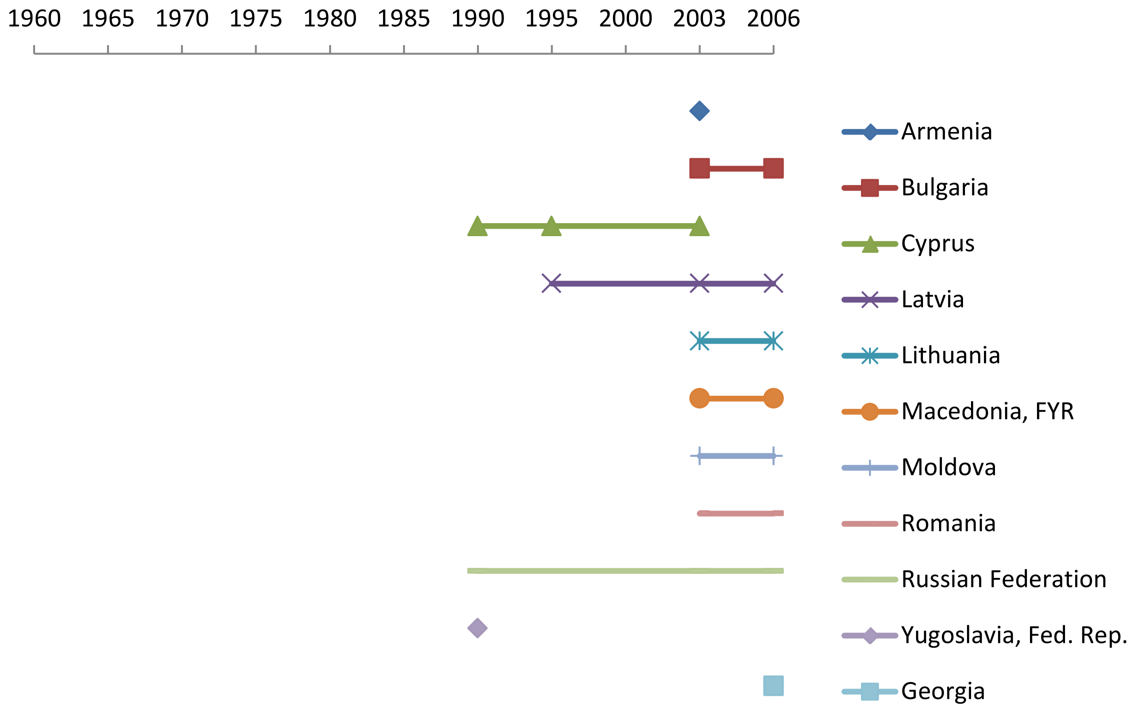
OECD Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure A2: Test Score Availability by Country – Primary Scores Averaged over Subjects

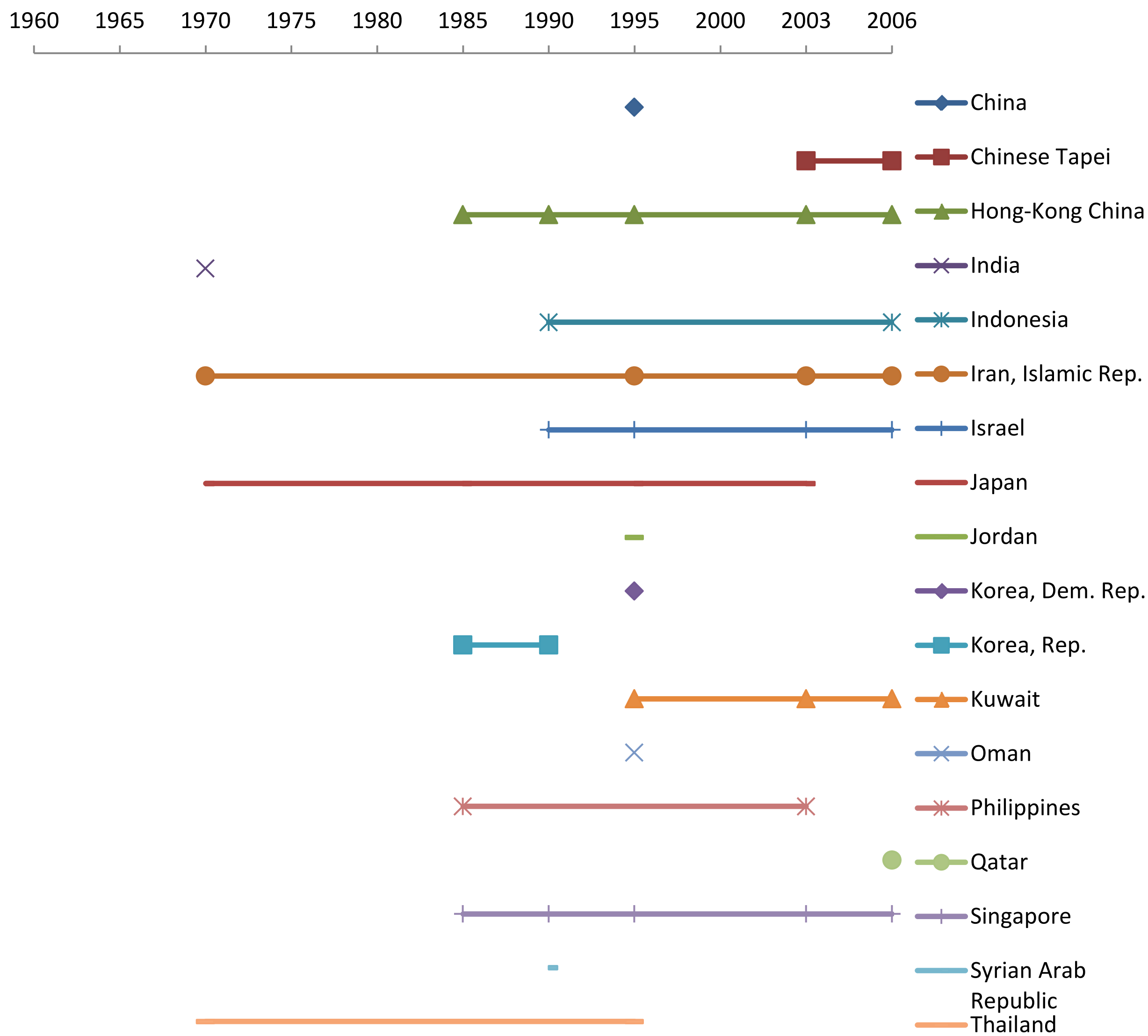
Non-OECD European Countries



Notes: Every Marker indicates the availability of a test score for the respective five-year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure A3: Test Score Availability by Country – Primary Scores Averaged over Subjects

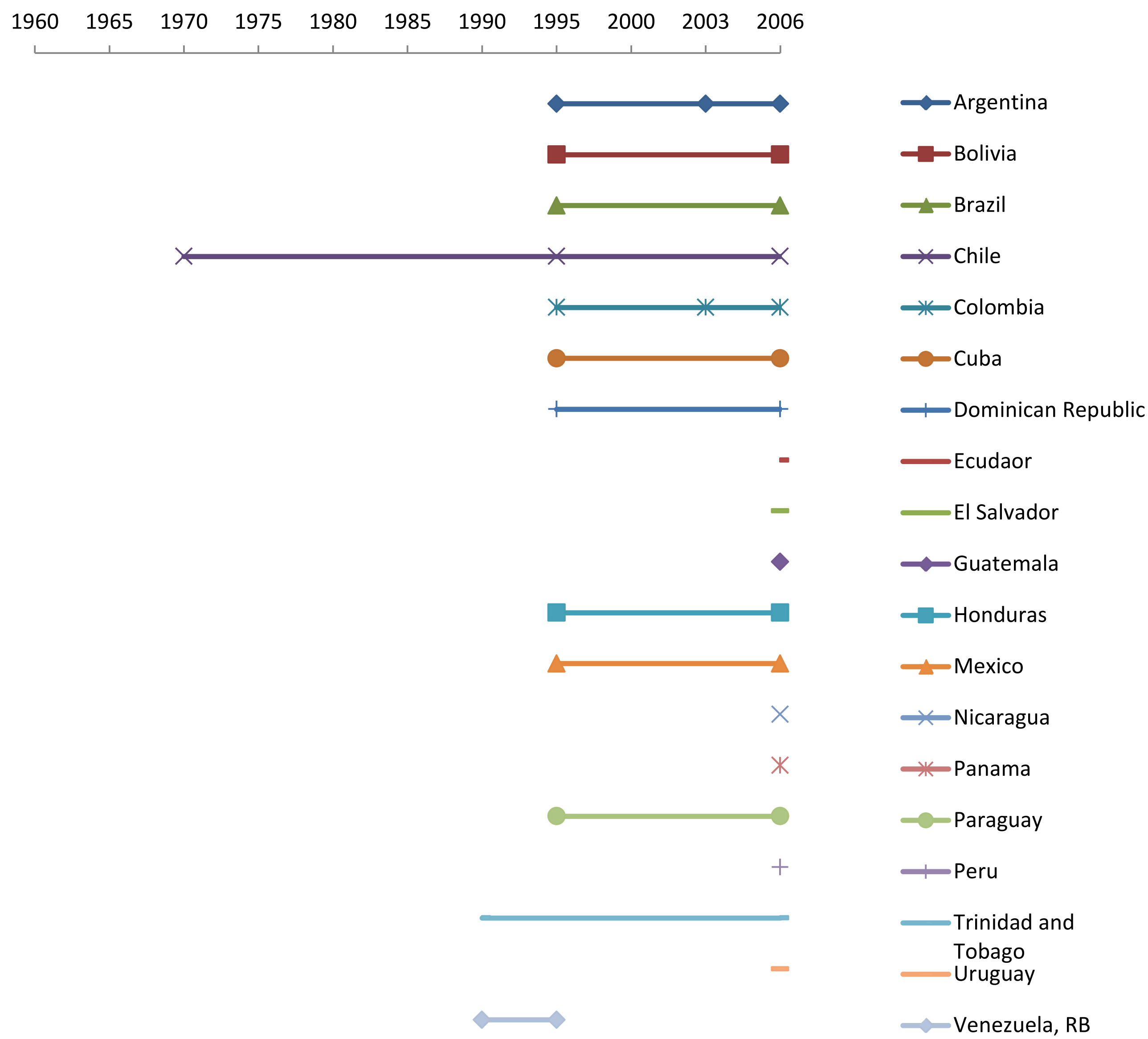
Asian Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure A4: Test Score Availability by Country – Primary Scores Averaged over Subjects

