International Student Assessment
Performance and Spending

John M. Krieg

FRASER INSTITUTE
2019

2019
Fraser Institute

International Student Assessments
Performance and Spending

by John M. Krieg

Contents

Executive Summary

The Programme for International Student Assessment (PISA) is an international series of tests that is commonly used to compare a nation's academic performance against other nations. A nation's PISA results are a function of a number of factors: the quality and access of education services, the composition of students, health, and nutrition to name a few. This report explores the role of national income and national spending on education to identify cross-country differences in PISA scores. Specifically, this research studies 72 countries over a 15-year period to explore the impact on PISA scores of two measures of national income and per-capita education spending, both corrected for inflation.

There are two findings about per-capita education spending. First, there is a positive relationship between education spending and PISA results—countries that spend little on education have the lowest PISA outcomes while countries that spend a lot tend to have higher PISA scores. Second, the impact of additional educational spending on PISA outcomes is different for low-spending countries than it is for high-spending countries. Specifically, an increase in educational spending by a currently low-spending country raises PISA scores significantly. A back-of-the-envelope calculation suggests that the costs of increasing educational spending for the currently lowest-spending countries could be recouped by increased economic growth rates within a little more than a year. On the other hand, increased education spending by high-spending countries is expected to change PISA results very little. Some of the difference between low- and high-spending countries is expected because of diminishing returns. High-spending countries have already purchased many educational amenities that have contributed to their already high test scores and, unlike the case of low-income countries, purchasing additional educational services is unlikely to contribute significantly to education outcomes.

This report also answers a basic question: what is the relationship between national income and PISA scores? PISA scores are a positive function of real GDP per capita: wealthy nations score higher on the PISA than poor ones. However, this fact too is nuanced. As national income rises, the impact on PISA scores is much larger for poor countries than rich ones. In other words, increasing national income has a much more positive effect for low-income countries than high-income countries—a fact likely related to the diminishing returns of education spending on PISA outcomes. This is also important when considering international rankings of educational attainment. A rough estimate is that real GDP per capita explains about ⅓ of the differences in PISA test scores among countries. This suggests that failing to account for national income would cause one to overstate the impact of educational systems when comparing PISA outcomes.

For policy makers, these results present a number of possibilities. First, from an international perspective, the fact that low-spending and low-income countries benefit most from additional educational intervention suggests that resources are better spent in these countries than in others. From a national perspective, the ability of a nation to climb the PISA rankings is limited: the data suggest that a nation's scores vary little over time and that significant movements would require large changes in educational resources, especially for well-developed countries. However, care must be taken in any policy prescription based upon international comparisons using standardized tests. Comparing nations with different educational standards, environments, and cultures is inherently difficult. Indeed, much of the difference in PISA scores among countries is attributable to unmeasured, country-specific factors. Policies to increase educational effectiveness need to account for these differences.

Introduction

The Programme for International Student Assessment (PISA) is a triennial international series of academic tests sponsored by the Organisation for Economic Co-operation and Development (OECD, 2018b: http://www.oecd.org/pisa/). Along with the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), the PISA is commonly used to compare a nation's academic performance against that of other nations. In making these comparisons, it is natural to inquire what national factors contribute or explain the observed differences. This paper explores the role of national income, measured by per-capita GDP, and the impact of national-level educational spending on PISA outcomes.

Data

The OECD Programme for International Student Assessment (PISA) is a collaborative effort among OECD countries to measure the academic preparation of 15-year-olds. The assessment is intended to be forward looking by focusing on the extent to which students use their knowledge to meet real-life challenges. The first PISA was conducted in 2000 in 32 countries using written tasks answered in schools under supervised test conditions. Since 2000, the PISA has been given six additional times (every three years) with the most recent version having been given in 2018. However, the 2018 data has yet to be released so this study focuses on the six tests starting in 2000 and concluding in 2015. In 2015, the PISA was given in 35 OECD countries and 37 "partner" countries. While the PISA has been regularly given in OECD countries, the partner countries can elect to participate in the PISA. The resulting addition and attrition of countries over time mean that the PISA has been given in over 90 countries since 2000. Each country must test at least 5,000 students unless the country has less than that number in a grade, in which case the PISA reports all students in a cohort. In early years, the PISA test was expected to take about six to seven hours of time to complete. Modern, computer-based examinations take significantly shorter time for students to finish.

The PISA data come in one of two formats: individual-level records and country-level aggregated reports. The individual records contain response information from students linked to classroom, school, and country characteristics. This data has been used to examine a number of educational questions, including the relative effectiveness of public versus private schools (Vandenberghe and Robin, 2004), class size impacts (Denny and Oppedisano, 2013), the effect of grade retention (Garcia-Gracia-Perez, Hidalgo-Hidalgo, and Robles-Zurita, 2014), the role of immigrants in education outcomes (Jensen and Rasmussen, 2011), tutoring (Choi, Calero, and Excardibul, 2012), peer effects (Schneeweis and Winter-Ebmer, 2008), and tracking (Hanushek and Woessman, 2006).

The country-level PISA data consists of average reading, science, and mathematics scores for the entire tested population of the country. It also includes the associated standard errors of each country's average score.[1] In 2000, the student-level raw scores of the PISA are scaled in such a way that the average score in each domain (math, science, and reading) averages 500 with a student-level standard deviation of 100. In subsequent cycles of the PISA, item response theory[2] and mean-mean linking[3] has been employed to maintain a similar student-level average and standard deviation, allowing for comparison

1. These are publicly available at the *PISA International Data Explorer* (National Center for Education Statistics, 2017: <http://nces.ed.gov/surveys/pisa/idepisa>.

2. Item response theory is a mechanism to account for the difficulty of a test item, student ability, and the likelihood of randomly guessing the correct answer. It can be used to standardize different exams to make them comparable across time or people. See Mazzeo and von Davier (2013) for its application to the PISA.

3. Mean-mean linking is used on a subsample of PISA questions that are repeated from one test to the next in order to ensure the tests are similarly scored over time. See Robitzsch and Ludtke (2018) for details.

of country results over time. Importantly, these averages are constructed in such a way that similarly able students taking the PISA at different points in time should have similar scores. **Table 1** reports the country-level means and standard deviations of the PISA data. At least two things stand out in this table. First, PISA scores for mathematics and science are unavailable for early years of the test (the math test was introduced in 2003, the science in 2006). Second, the number of participating countries has increased over time. What is not apparent in table 1 is that a number of partner countries have exited and entered the PISA at different times. Thus, some of the instability of the mean scores across time have to do with the composition of countries entering and exiting the PISA. For instance, much of the decrease in average score over time is because new countries have entered the PISA. Since the original set of countries were the well-to-do OECD nations with well-developed education systems, most of the later entries were poorer countries that tended to reduce the overall average scores. The average reading score for countries that participated in all six test regimes is 488 while the average country score for the seven countries who just started participation in 2015 is 380.[4]

Table 1: Means and standard deviation of country-level PISA results

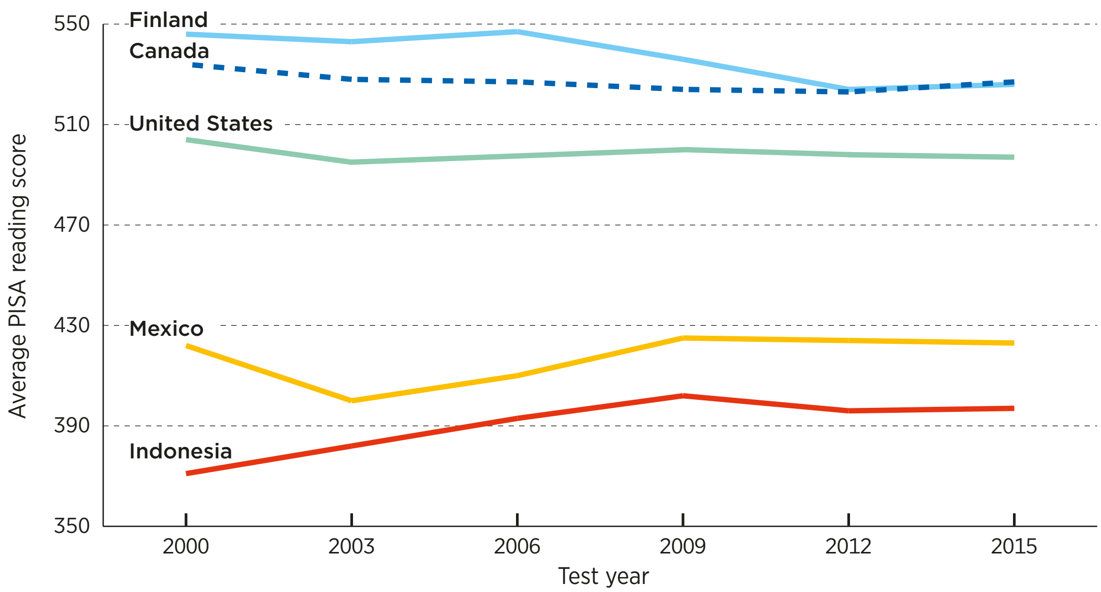
	Reading	Math	Science	N
2000	470.07 (53.72)			41
2003	482.02 (41.26)	486.33 (53.30)		39
2006	458.85 (58.15)	467.58 (58.81)	472.08 (54.48)	56
2009	462.36 (50.90)	464.31 (57.49)	469.25 (55.06)	63
2012	471.63 (45.85)	469.71 (52.29)	476.20 (49.55)	63
2015	460.34 (51.86)	459.46 (54.77)	464.37 (49.95)	67

Note: Country-level standard deviation in parenthesis.

It is necessary to understand two additional facts about PISA scores prior to further investigation. First, there are large variations in scores across countries and second, the variation in scores within countries over time is small. **Figure 1** presents average reading scores for a selection of countries observed in all six PISA rounds: the three North American countries, the highest scoring country (Finland), and the lowest scoring country (Indonesia). The variation across countries is quite large: Finland and Indonesia are separated by almost 200 points—roughly two student-level standard deviations. Secondly, the variation within countries is relatively small.

4. The seven new countries that entered PISA in 2015 were Algeria, Dominican Republic, Georgia, Kosovo, Lebanon, Malta, and the Republic of Moldova.

Figure 1: Average PISA reading scores, selected countries, 2000–2015



Source: OECD, 2000–2015.

Among the countries in figure 1, the largest variation over the six observed years was in Indonesia, which had a 31-point difference between 2000 and 2012. Canada, on the other hand, had an 11-point difference between those same years. For the entire sample, the standard deviation in reading scores between countries was 57, almost five times as large as the standard deviation within countries of 11.5. The relative sizes of these variations suggest that between-country analysis of PISA scores is more likely to find an impact of GDP on PISA scores than a within-country analysis.