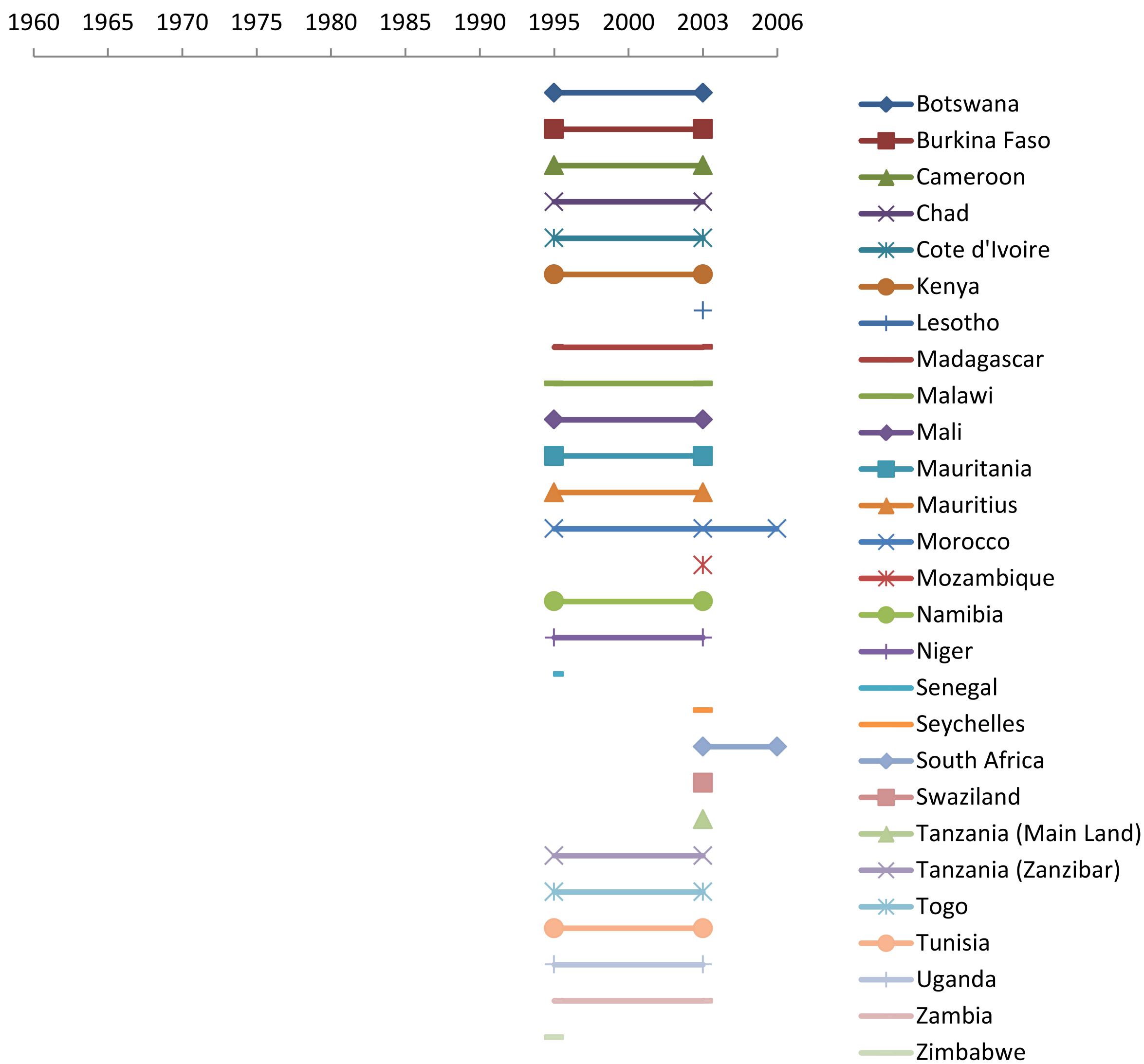
Latin American and Caribbean Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006

Figure A5: Test Score Availability by Country – Primary Scores Averaged over Subjects

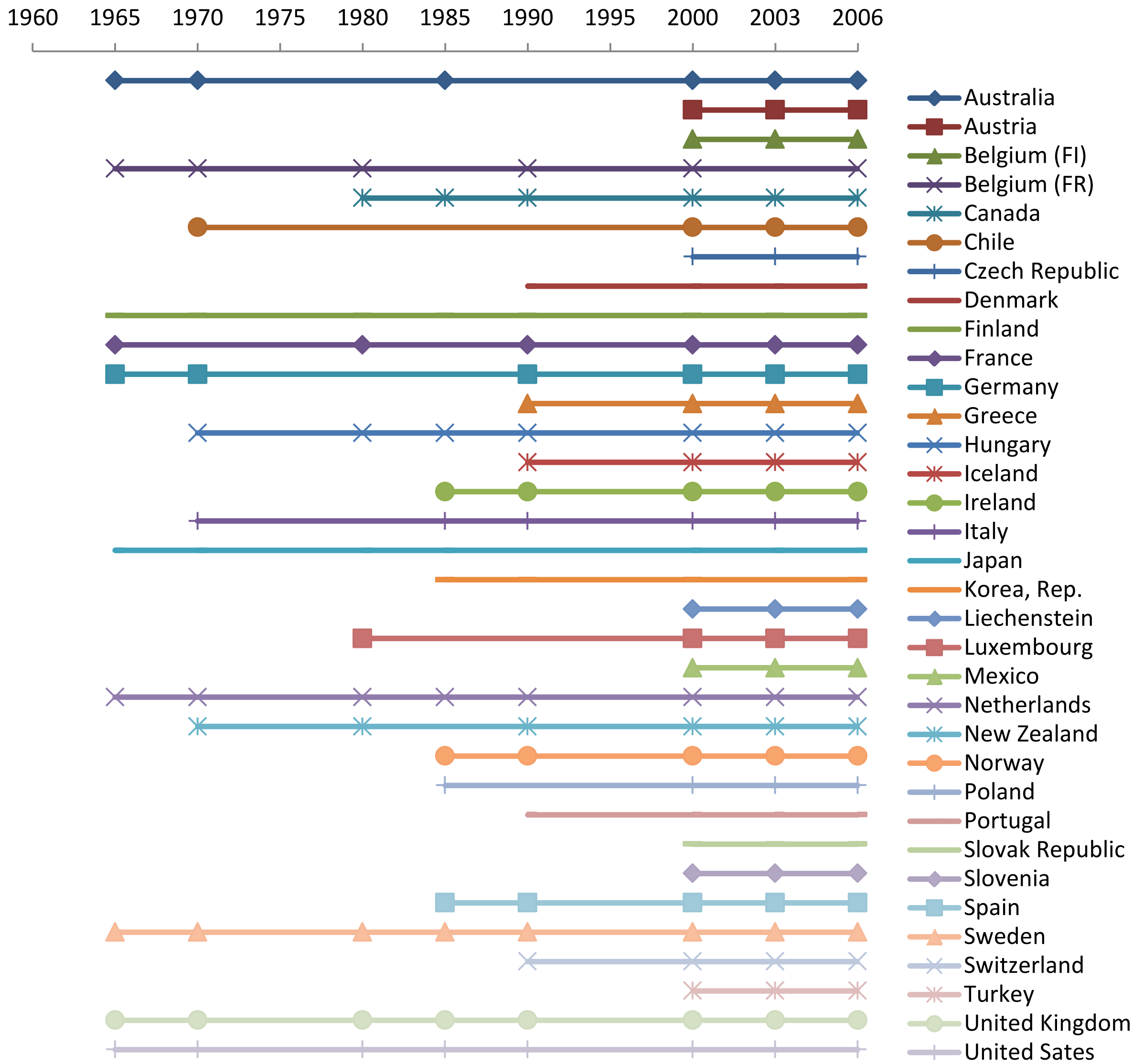
African Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure B1: Test Score Availability by Country – Secondary Scores Averaged over Subjects

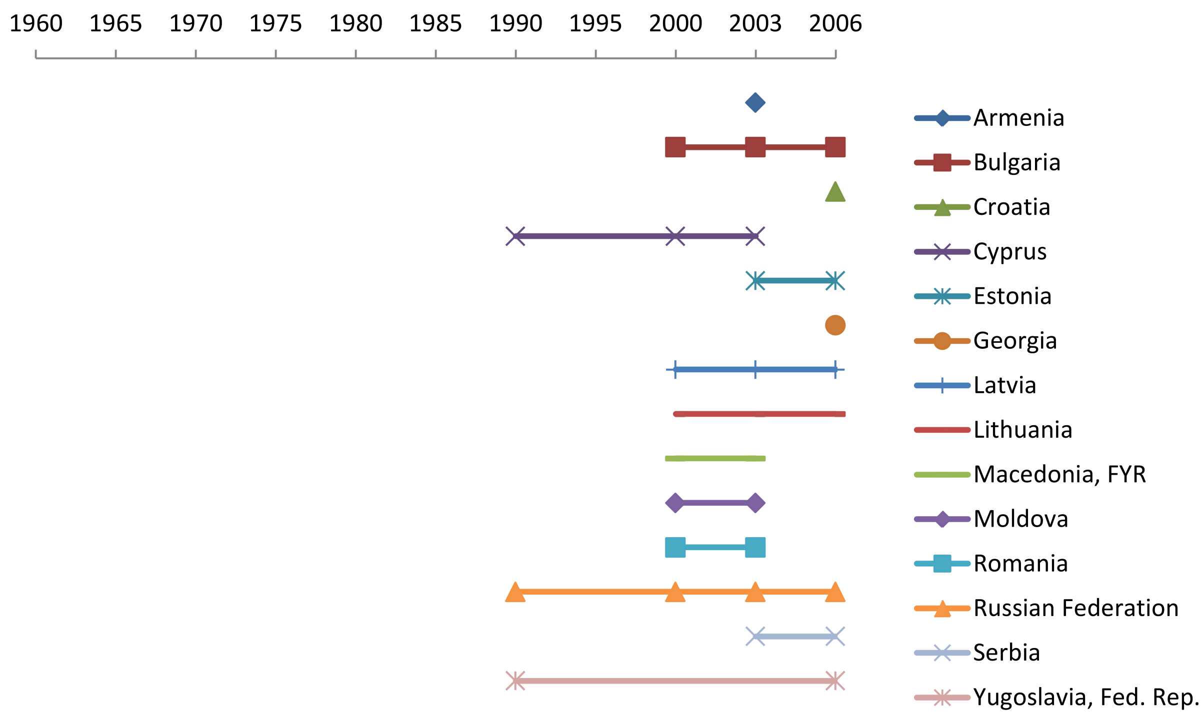
OECD Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure B2: Test Score Availability by Country – Secondary Scores Averaged over Subjects

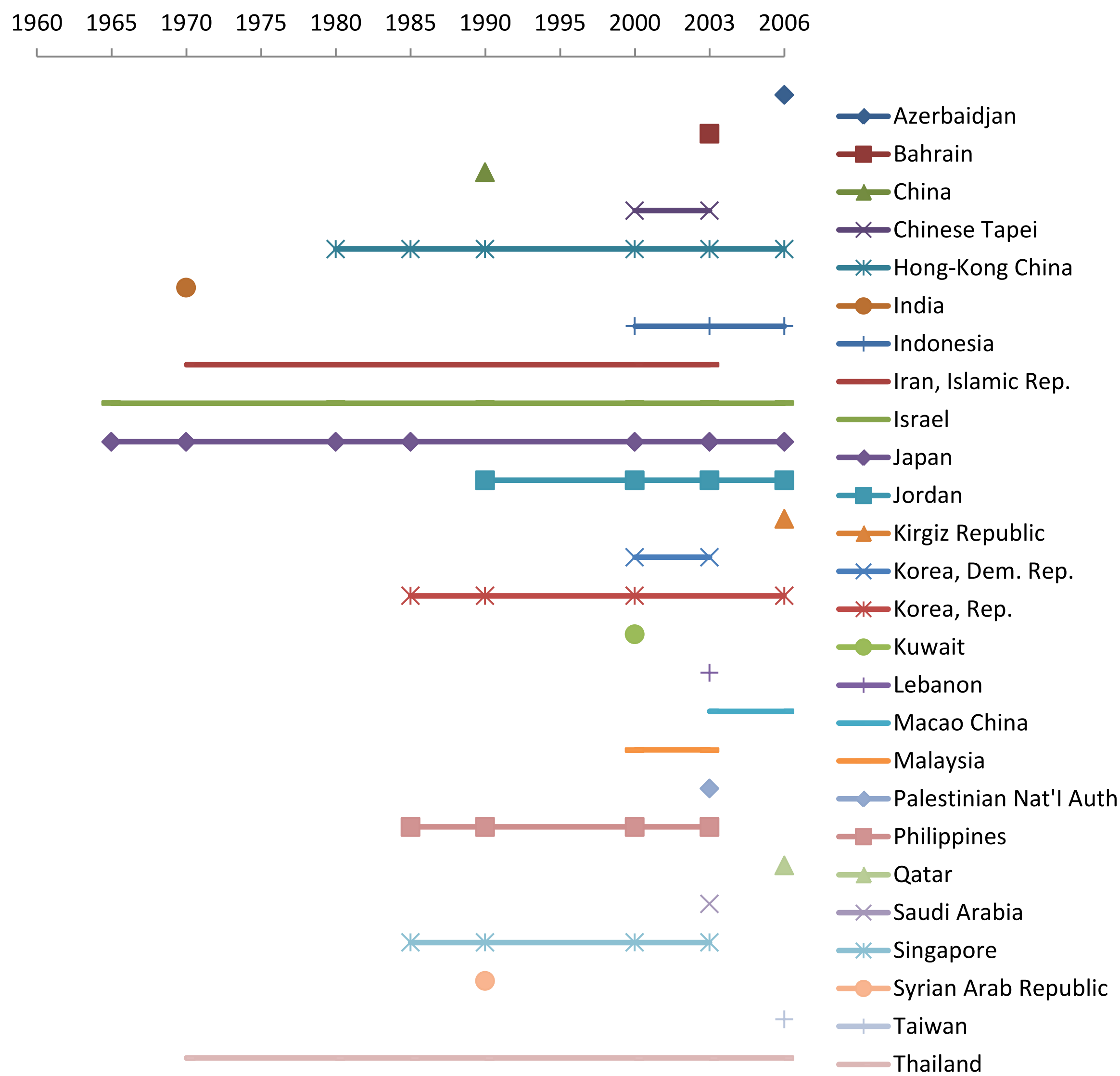
Non-OECD European Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure B3: Test Score Availability by Country – Secondary Scores Averaged over Subjects

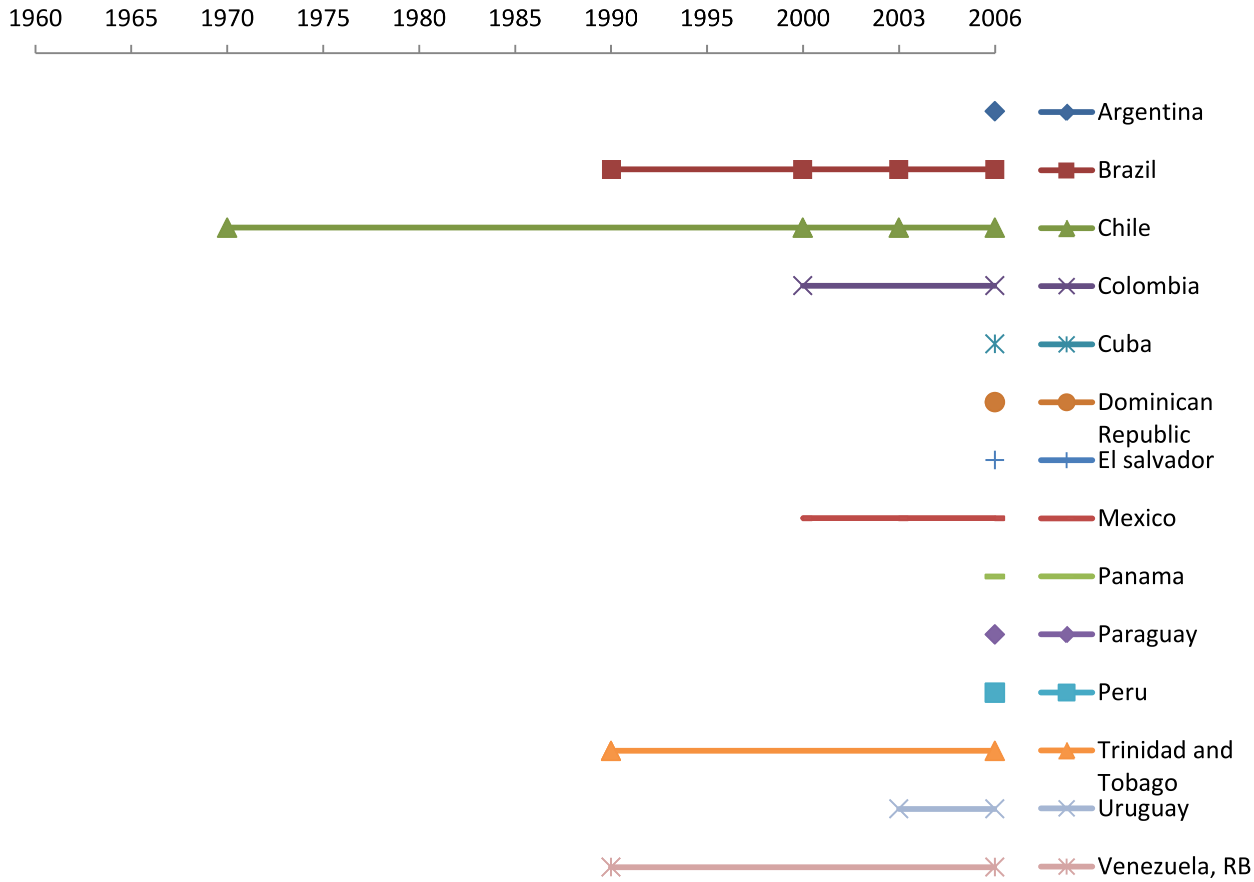
Asian Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure B4: Test Score Availability by Country – Secondary Scores Averaged over Subjects

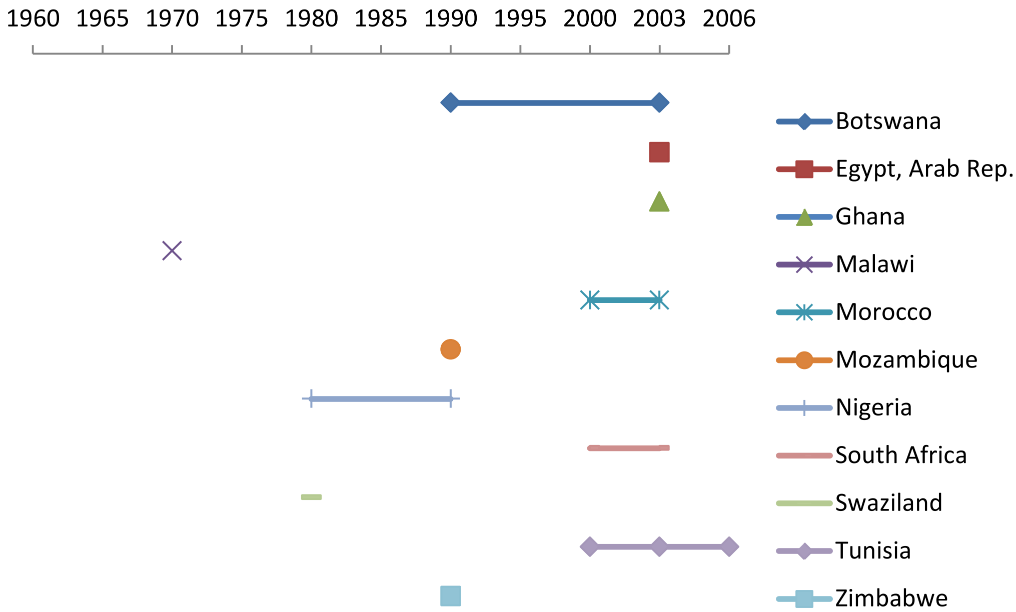
Latin American and Caribbean Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure B5: Test Score Availability by Country – Secondary Scores Averaged over Subjects