In measuring economic activity, I construct three measures that plausibly influence PISA test scores. The first two are different measures of real GDP per capita (in 2011 US dollars).[5] The first of these measures is real GDP per capita on the expenditure side, or GDP^e. As described in Feenstra, Inklaar, and Timmer (2015) GDP^e is intended to provide a comparable measure of the standard of living across countries. GDP^e is constructed as nominal GDP, deflated by the prices that are converted from national currencies to US dollars using the purchasing-power-parity exchange rate and adjusted to allow for comparisons over time using a chain-weighted price methodology. The second measure of GDP per capita, GDP^o, is a measure of output-based real GDP that is intended to capture the productivity differences among countries. Unlike GDP^e, GDP^o includes exports and imports and their associated prices in computing output and will differ from GDP^e to the extent that a country differs in its terms of trade. Because there are a handful of countries that have high levels of international trade in our data, the effect of GDP^o on PISA scores may produce different results than the effect of GDP^e. With that said, the pooled, cross-sectional correlation between GDP^o and GDP^e is .79 so there is likely small scope for differences when comparing these measures.

5. Both per-capita measures are available through the *Penn World Table*, version 9 (Groningen Growth and Development Centre, 2019; Feenstra, Inklaar, and Timmer, 2015).

The third measure of economic activity is a country's per-capita spending on non-tertiary education (in 2011 US dollars).[6] The advantage of this measure is that per-capita expenditures on education should be more closely tied to education outcomes than measures of per-capita GDP. The disadvantage of this data is that it covers only 84.7% of observed country-years. For the remaining 15.3% of observations, I impute missing observations.[7] **Table 2** reports summary statistics for these three measures by PISA-year. It is important to remember that the composition of countries changes over time in table 2, as it does in table 1. Specifically, more countries participate in the PISA over time and, as the PISA was originally given in OECD countries, the additional countries that join in later periods tend to have lower per-capita GDP and spend lower amounts on education.

Table 2: Means and standard deviation of economic activity measures

	GDP^e (real 2011 US$)	GDP^o (real 2011 US$)	Per-capita GDP (real 2011 US$) spent on non-tertiary education	N
2000	25,163 (14,726)	24,217 (13,815)	1262 (855)	41
2003	27,796 (13,716)	27,013 (13,073)	1413 (823)	39
2006	28,345 (19,933)	27,610 (19,552)	1392 (1002)	54
2009	29,753 (19,191)	28,662 (18,976)	1492 (978)	62
2012	34,074 (24,754)	33,347 (25,444)	1722 (1114)	62
2015	33,662 (21,528)	31,401 (19,688)	1844 (956)	67

Note: Country-level standard deviation in parenthesis.
Source: Groningen Growth and Development Centre, 2019: *The Database: Penn World Table version 9.1*; Feenstra, Inklaar, and Timmer, 2015.

6. The OECD provides estimates of the fraction of GDP spent on non-tertiary education. I multiply this fraction by $RGDP^e$ to calculate per-capita spending on non-tertiary education. The fraction of GDP spent on non-tertiary education is available from the OECD (2018a, 2019). I multiply by GDP^e rather than GDP^o because GDP^e is the best proxy for a nation's standard of living and does not account for imports and exports, which are likely to play a small role in education spending.
7. I imputed this data in two steps. The OECD provides country-year data on the percentage of GDP spent on all education. Because this data includes tertiary education that occurs after the PISA, I used it only for imputing missing World Bank data. When a World Bank observation of non-tertiary education spending was missing, I calculated the annual growth rate using the OECD measure and applied that growth rate to observed World Bank data to fill in missing observations. For most of the remaining missing observations, I linearly extrapolated any missing values. The few remaining observations that are missing are at the beginning or end of the data set, which precludes linear extrapolation or estimation with growth rates.

Methodology

As an initial exploration, I begin by investigating the relationship between PISA scores and expenditures by estimating

$$PISA_i = \beta_0 + \beta_1 Expenditures_i + \varepsilon_i \quad (1)$$

where PISA is one of the three average PISA scores (Reading, Math, or Science) in country i, and Expenditures are country i's measures of GDP^e, GDP^o, or the per-capita spending on education. I estimate equation (1) separately for each PISA-year to avoid inefficiencies that result from pooling countries over time.

The OECD (2012) explored a variant of equation (1) and found that countries with low levels of GDP per capita have a positive value of β_1 while countries with medium or high levels of GDP per capita have values of β_1 no different from zero. The OECD interpreted this as meaning that increasing national wealth for low-income countries raised education performance while national income had no impact on education outcomes for high-income countries. In order to follow up on the OECD's result, I also estimate:

$$PISA_i = \beta_0 + \beta_1 Expenditures_i + \beta_2 High_i + \beta_3 Expenditures_i \times High_i + \varepsilon_i \quad (2)$$

where $High_i$ is a binary variable equal to 1 if a country is a high-income country and 0 if a low-income country. Effectively, equation (2) allows for a different slope and intercept for high-income countries relative to low-income countries. If the OECD's argument that low-income countries transform additional Expenditures into higher PISA results while high-income countries do not, I would expect $\beta_1 > 0$ and $\beta_3 = -\beta_1$. In effect, equation (2) is a spline regression with the discontinuity in the spline occurring at the boundary between "high-income" and "low-income" countries. In this case, I perform a grid search over different levels of Expenditures to find the boundary that produces equation (2) with the smallest residual sum of squares.[8] I constrain these models in such a way that the linear estimate for low-income countries meets the estimate for high-income countries at the boundary between low and high income.

Equations (1) and (2) present unbiased estimates of β_1 only if the unexplained portion of PISA scores (ε_i) are not correlated with a country's Expenditures. There are a number of reasons that one might be suspicious of this approach. First, it is likely that education attainment increases Expenditures. In the case where Expenditures is measured as per-capita GDP, it is likely that countries with high education outcomes generate higher per-capita income. If this is the case, then β_1 will not represent the impact of per-capita GDP on PISA scores but instead simply represent a correlation between high GDP countries and high PISA-scoring countries. This is also a concern

8. I recognize that a search process finding the minimum residual sum of squares invalidates classical statistical inference; however the purpose of this is simply to best describe the relationship between Expenditures and test results rather than test competing hypotheses.

when the measure of Expenditures is the per-capita level of spending on education. If higher test scores (or education achievement) generate political enthusiasm in a country for education spending, then β_1 will, again, be biased. Both of these examples will cause the estimated β_1 to be larger than it really is—in other words equation (1) and (2) will overstate the impact of Expenditures on PISA scores. Of course, it is possible that higher PISA scores actually reduce political enthusiasm for education spending—after all students are doing relatively well so why should there be more spending? If this were the case, then the estimates of β_1 will be smaller than they actually are.

A second concern arises in that these equations omit a number of variables that likely affect PISA scores. If these omitted variables are unrelated to Expenditures, the effect will be to simply reduce the precision with which these equations are estimated. However, a more insidious case arises if the omitted variables are correlated with Expenditures: in this case the model estimates are biased. A plausible example has to do with secondary-school enrollment rates. It is well documented that secondary-school enrollment rates are lower in poorer countries (Glewwe and Lambert, 2010). If only the best students in poor countries enroll in secondary school and are then tested on the PISA, then we would measure a different selection of students in poor countries relative to wealthy ones where enrollment rates approach 100%. The resulting impact on these models are often difficult to determine, but plausibly mean that PISA scores over-estimate education attainment in poor countries.

One approach to lessening these concerns is to make use of the repeated nature of the PISA tests. In this data set, we observe 41% of countries in six periods and an additional 15% countries in five periods. The repeated observations of the same country allows for exploration of the data longitudinally. Specifically, I estimate

$$PISA_{it} = \beta_0 + \beta_1 Expenditures_{it} + \alpha_i + \varepsilon_{it} \quad (3)$$

where the t subscript indicates the PISA test period and the α_i represents a country fixed effect. The addition of country fixed effects mitigates some, but not all, of the concern with bias arising from omitted variables and education outcomes causing Expenditures. The addition of fixed effects means that the year-to-year change in a country's PISA scores are explained by year-to-year changes in Expenditures. Thus, only if *changes* in Expenditures are responsive to *changes* in PISA scores would bias arise. The fixed effects also account for any country-specific effect that does not change over time. For instance, this approach will account for differences in PISA scores caused by differences across countries in enrollment rates as long as those enrollment rates are relatively stable within a country over time. This approach also accounts for stable characteristics of countries like culture, attitudes towards education, environment, and workforce quality. However, this benefit comes at a cost. Recall from figure 1 that within a country, there is little variation in PISA scores over time. The fixed-effects approach is dependent upon being able to explain this within-country variation. Since there is little within-country variation, the fixed-effects approach will identify a relationship between Expenditures and PISA scores only if a very strong relationship exists.

One final concern has to do with the fact that PISA scores are average scores across individual students within a country. These average scores are measured with error partly because only a sample of a nation's students are observed and also because each student taking an exam does so under different conditions.[9] Because of this, the OECD provides standard errors for each country's test results. As a result, I weight each equation by the inverse of these standard errors. This results in countries that have more precisely measured PISA results as having more influence on the regression results than countries with less precisely measured test scores.

9. There are a number of other reasons that the national average PISA score is measured with error. See Robitzsch and Ludtke, 2018 for details.

Results

We begin our exploration of Equation (1) in **table 3**, which displays estimates of the impact of Expenditures on PISA scores for each test and each year of the test. Panel 3A presents results using the expenditure-side real GDP per capita as the measure of Expenditures; Panel 3B uses the consumption-side measure of per-capita real GDP; and Panel 3C uses per-capita real GDP spent on non-tertiary education. The messages of Panels 3A and 3B is straightforward: countries that have higher per-capita real GDP score higher on all three PISA tests and do so in every year that these tests are given. For instance, Panel 3A's 2000 reading estimate is 2.71 indicating that, for each increase of $1000 in per-capita real GDP, a country can expect their average reading PISA score to increase by 2.71 points. While all of the estimates in Panels 3A and 3B are statistically significant, it is less obvious that these coefficients are quantitatively important. For instance, in 2000 the cross-country standard deviation in reading PISA scores was 53 points (table 1) so an increase of 2.71 points represents an increase of about .05 standard deviations. In other words, a country would need to increase its per-capita real GDP by $20,000 to achieve a one (country-level) standard deviation increase in reading PISA scores. This 2.71 estimate is rather high; among other estimates in Panel 3A there is only one higher—Math, 2003—and many estimates are about ½ the size of this figure so it is likely this is overstates the effect of per-capita real GDP on PISA scores.

Panel 3C tells a slightly different story. Panel 3C presents results from equation (1) where the per-capita spending on non-tertiary education is used to explain PISA scores. Taking the 2000 reading coefficient of .043 as an example, this indicates that a nation that increases the per-capita spending on education by $100 can expect their reading PISA average to increase by 4.3 points. Given the average per-capita spending on education in 2000 was $1,262, this increase in spending represents an 8% increase in spending—to achieve a 4.3 increase in the PISA score, which represents about 1⁄10 of a within-country standard deviation. It is important to note that all coefficients in Panel 3C are statistically significant and about the same magnitude as found for reading in 2000, suggesting that countries that spend more on education receive higher PISA results.