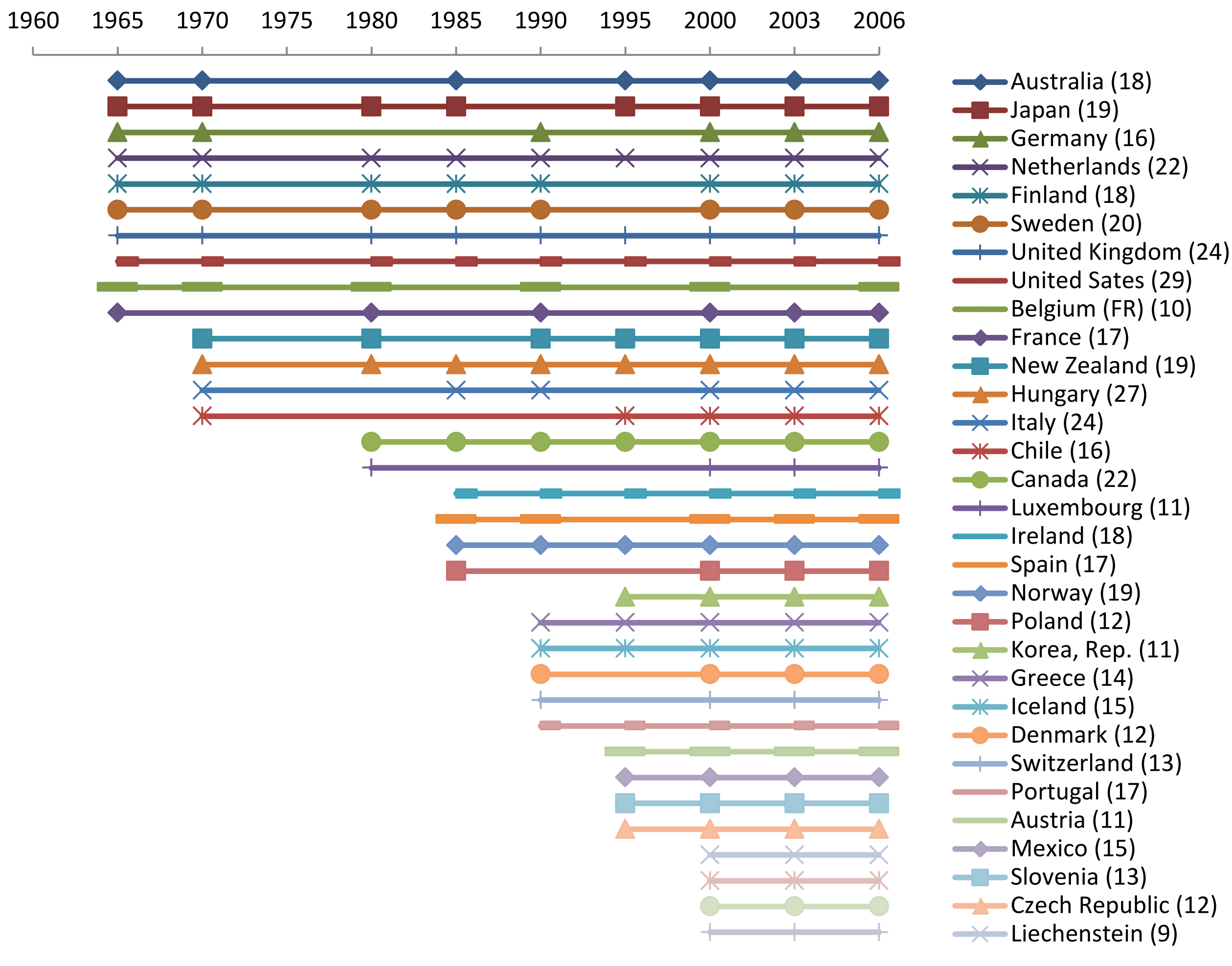
African Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2003.

Figure C1: Test Score Availability by Country – Scores Averaged over Grades and Subject

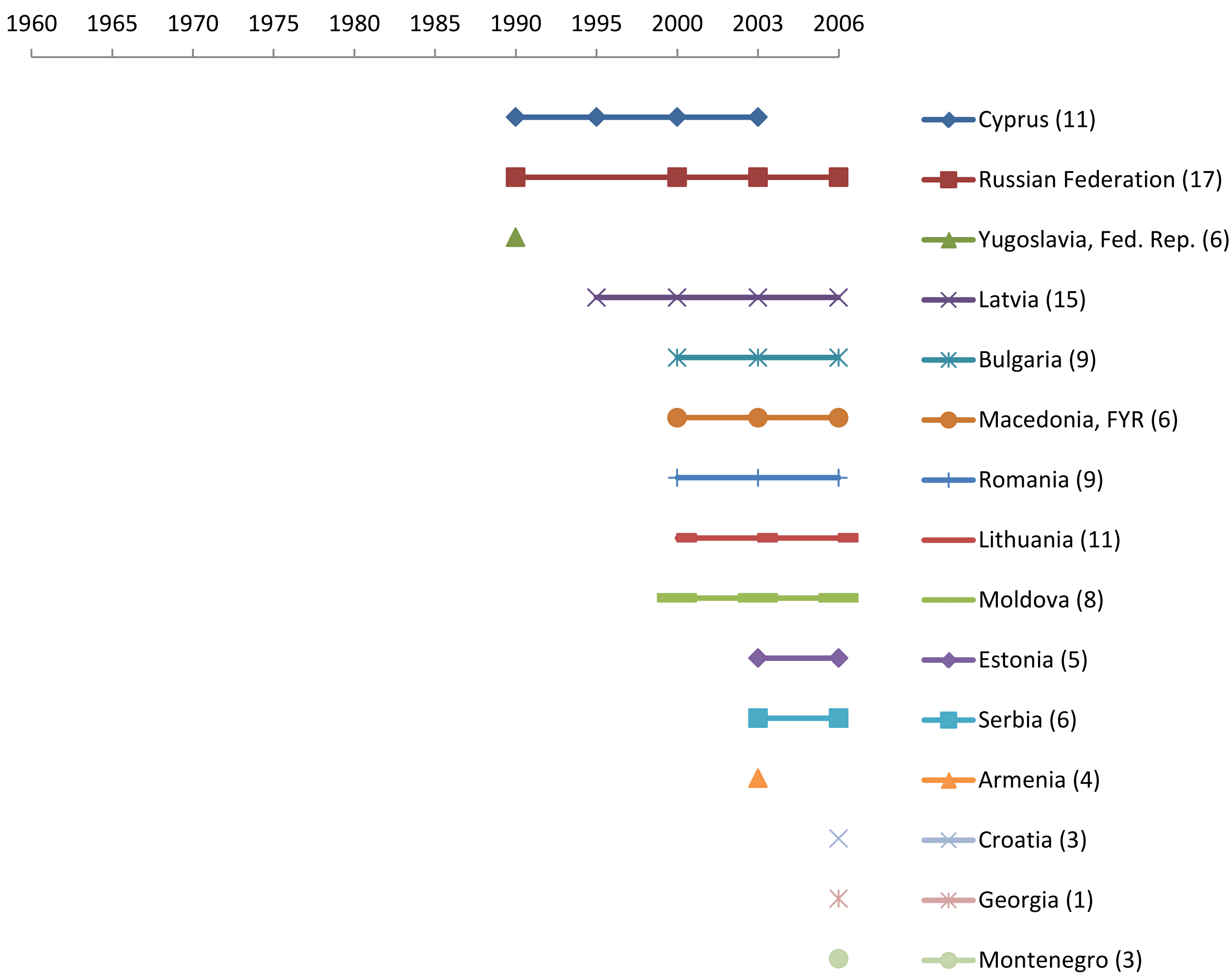
OECD Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure C2: Test Score Availability by Country – Scores Averaged over Grades and Subject

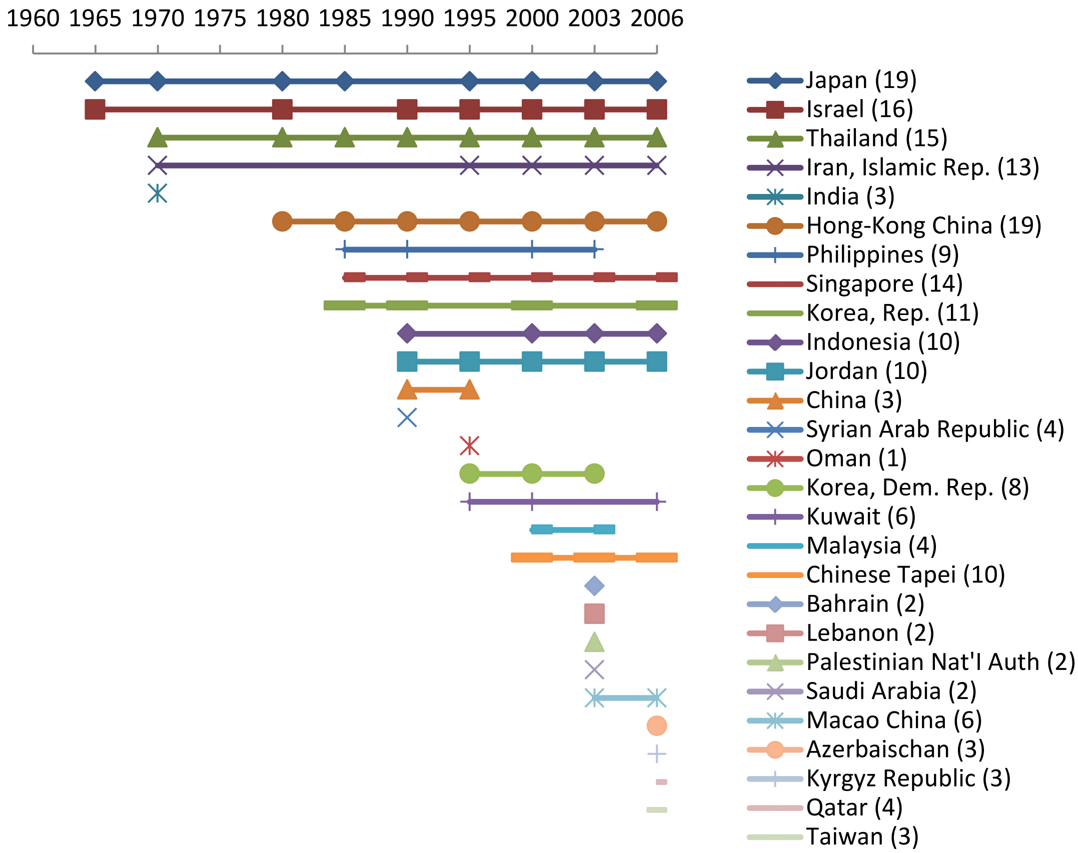
Non-OECD European Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure C3: Test Score Availability by Country – Scores Averaged over Grades and Subject

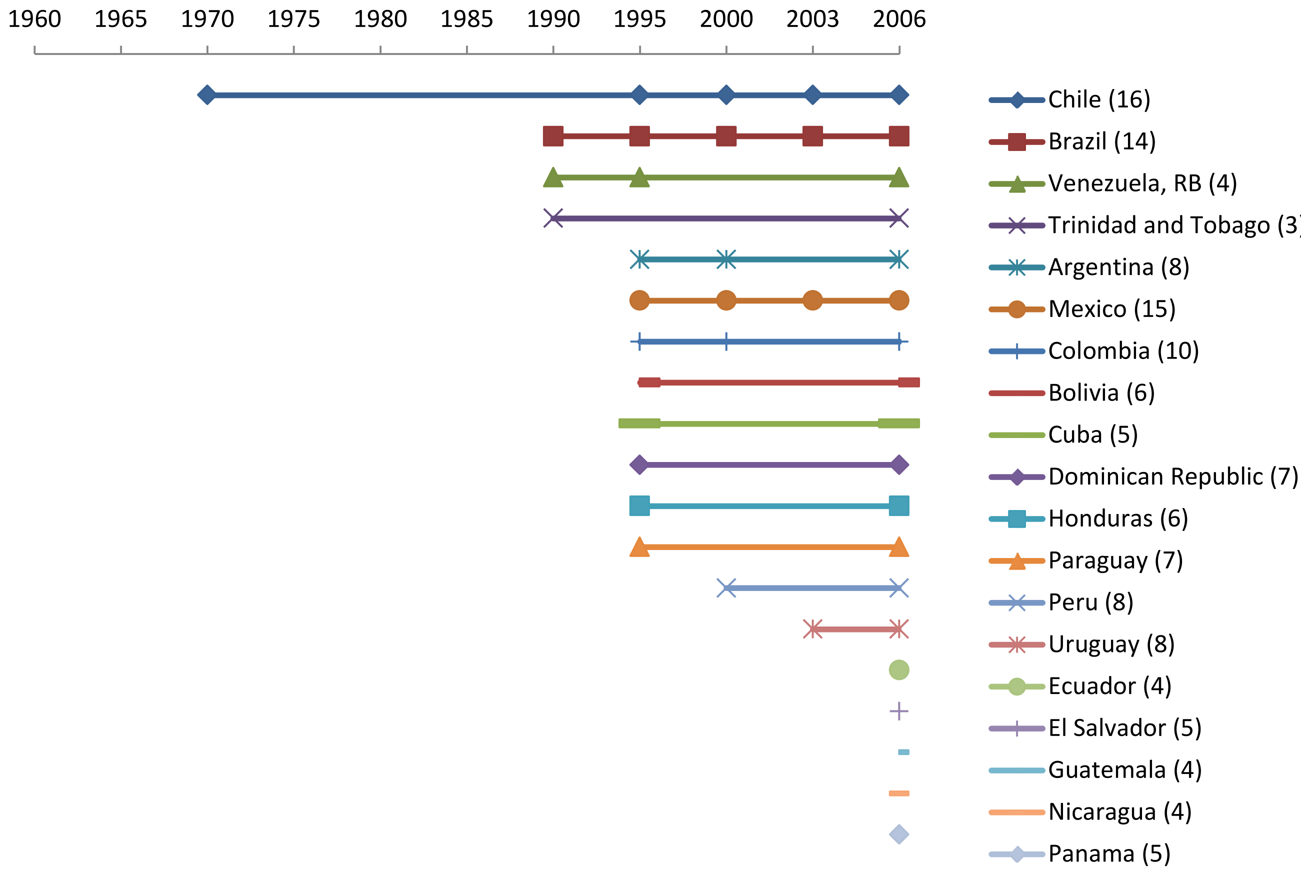
Asian Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure C4: Test Score Availability by Country – Scores Averaged over Grades and Subject