One possibility is that these results differ across countries based upon the country's level of Expenditures. This possibility is explored using Equation (2). Because these results are more difficult to interpret in a table, I present the outcomes in **figures 2a, 2b, and 2c** (reading), **figures 3a, 3b, and 3c** (math), and **figures 4a, 4b, and 4c** (science) for selected years. These figures present each country's Expenditure-PISA score combination (with codes described in the **Appendix**, p. 29), the best fitting spline regression (solid horizontal line), and the demarcation between "high-income" countries and "low-income" countries (dashed vertical line) that minimizes the sum of squared residuals of the spline regression.

Table 3: Estimates of the impact of expenditures on PISA scores

	Reading PISA	Math PISA	Science PISA
Panel 3A: Expenditure-side real GDP per capita (1000s of 2011 US$)			
2000	2.71***		
	(.559)		
2003	2.32***	3.26***	
	(.458)	(.593)	
2006	2.02**	2.09***	1.97***
	(.793)	(.727)	(.614)
2009	2.23***	2.61***	2.33***
	(.478)	(.475)	(.499)
2012	1.23***	1.37***	1.24***
	(.424)	(.499)	(.445)
2015	1.67***	1.72***	1.49***
	(.341)	(.392)	(.344)
Panel 3B: Consumption-side real GDP per capita (1000s of 2011 US$)			
2000	2.89***		
	(.524)		
2003	2.39***	3.24***	
	(.393)	(.586)	
2006	1.94**	2.01***	1.87***
	(.798)	(.741)	(.625)
2009	2.16***	2.51***	2.23***
	(.520)	(.535)	(.557)
2012	1.11***	1.20***	1.11**
	(.413)	(.478)	(.432)
2015	1.73***	1.75***	1.51***
	(.382)	(.443)	(.382)
Panel 3C: Spending per-capita on non-tertiary education (2011 US$)			
2000	.043***		
	(.008)		
2003	.035***	.047***	
	(.006)	(.009)	
2006	.036***	.036***	.034***
	(.010)	(.009)	(.008)
2009	.040***	.044***	.041***
	(.006)	(.006)	(.007)
2012	.024***	.026***	.024***
	(.005)	(.007)	(.006)
2015	.029***	.032***	.026***
	(.005)	(.006)	(.005)

Notes: Robust standard errors in parenthesis. * p<.1, **p<.05, ***p<.01. All regressions weighted by the inverse of the standard deviation of the average test score.

Figure 2a: Effects of per-capita real GDP on PISA scores for **reading**, spline regressions, 2003, 2015—**expenditure**-side real GDP per capita

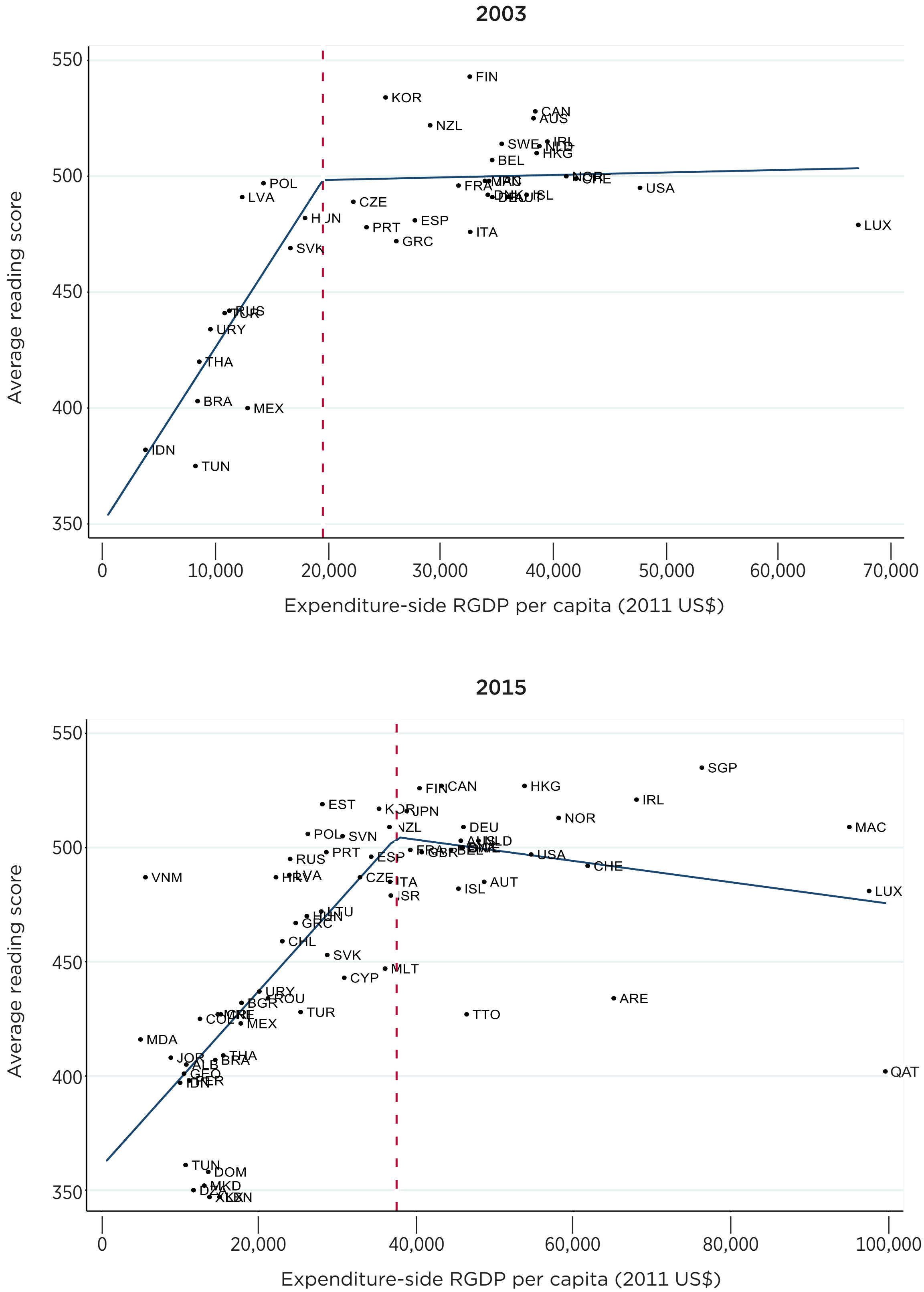


Figure 2b: Effects of per-capita real GDP on PISA scores for **reading**, spline regressions, 2003, 2015—**consumption**-side real GDP per capita

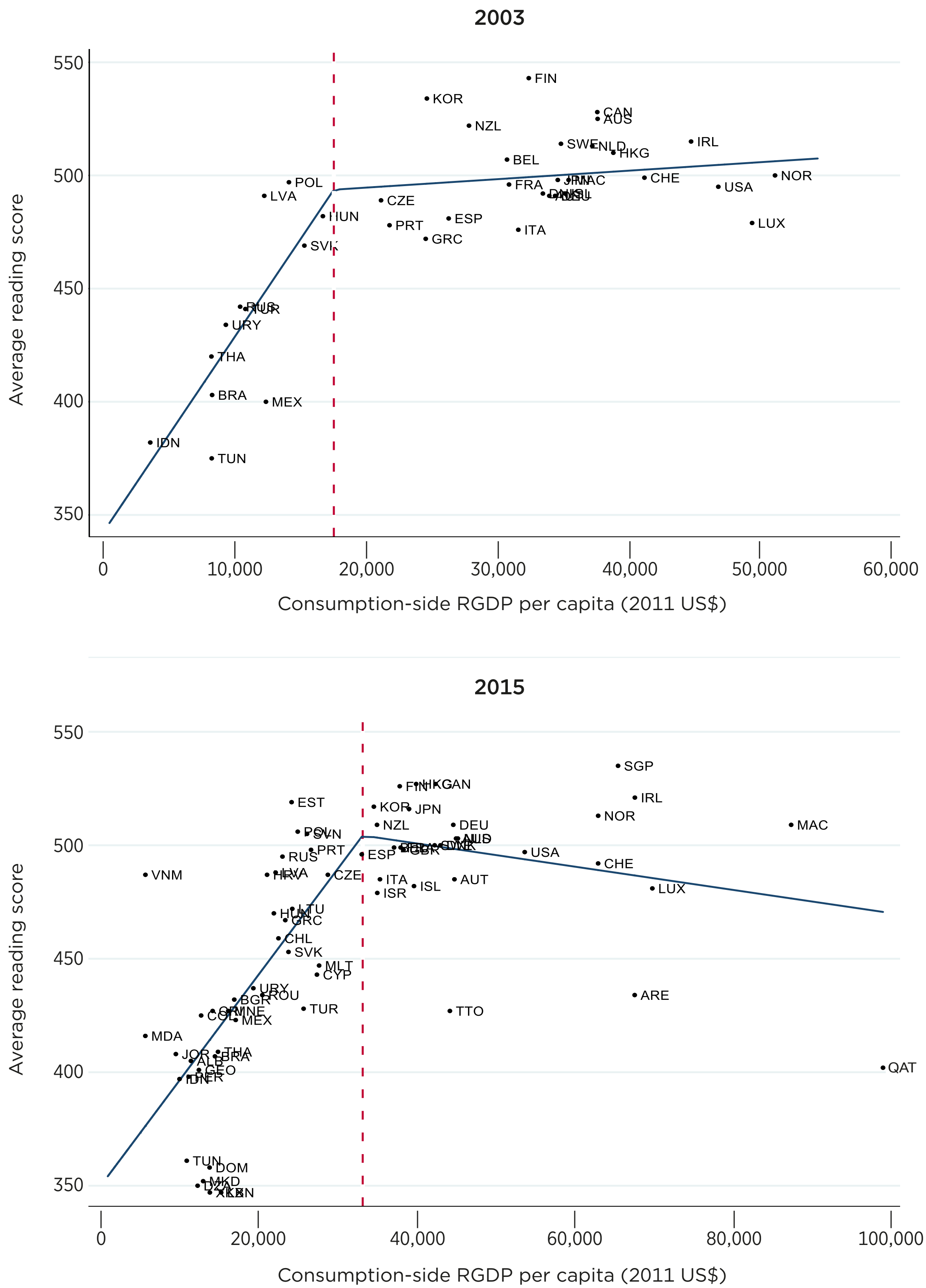


Figure 2c: Effects of per-capita spending on PISA scores for **reading**, spline regressions, 2003, 2015—expenditure-side **education spending** per capita