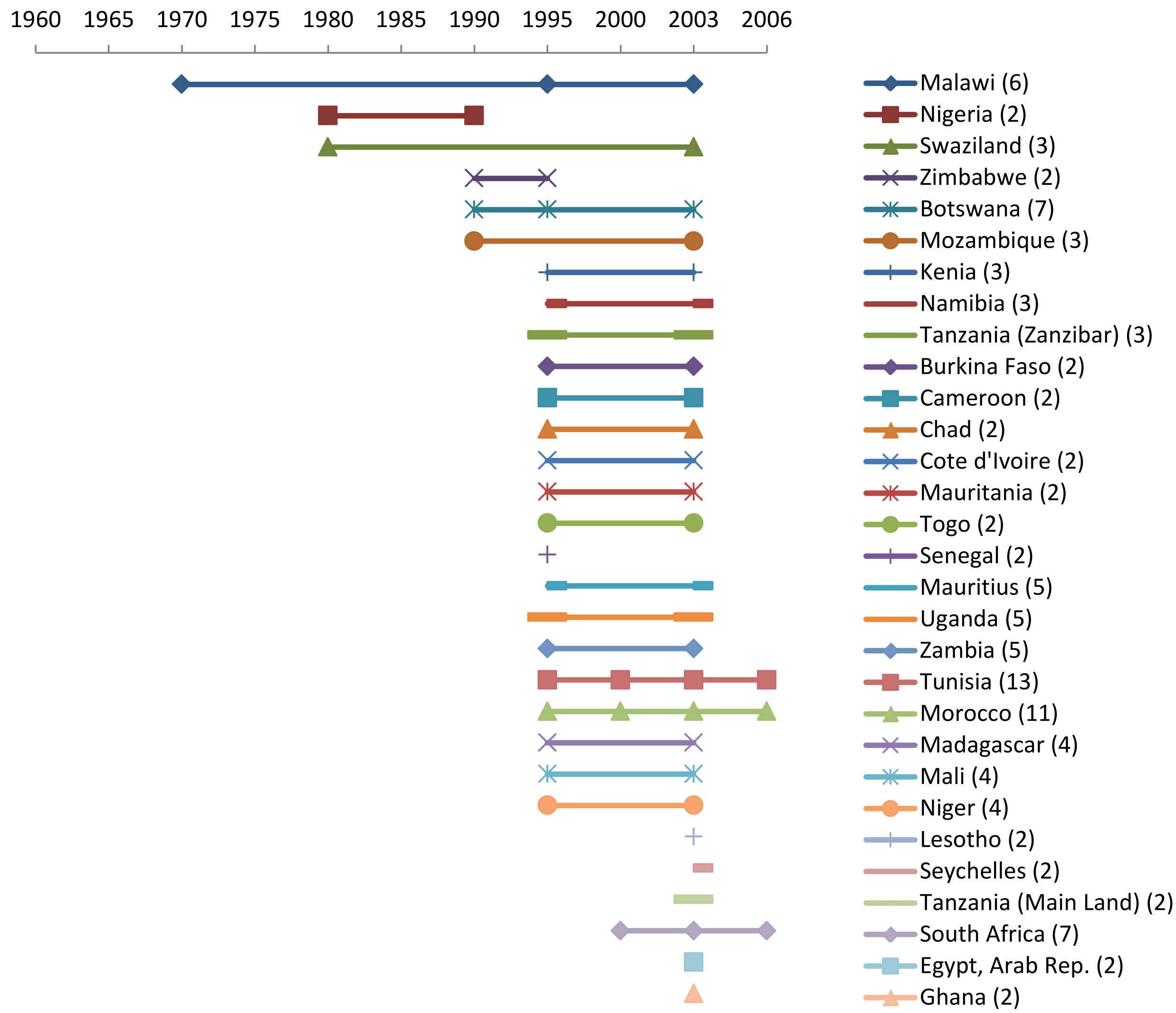
Latin American and Caribbean Countries



Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

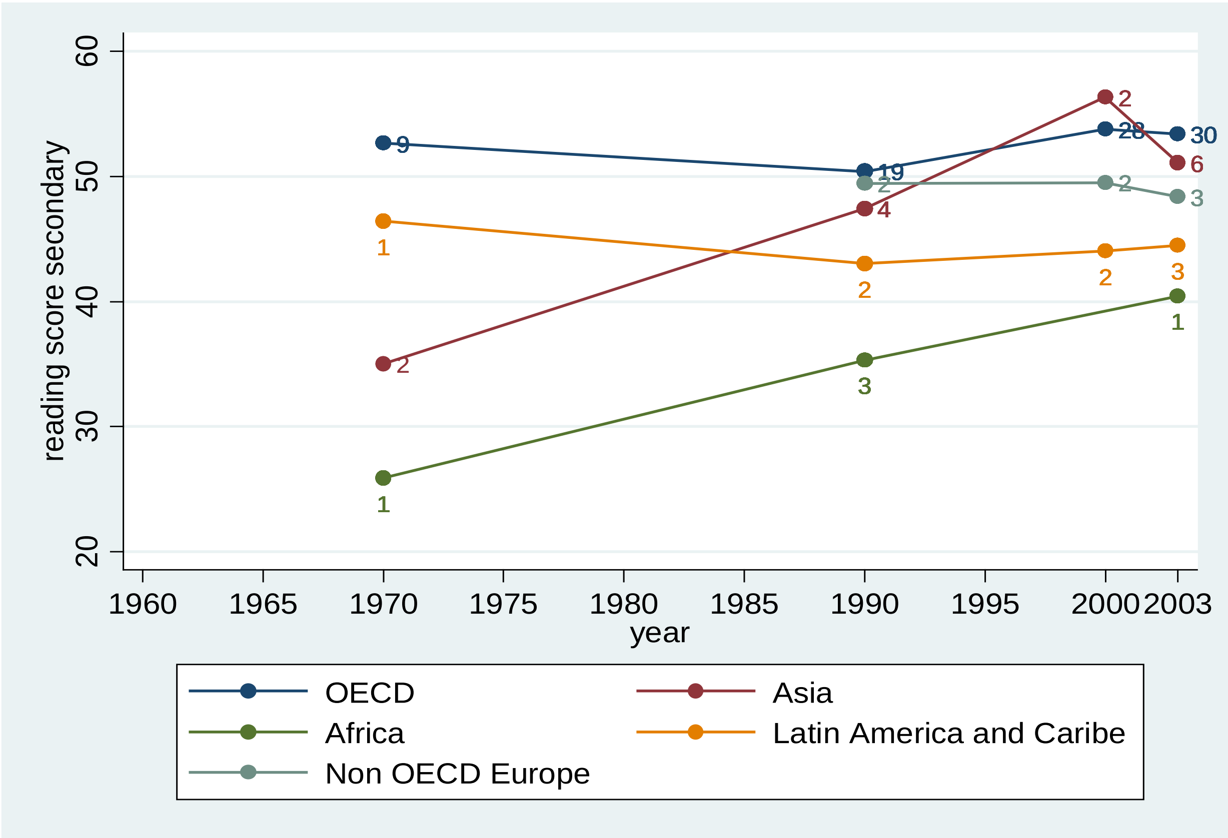
Figure C5: Test Score Availability by Country – Scores Averaged over Grades and Subject

African Countries



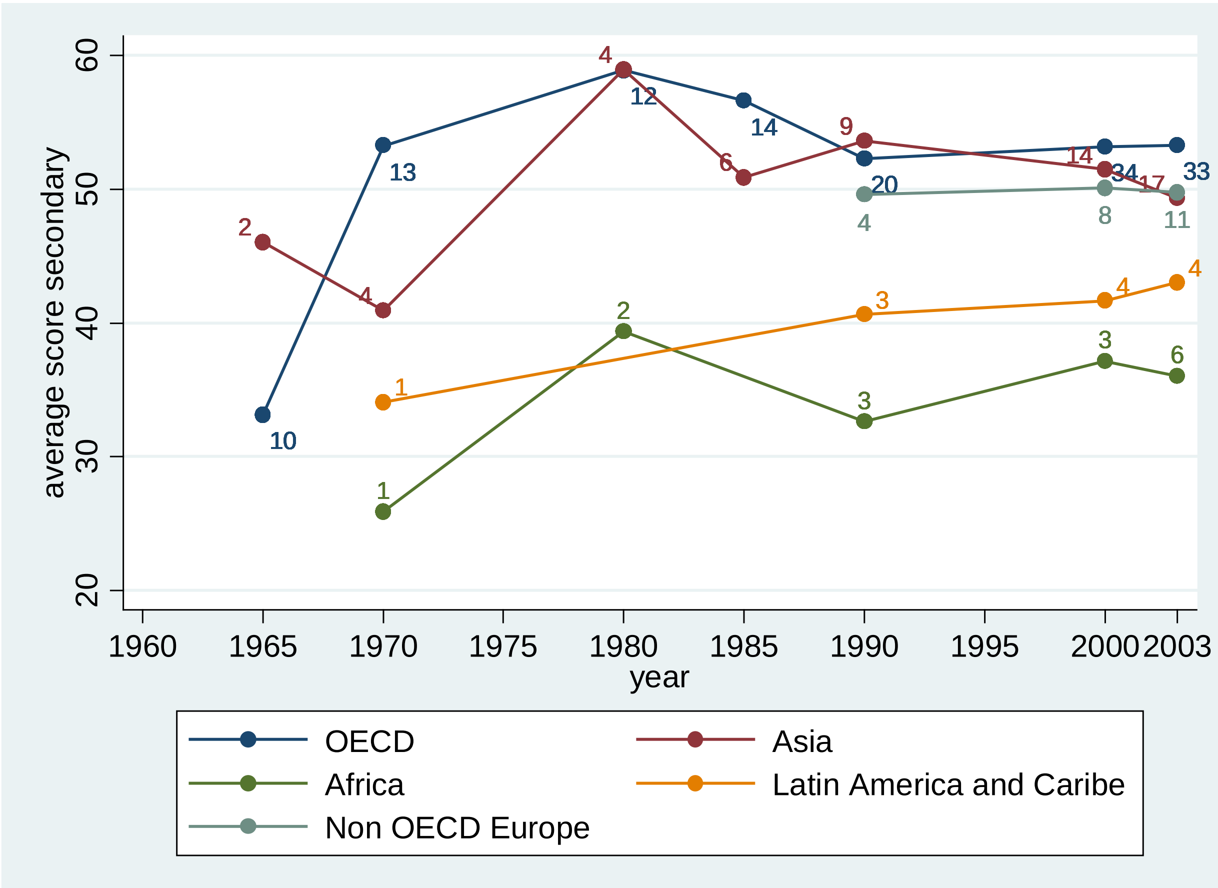
Notes: Every Marker indicates the availability of a test score for the respective five year span. The number in brackets behind countries' names show the overall ability of single test scores (by subject and by grade) for the respective country over the whole span between 1965 and 2006.

Figure D1: Test Score Trends over Time averaging over all Test Domains (on Primary School Level)



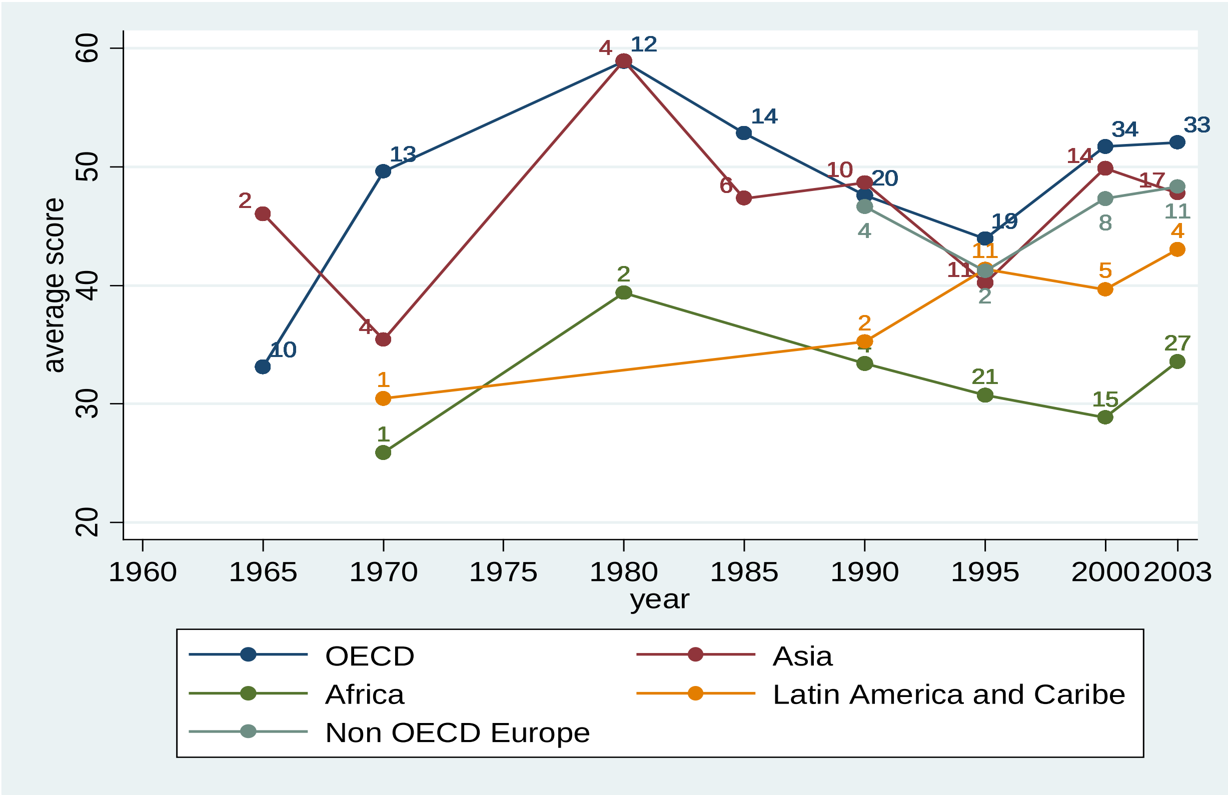
Notes: Every Marker indicates the average test score for the respective world region averaged over all test domains (Math, Reading and Science), only including tests with primary school students. The numbers at the markers indicate the number of countries over which the average is computed.

Figure D2: Test Score Trends over Time averaging over all Test Domains (on Secondary School Level)



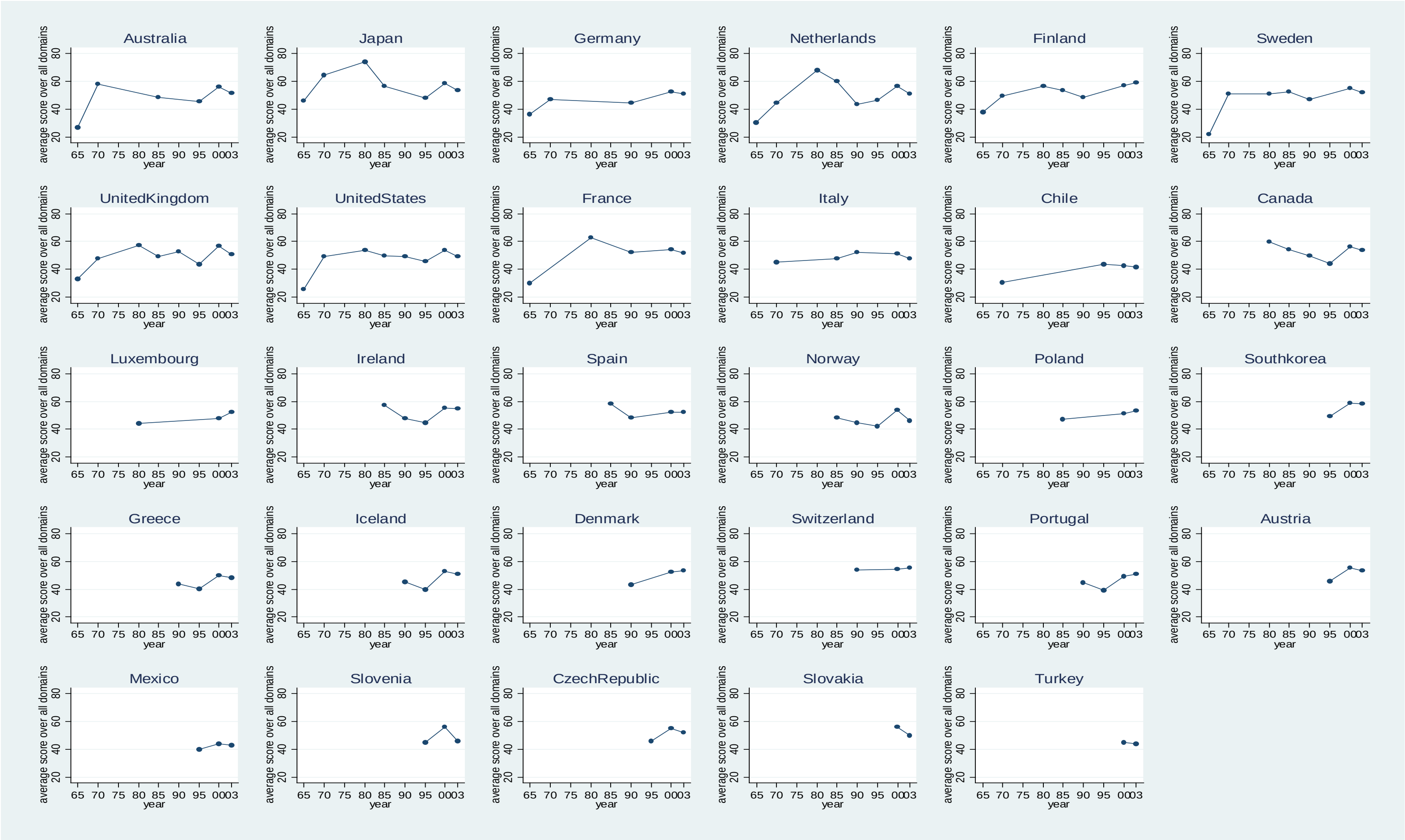
Notes: Every Marker indicates the average test score for the respective world region averaged over all test domains (Math, Reading and Science), only including tests with secondary school students. The numbers at the markers indicate the number of countries over which the average is computed.

Figure D3: Test Score Trends over Time averaging over all Test Domains (on Secondary School Level)



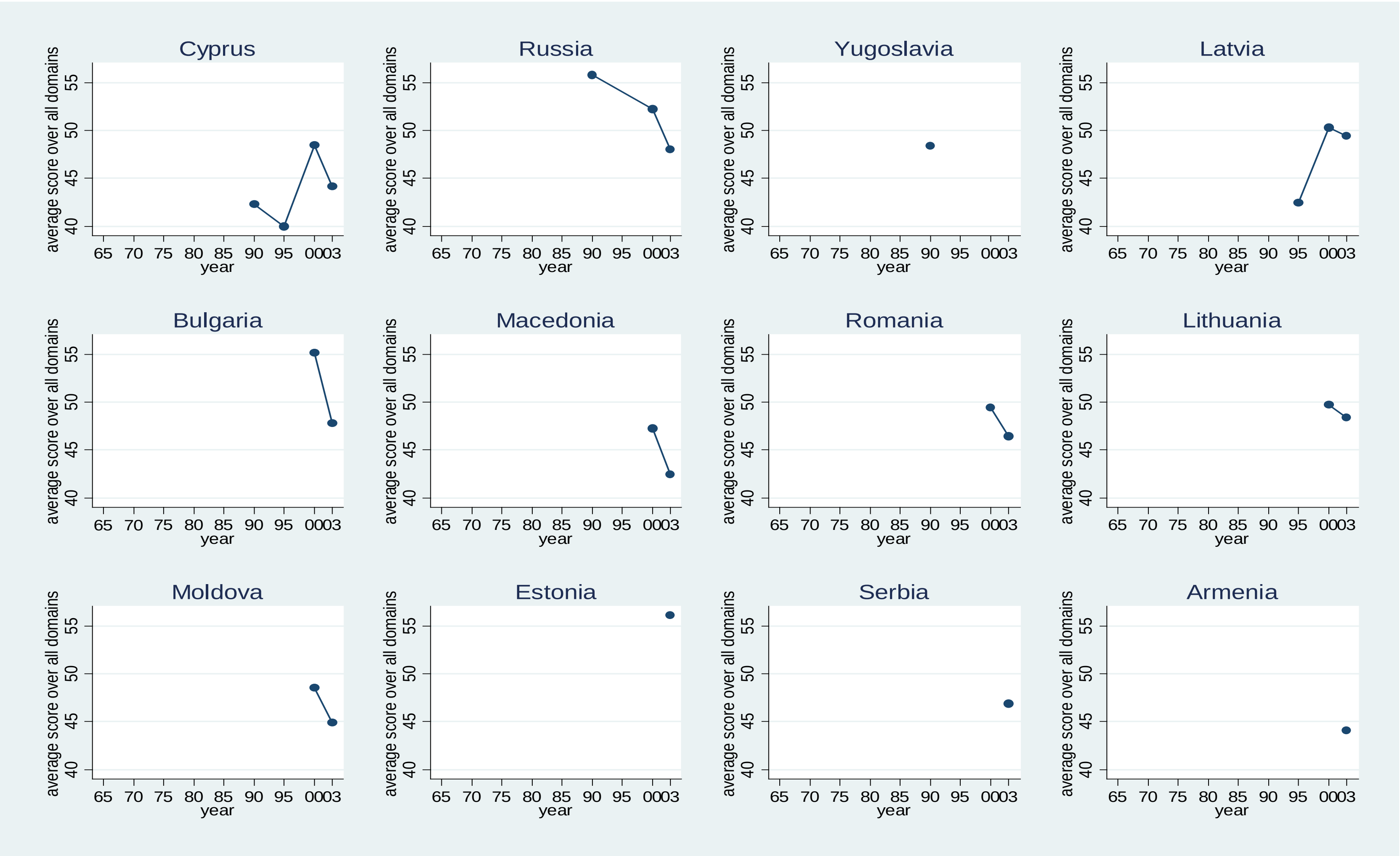
Notes: Every Marker indicates the average test score for the respective world region averaged over all test domains (Math, Reading and Science) and all grade levels. The numbers at the markers indicate the number of countries over which the average is computed.

Figure E1: Test Score Trends over Time averaging over all Test Domains and Grades – OECD Countries



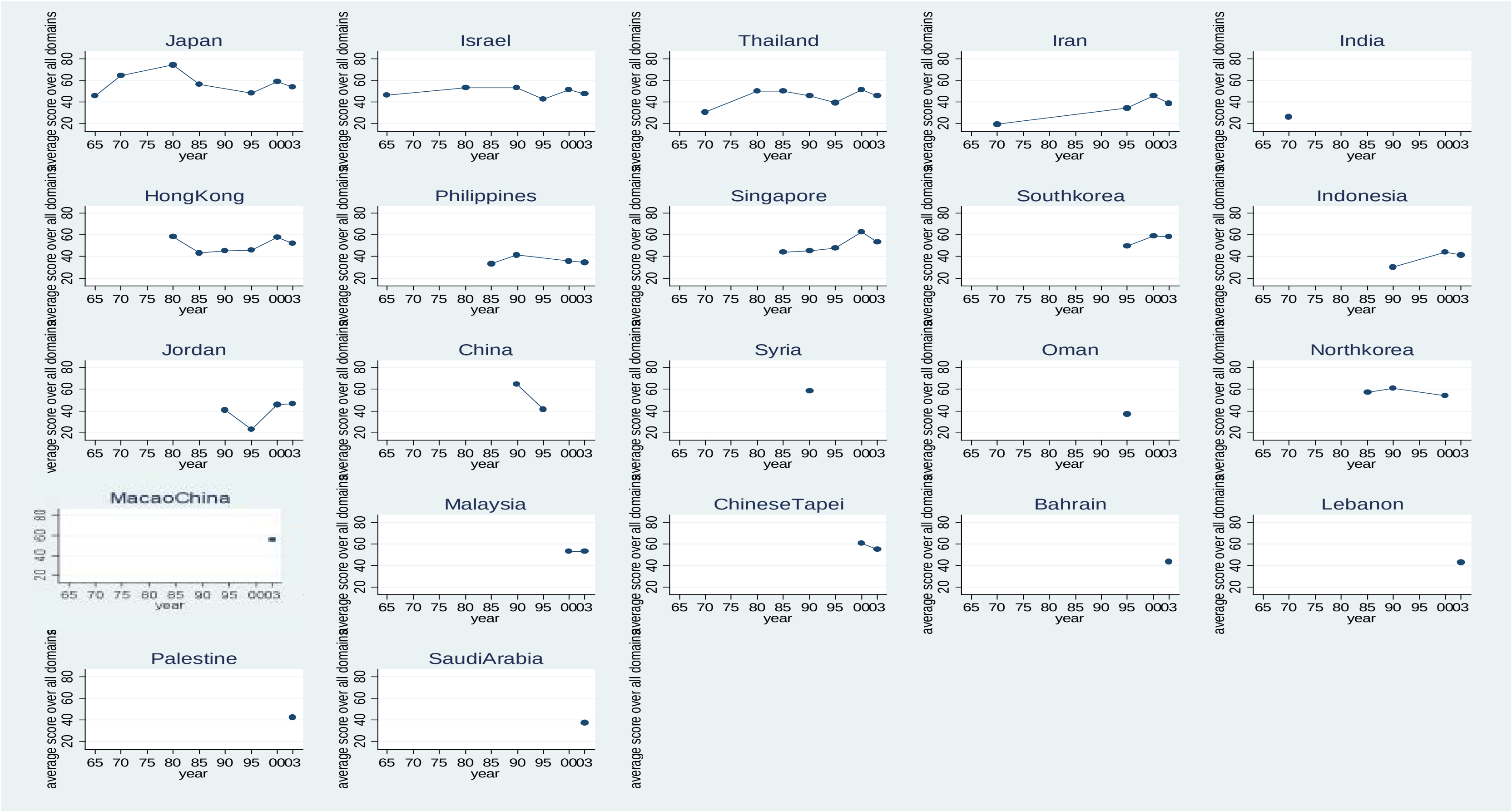
Notes: Every Marker indicates the average test score for the respective country averaged over all test domains (Math, Reading and Science) and all grade levels