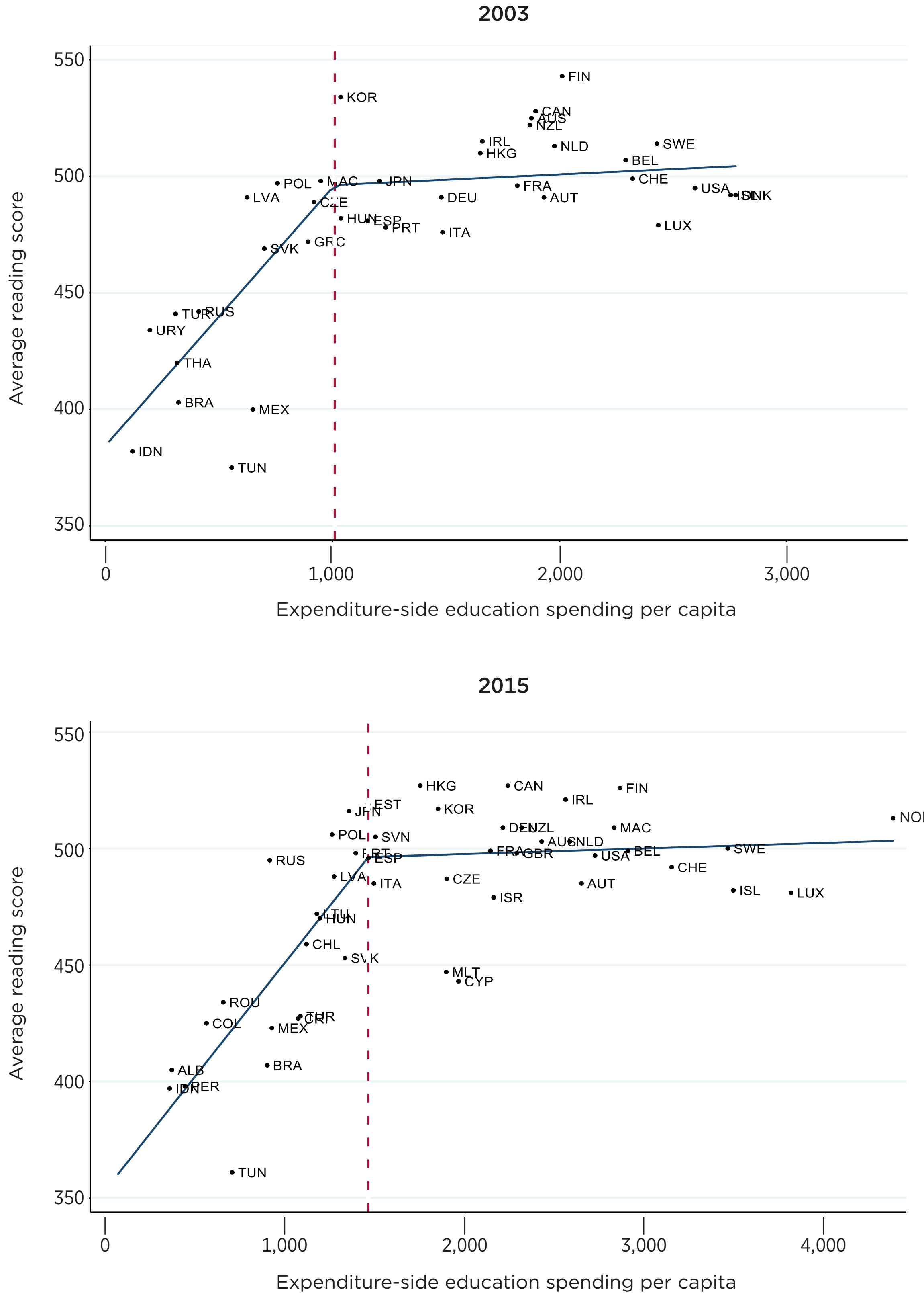


Figure 3a: Effects of per-capita real GDP on PISA scores for **math**, spline regressions, 2003, 2015—**expenditure**-side real GDP per capita

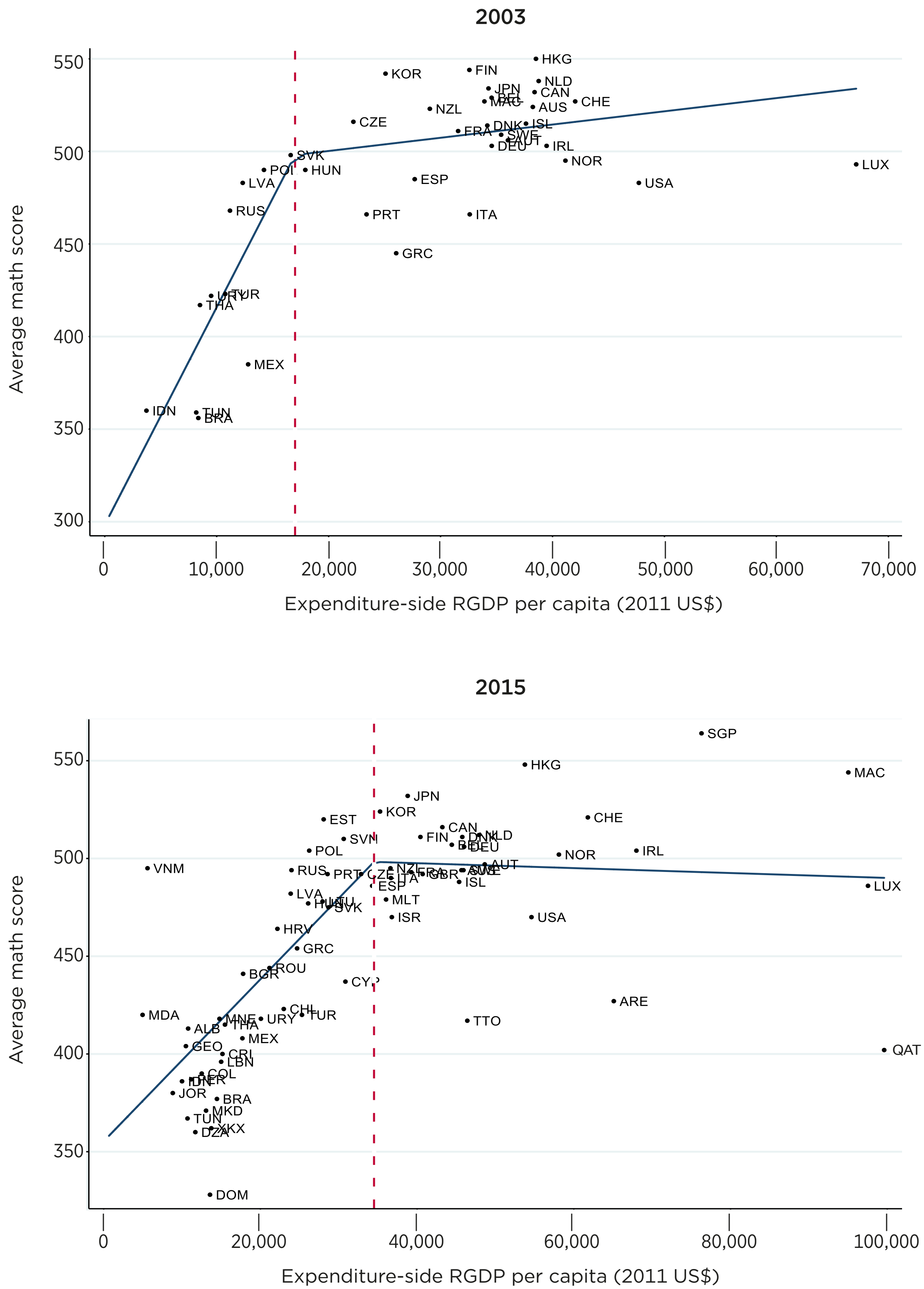
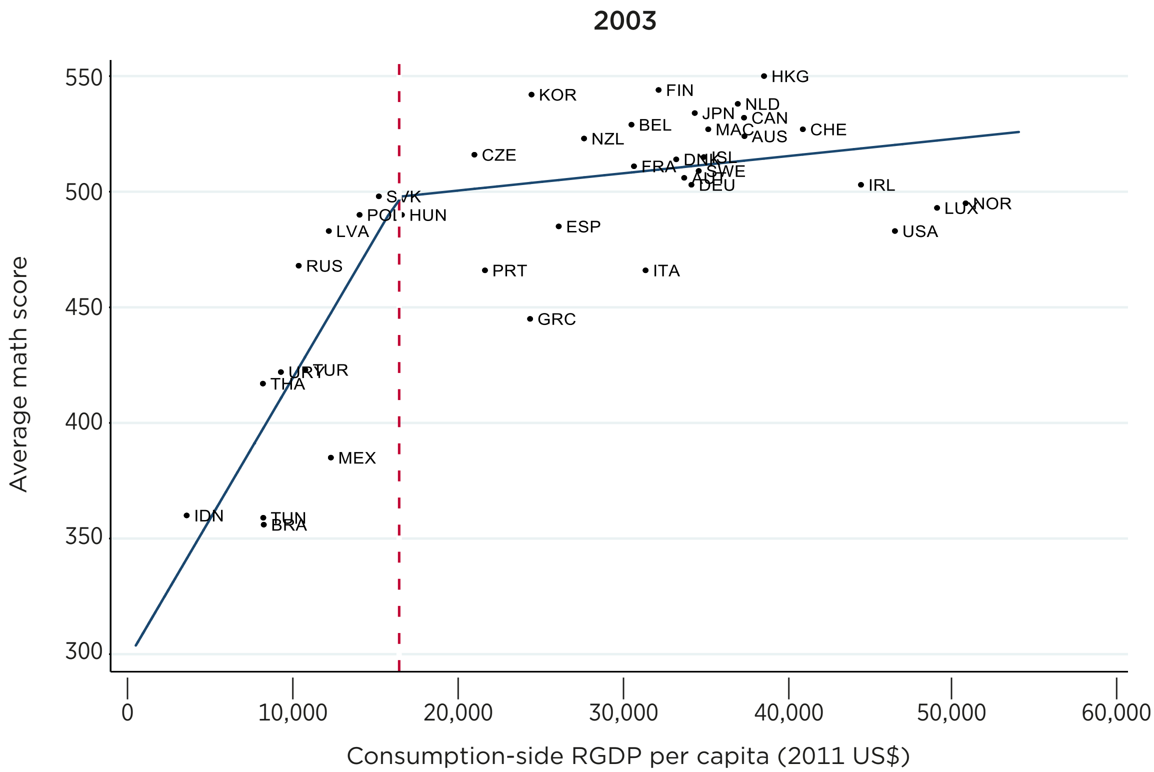


Figure 3b: Effects of per-capita real GDP on PISA scores for **math**, spline regressions, 2003, 2015—**consumption**-side real GDP per capita



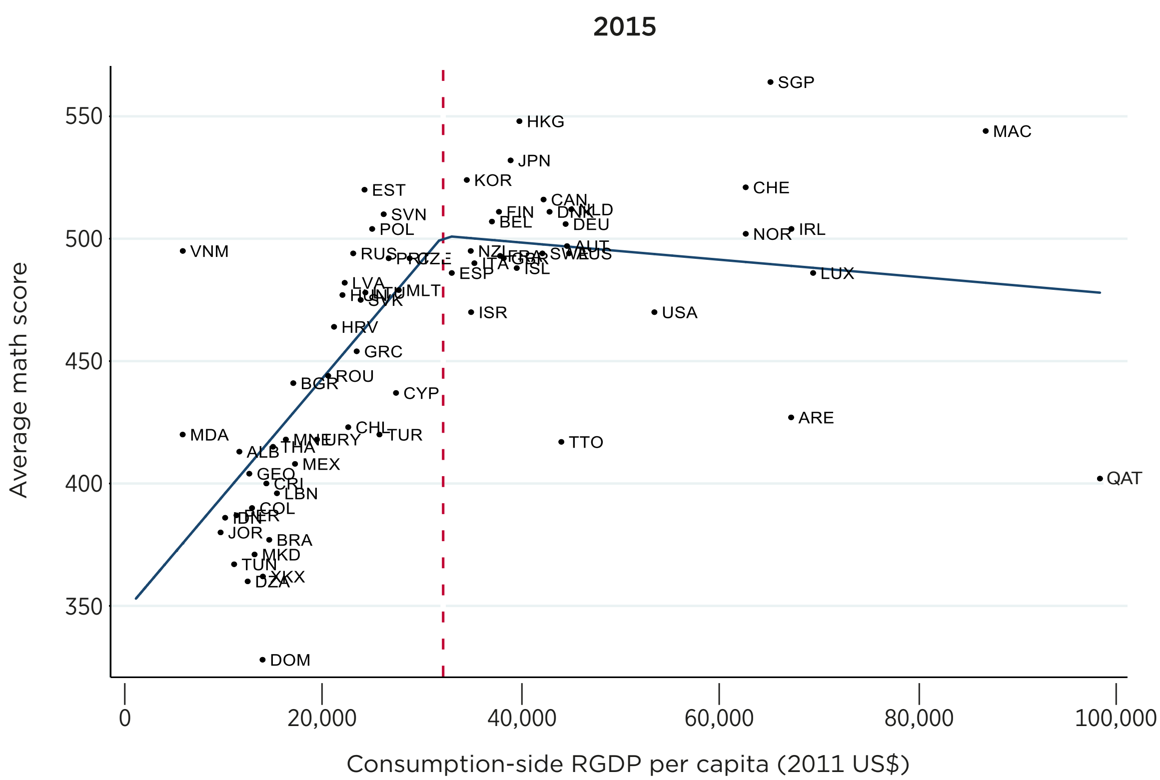


Figure 3c: Effects of per-capita spending on PISA scores for **math**, spline regressions, 2003, 2015—expenditure-side **education spending** per capita

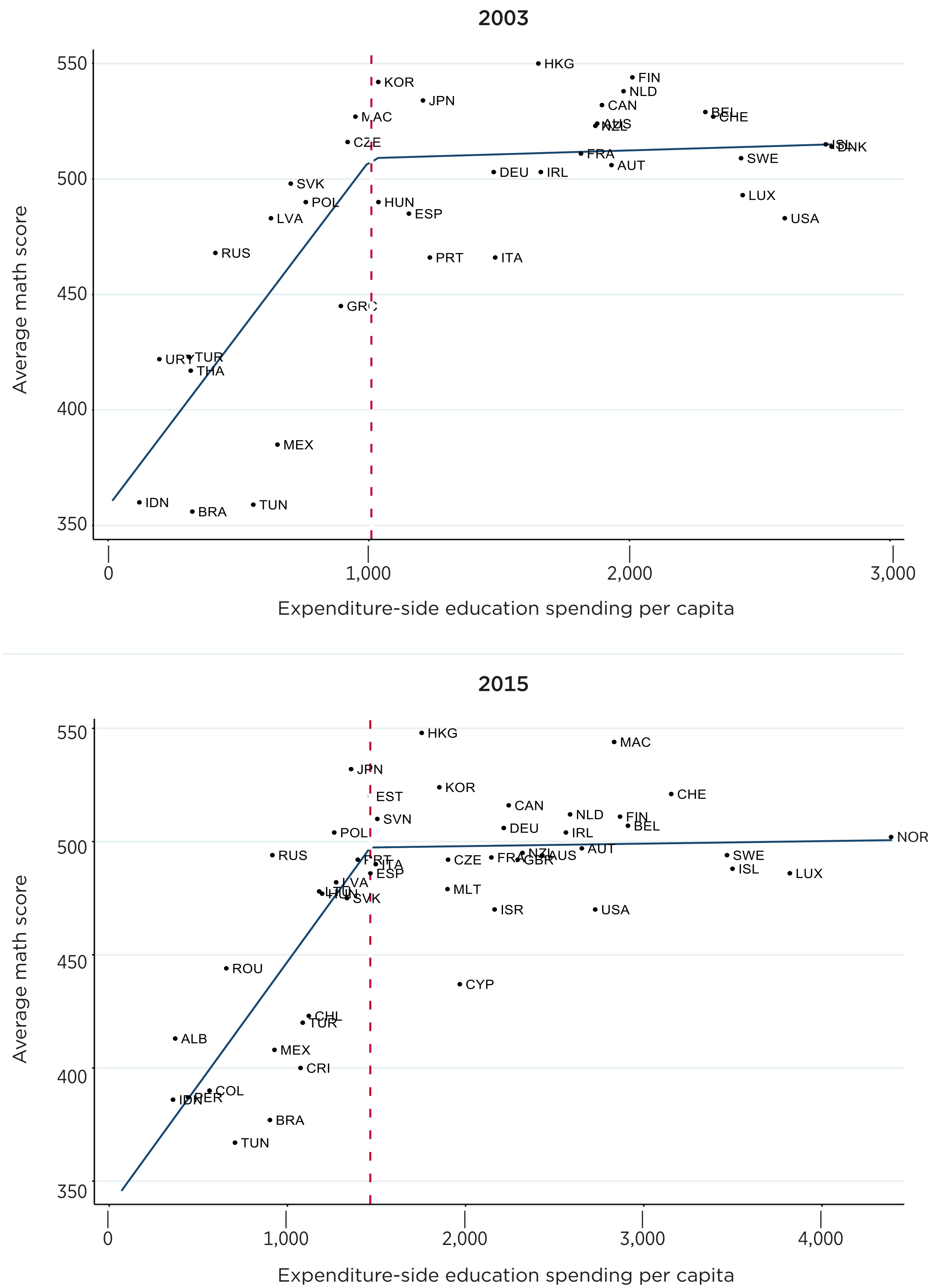


Figure 4a: Effects of per-capita real GDP on PISA scores for **science**, spline regressions, 2006, 2015—**expenditure**-side real GDP per capita

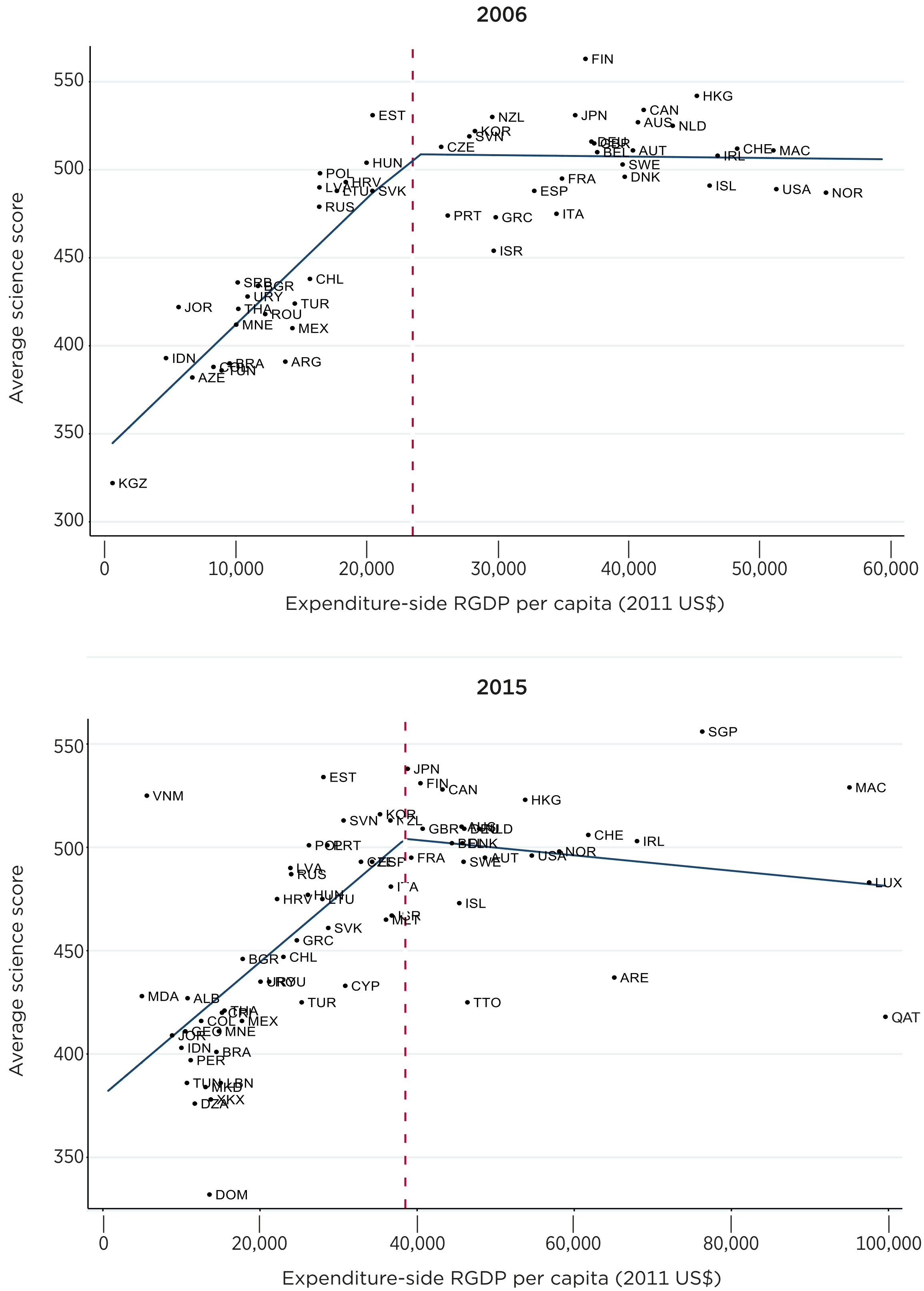
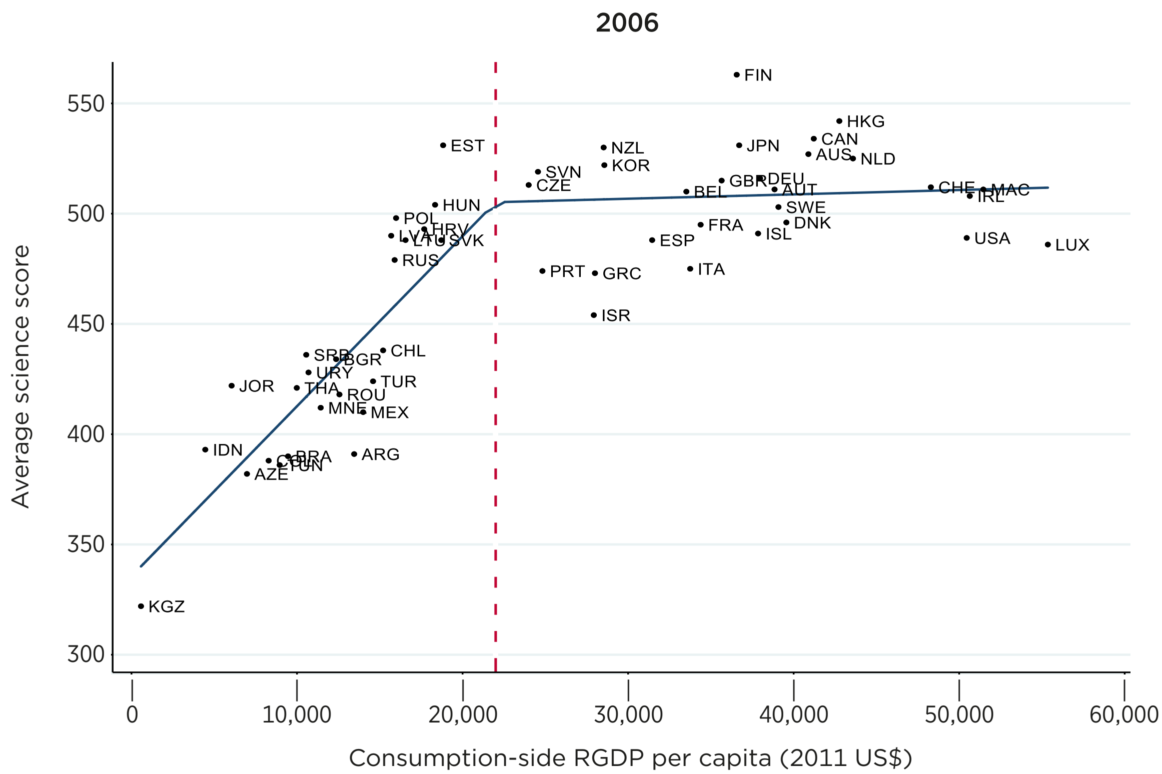