Figure E2: Test Score Trends over Time averaging over all Test Domains and Grades – Non-OECD European Countries



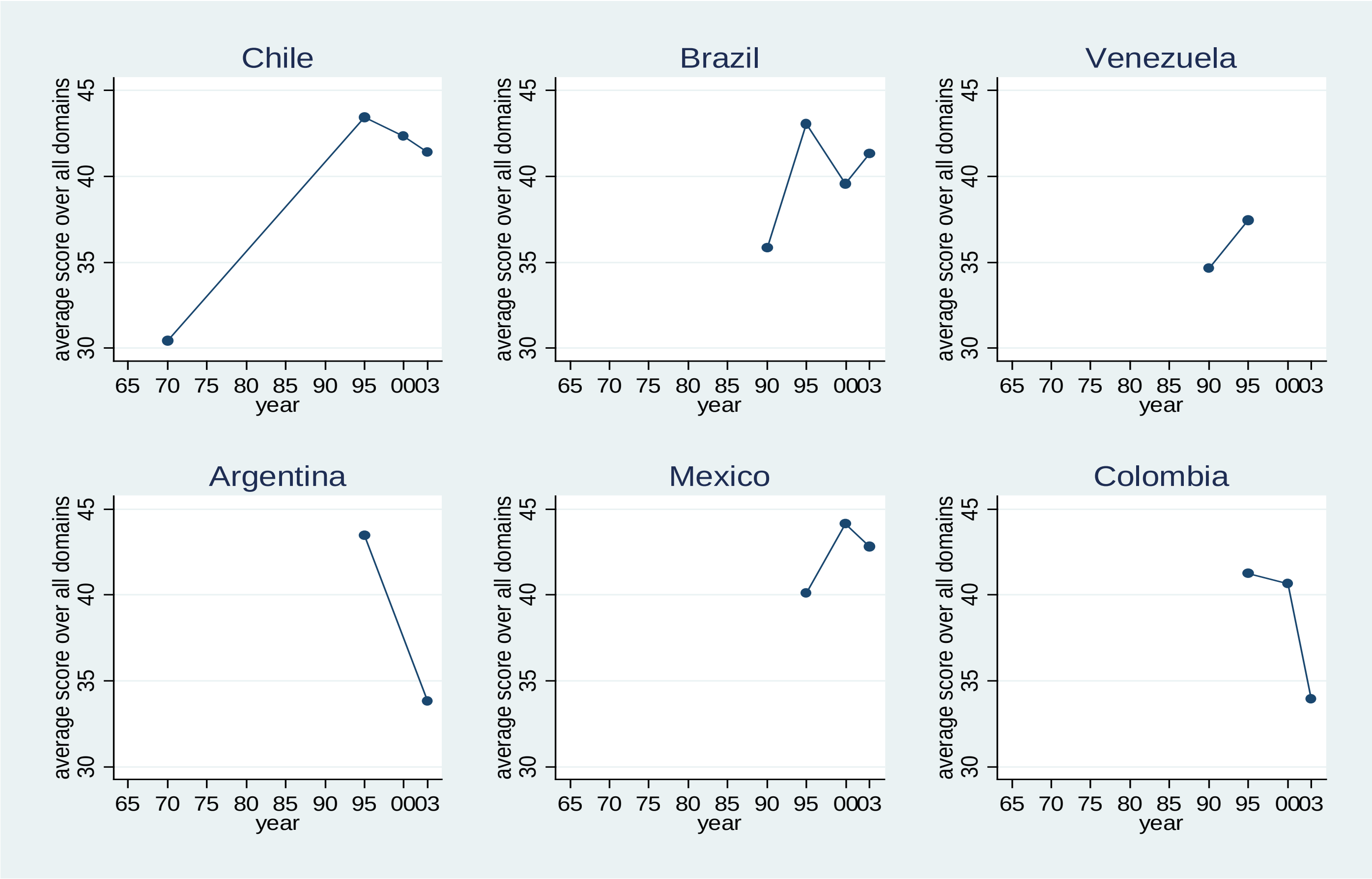
Notes: Every Marker indicates the average test score for the respective country averaged over all test domains (Math, Reading and Science) and all grade levels

Figure E3: Test Score Trends over Time averaging over all Test Domains and Grades – Asian Countries



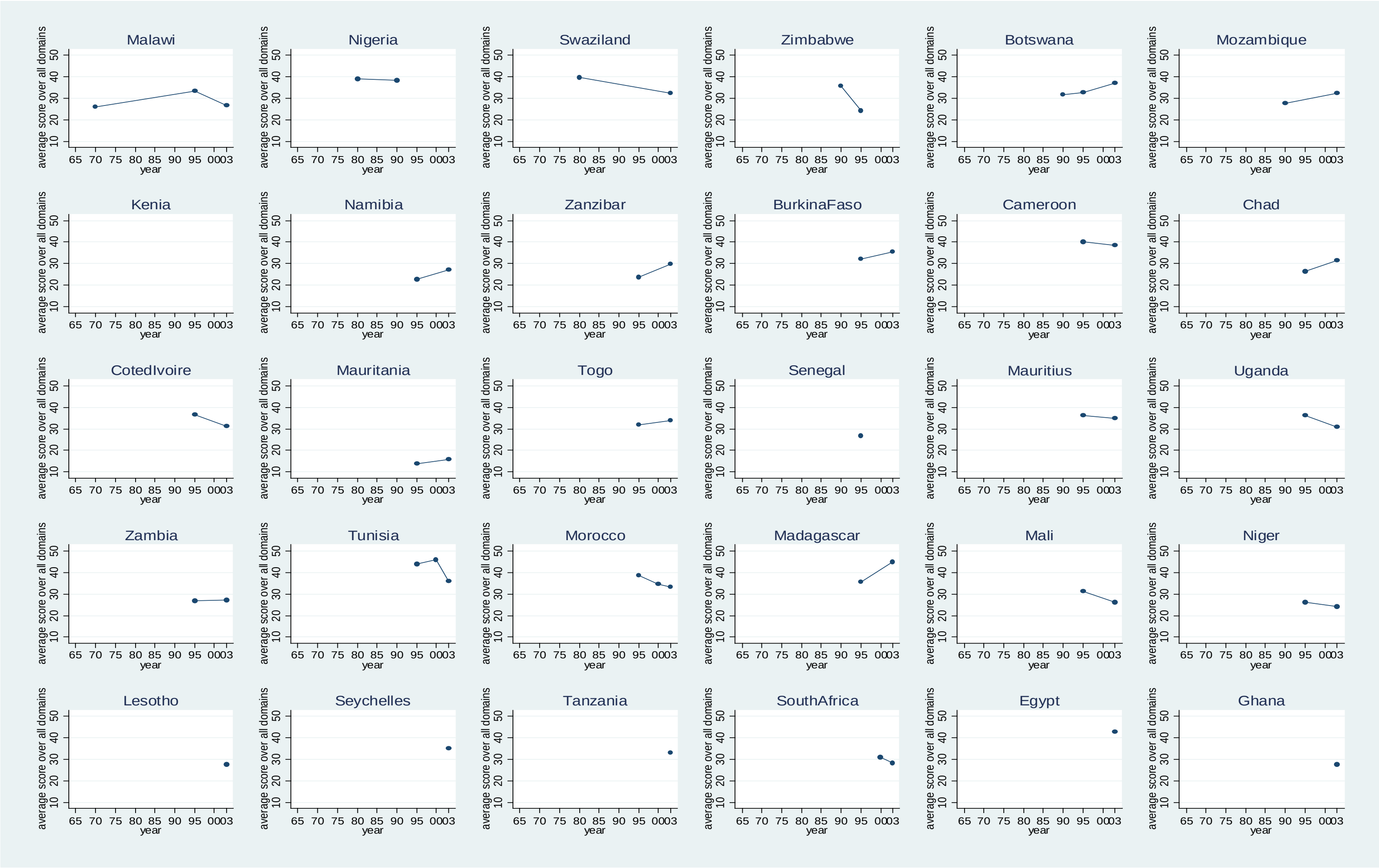
Notes: Every Marker indicates the average test score for the respective country averaged over all test domains (Math, Reading and Science) and all grade level

Figure E4: Test Score Trends over Time averaging over all Test Domains and Grades – Latin American and Caribbean Countries



Notes: Every Marker indicates the average test score for the respective country averaged over all test domains (Math, Reading and Science) and all grade levels

Figure E5: Test Score Trends over Time averaging over all Test Domains and Grades – African Countries



Notes: Every Marker indicates the average test score for the respective country averaged over all test domains (Math, Reading and Science) and all grade levels