Figure 4b: Effects of per-capita real GDP on PISA scores for **science**, spline regressions, 2006, 2015—**consumption**-side real GDP per capita



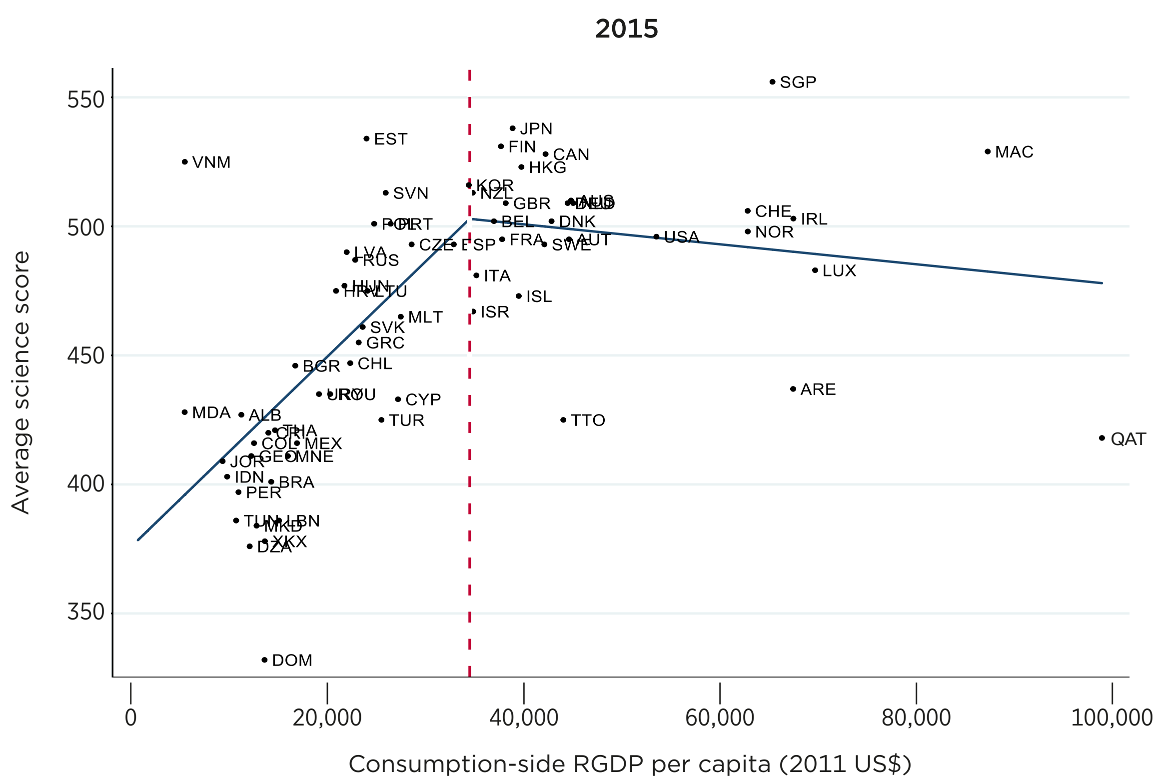
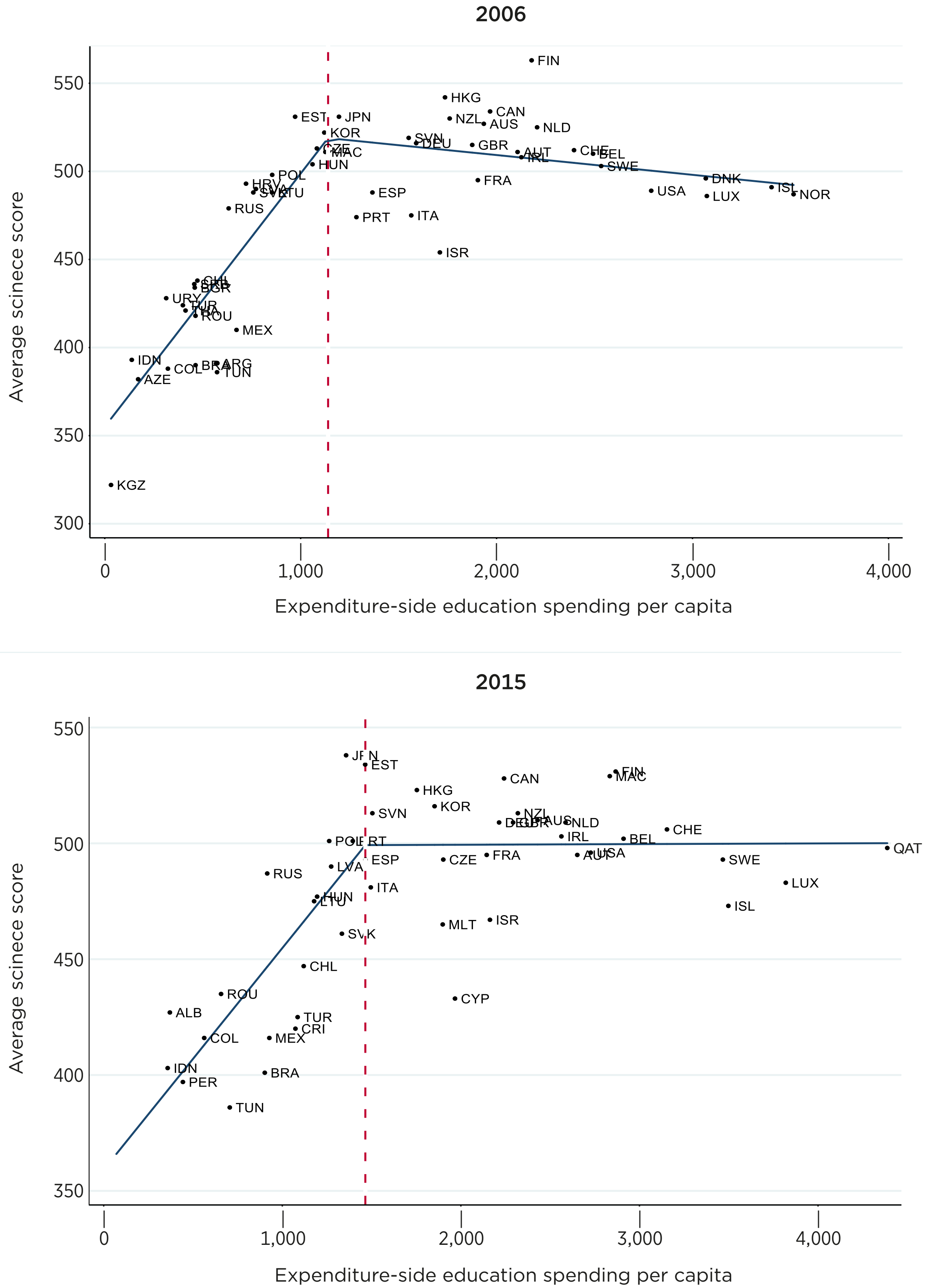


Figure 4c: Effects of per-capita spending on PISA scores for **science**, Spline Regressions, 2006, 2015—expenditure-side **education spending** per capita



Figures 2a–c, 3a–c, and 4a–c demonstrate one clear finding: countries with low GDP per capita have a positive relationship between Expenditures and their average PISA score. This is consistent across years, tests, and the three measures of Expenditures. Taking the reading, 2015, expenditure-side real GDP results (figure 2a, 2015) as an example, the slope of the spline regression to the left of the demarcation is .0038 (se = .0004, t = 9.2) indicating that each additional $1,000 of real GDP per capita these countries earn, the average PISA score is expected to rise by 3.8 points. This is larger than the effect of 2.71 estimated in table 3. In every graph of these figures, the slope of the line for low-GDP countries is positive and statistically different from zero. While these results do not necessarily speak to causation, the take-away is that increasing income among poor countries is related to better PISA test scores.

A second finding is that among high-income countries, that is those to the right of the demarcation line, there is little consistency between Expenditures and PISA scores. Indeed, in every graph of figures 2a–c, 3a–c, and 4a–c, no spline regression to the right of the demarcation line has a slope that statistically differs from zero.[10] This suggests that there is little or no relationship between per-capita GDP and PISA scores for high-income countries. Again, I caution against inferring causation, or a lack of it, because of possible omitted variables and reverse causation between per-capita GDP and education outcomes.

As mentioned earlier, one problem with equations (1) and (2) is that they ignore repeated observations of countries that take the PISA multiple times. **Table 4** presents results from equation (3), which explicitly controls for the time-invariant characteristics of countries through a fixed-effects framework. The interpretation of these coefficients is similar to those found in table 1—each coefficient in panels 4A and 4B represent the impact on a nation's average PISA score given an increase in real GDP per capita of $1,000. The coefficients in panel 4C represent the change in average PISA scores given an increase in per-capita GDP spent on non-tertiary education of one dollar.

The results from table 4 differ from earlier results in a number of ways. First, and most obviously, the coefficients are considerably smaller than the OLS coefficients of tables 2 and 3. For instance, the largest coefficient in table 4, the reading coefficient associated with expenditure-side real GDP per capita of .471, is about ⅓ to ⅕ the size of corresponding reading coefficients in table 3. The smaller coefficient in the presence of fixed effects suggests that much of the earlier measured impact of real GDP on PISA scores is not related to real GDP, but instead to country-specific characteristics. In other words, about ⅔ to ⅘ of the earlier, naïve results were actually caused by a country's time-invariant attributes rather than their levels of real GDP per capita. Second, there is no statistically significant impact of any of the Expenditure measures on math or science PISA scores. The only statistically significant relationships remaining after controlling for country fixed effects are those related to real GDP per capita

10. There are a few outliers like Qatar that tend to reduce the estimated slope in the per-capita real GDP regressions. Even after removing these outliers, all slopes for high-income countries in figures 2a–c, 3a–c, and 4a–c remain statistically no different from zero.

Table 4: Fixed-effects estimates of the impact of expenditures on PISA scores

	Reading	Math	Science
Panel 4A: Expenditure-side real GDP Per Capita (1000s of 2011 US$)			
B_1	.471*** (.153)	.099 (.127)	.014 (.160)
N	323	284	247
R^2	.949	.975	.977
Panel 4B: Consumption-side real GDP per capita (1000s of 2011 US$)			
B_1	.383** (.154)	.105 (.128)	.079 (.159)
N	323	284	246
R^2	.949	.975	.977
Panel 4C: Spending per capita on non-tertiary education (2011 US$)			
B_1	.005 (.003)	−.001 (.002)	.0002 (.003)
N	290	251	213
R_2	.945	.974	.977

Notes: Robust standard errors in parenthesis. * p<.1, **p<.05, ***p<.01. All regressions weighted by the inverse of the standard deviation of the average test score.

and reading scores. However, in both of these cases, while the coefficients are statistically significant, it could be argued that they are relatively unimportant. For instance, the Panel 4A reading coefficient of .471 indicates that an increase in per-capita real GDP of $1,000 raises the average test score by .47 points—a very small amount relative to the 50-point variance in reading scores between countries.

As a final exploration using the fixed effects approach, I replicate the spline regression approach for each measure of *Expenditure*. The results are produced in **figures 5a and 5b** for the levels of real GDP and in **figure 6** for the per-capita GDP spent on education. Concentrating first on figures 5a–b, like the earlier spline results, countries at the low end of real GDP per capita have a positive relationship between income and PISA results for all three tests and both measures of income. In each graph of figures 5a–b, the slope of the line to the left of the demarcation line (the countries with low incomes) has a statistically significant slope. For instance, the slope of the reading line for low expenditure-side real GDP per capita is .0022, which suggests that PISA reading scores increase by 2.2 points for each $1,000 increase in per-capita income. Like the earlier fixed effects results, this a significantly smaller slope relative to the OLS approach used in figures 2a–c, 3a–c, and 4a–c. The decrease in slope suggests that the fixed effects approach controls for country-specific characteristics that the earlier regressions attributed to the relationship between real GDP per capita and PISA results.

Figure 5a: Effects of per-capita GDP on PISA scores, fixed-effect spline regressions—**expenditure**-side real GDP per capita

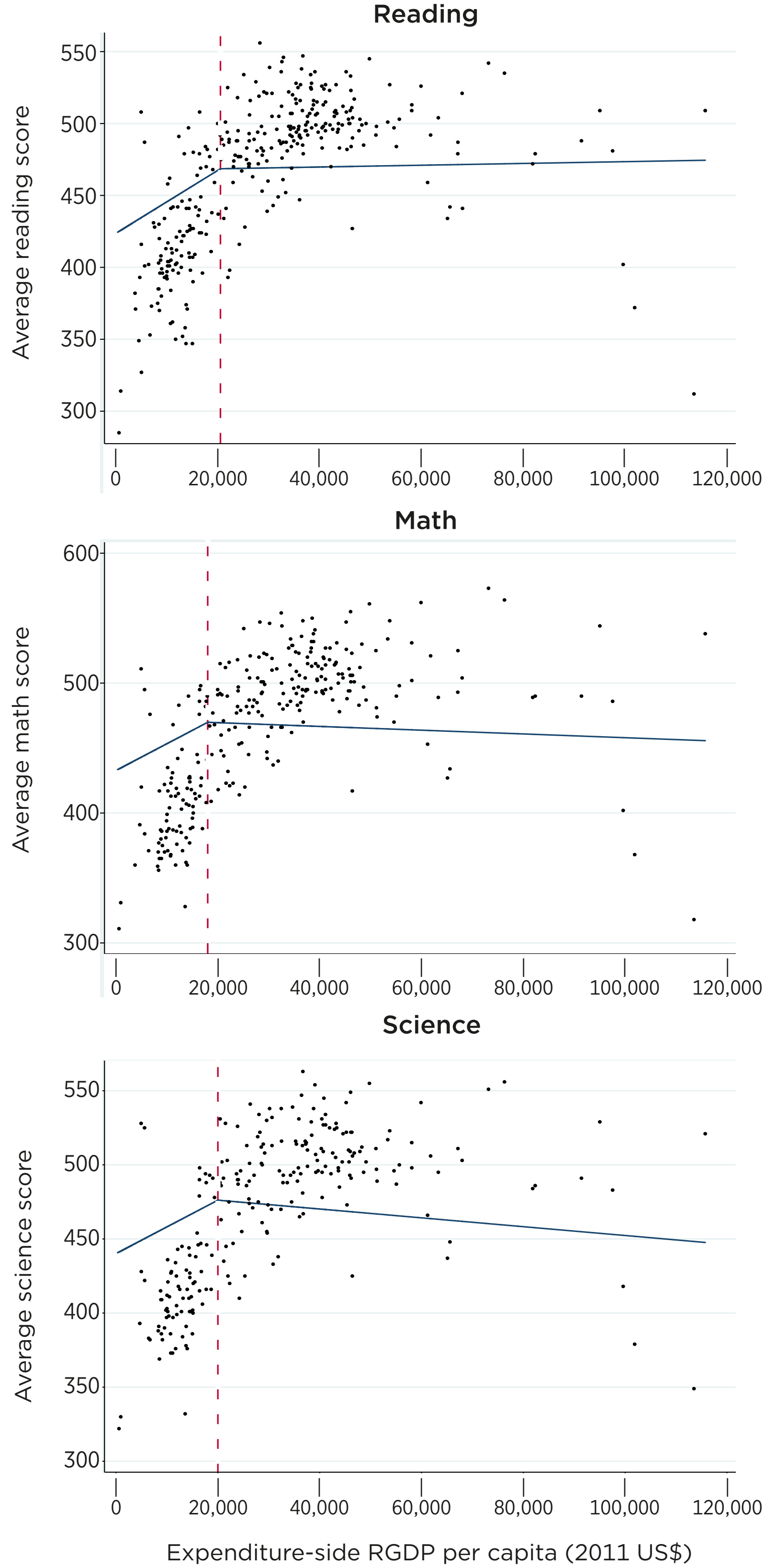
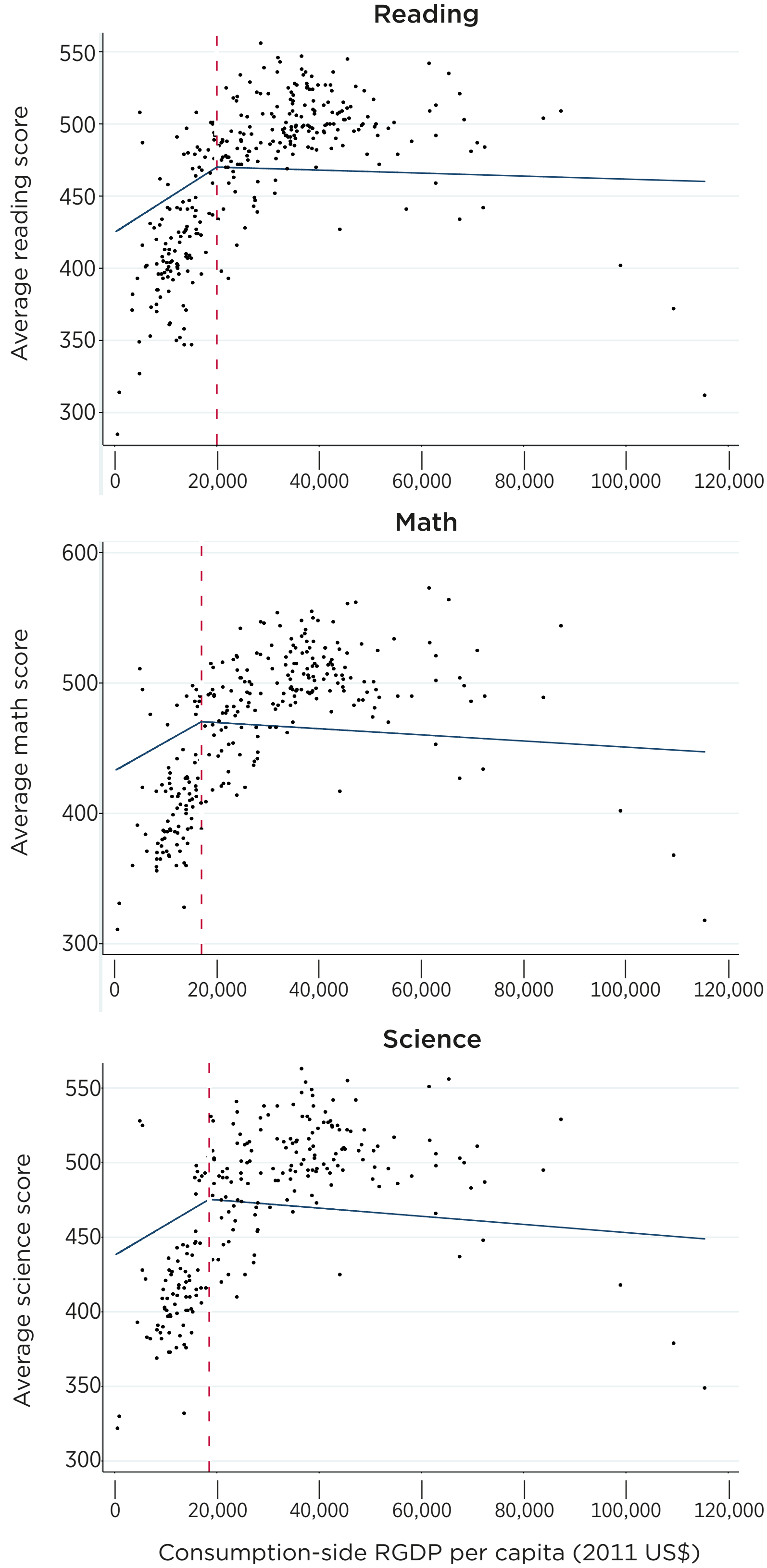


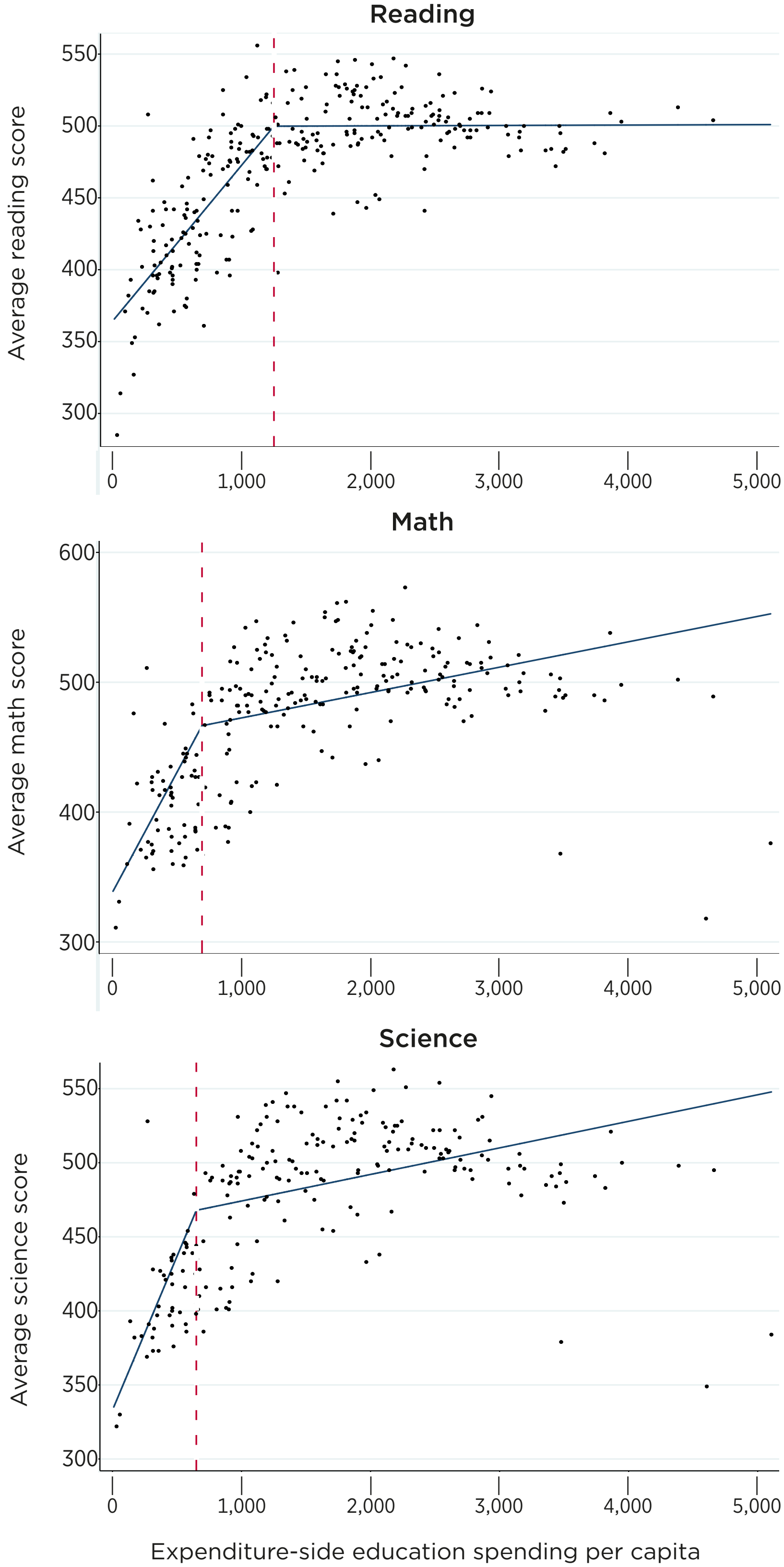
Figure 5b: Effects of per-capita GDP on PISA scores, fixed-effect spline regressions—**consumption**-side real GDP per capita



Also, similar to the earlier spline results, each line to the right of the demarcation line in figures 5a–b has a slope that is not statistically different from zero suggesting that increases in per-capita income for relatively rich countries does not alter PISA scores.

I replicate the spline approach using per-capita spending on non-tertiary education in **figure 6**. Unlike the spline results ignoring the fixed effects, the results of figure 6 suggest that increased spending raises most test scores for both low- and high-spending countries, though high-spending countries appear to benefit less from increased spending than low-income countries. Consider the math results. Low-income countries that increase their spending by $100 per capita are expected to raise their average math PISA scores by 18 points; high-income countries doing the same can be expected to raise their math scores by 2 points. Both results are statistically significant as are those for the science test (slope of .20 for low income and .018 for high income) and for low-income, reading (slope of .11). The only result in figure 6 that has a measured slope statistically not different from zero is for reading in high-income countries.

Figure 6: Effects of fraction of GDP spent on non-tertiary education on PISA scores, fixed-effect spline regressions—expenditure-side education spending per capita



Discussion and Conclusions

This study examines the relationship between PISA test scores and three measures of spending on education: two measures of real GDP per capita and real per-capita spending on non-tertiary education. One clear takeaway from this is that a positive relationship exists between PISA performance and per-capita GDP—as national income increases so does a nation's test scores. However, this positive relationship is driven almost completely by countries with low levels of per-capita GDP. In other words, as poor countries get richer, their PISA results rise. As rich countries get richer, their PISA results remain relatively stable. A more nuanced finding occurs with the level of per-capita spending on education. In the most refined model, a model taking into account country-level fixed effects, there are consistently strong positive relationships between spending on education and test scores for low-spending countries. As spending on education rises, this positive relationship continues at a much lower level as high-spending countries receive smaller increases in test scores for a unit increase in spending on education.

A second finding of this work is the relationship between per-capita spending on education and PISA results. In the preferred fixed-effects results, increases in per-capita spending for low-spending countries raises PISA results. This relationship is strong: a $100 increase in per-capita spending on non-tertiary education is expected to raise a nation's average test scores by around 20 points for math and science and 10 points for reading. In order to put these into context, the boundary between high- and low-spending countries was $700 in math so a $100 increase is equivalent to almost a 15% increase in education spending for these countries. However, a 20-point math increase is also non-negligible; it amounts to about 40% of a country-level standard deviation change in PISA scores.

The impact of increasing test scores on a nation's economy appears substantial. Hanushek and Kimko (2000) find statistically and economically significant impacts of an increase in educational quality on economic growth. Their measure of educational quality is a combination of international test scores, including the PISA. Their estimates suggest that a one country-level standard deviation higher in test performance yields about a one-percentage point higher annual economic growth rate, a fact that has been confirmed by several subsequent studies (e.g., Coulombe and Tremblay, 2006; Jamison, Jamison, and Hanushek, 2007). Taking this as a reasonable estimate, it suggests that a low-spending nation that increases its education spending per capita by $100 can lead to higher economic growth rates by about .4 percentage points. To get a better handle of the magnitude of this, consider Romania which, in 2015, had a per-capita real GDP of 21,000 and spent about $650 per capita on education. If Romania increased its education spending by $100, it would expect to increase its per-capita real GDP by about $84 per year ($21,000 × .004)—roughly equivalent to the amount spent on education.

While increasing education spending for low-spending countries appears to benefit those countries, it is less clear that this is the case for high-spending countries. Among these countries, increases in education spending appear to have much smaller impacts on math and science PISA scores, and negligible impacts on reading scores. The largest measured effect for these countries, that of math, suggests a $100 increase in per-capita education spending raises the average math PISA score by about two points. A $100 per-capita increase is about a five percentage-point increase in education spending in a country like Canada. A two-point increase in math scores represents about 4% of a country-level standard deviation in test scores. Using Hanushek and Kimbo's estimates, this 5% increase in education spending would generate an increase in economic growth by about ½ of ⅒ of a percentage point—a small change given a relatively substantially investment in education.

It is important to ask what drives these results. It is straightforward to consider the benefits of increased income on education for poor countries. Higher incomes allow more students to attend schools where they likely receive better instruction and better resources. Simple income-related changes in health and nutrition, for instance, can improve test scores. The more interesting question occurs at the upper end of the income distribution. Why do countries with high per-capita GDP appear to gain little benefit on PISA scores when income rises? A few suggestions have been forwarded. One possibility is that the education production function flattens out at a given level of resources. If education is produced using inputs like teachers, classrooms, books, and technology, then as nations gain wealth they add these inputs. However, these inputs have diminishing returns; for instance, adding a second teacher to a class probably increases learning by less than adding the first teacher. For instance, even a 5% increase in the quantity of teachers would probably reduce the average class size in most high-spending nations by one to two students—probably not enough to make significant differences in educational outcomes. Indeed, some simple interventions like decreasing class size and measuring its impact on student learning have remained surprisingly elusive.[11]

A second possibility is related to Baumol and Bowen's cost disease (1966). The idea behind this is that, in a growing economy, some sectors will become more productive than others. If the less productive sectors still hire workers, they will do so by paying a salary that competes against the more productive sectors and thus have situations where the wages paid in the less productive sectors are above the marginal productivity of the workers. This may lead to cases where nations raise wages (and spending) in a sector but fails to increase production in that sector. If education productivity has lagged behind other sectors' productivity, then one might expect to find little or no relationship between spending on education and measures of its output among nations with rapidly growing non-education sectors. Finally, in all countries, educational quality is a scarce resource. For developed nations to improve education

11. See Angrist, Lavy, Leder-Luis, and Shany (2017) for a description of the difficulties in measuring the impact of class size on student learning.

quality, it is assumed that schools must compete for teachers who have alternative career prospects, some of which may pay better or offer other amenities. Competing against these can be expensive and lead to marginal improvements in education.

Statistical reasons may also obscure the relationship between PISA scores and national well-being. One clear threat to this type of analysis is that it fails to control for two-way causation. While it seems likely that national income affects education outcomes, it is equally likely that educational outcomes affect national income. If this is the case, the measures of correlation, even those controlling for time-invariant within-country factors like country fixed effects, will not measure causation. This is most easily considered in a thought experiment. Consider the case where a country's per-capita income does not alter education attainment but education attainment leads to country's having greater per-capita income. In this case, we would observe wealthier countries having higher PISA scores not because wealth generates higher attainment, but because higher attainment generates wealth. The fixed-effects approach used within this paper mitigates some of this effect, but not necessarily all of it. In light of that, it is important to recognize that, at best, the results presented here are correlational in nature.

There is a second statistical reason that suggests caution in considering the relationship between Expenditures and PISA results: this has to do with the PISA test itself. The PISA is scored such that the average score of a test-taker in 2000 is 500 with a student-level standard deviation of 100 points. This consistency across time enables researchers to make comparisons among countries and within countries over time—much like what was done in this paper. However, these comparisons are necessarily relative in nature—the PISA mean is constructed to be unchanging over time so the PISA itself cannot provide measures of absolute academic growth. Consider the case where all nations grow in educational ability by 10% from one PISA testing regime to the next. Because the PISA's mean is unchanging over time, researchers would be unable to measure the growth in academic ability. Indeed, if all nations grew by 10% in ability, the country-level PISA results would look unchanged from prior years and a careless researcher would conclude that no academic growth took place.

In conclusion, there is a clear, positive correlation between PISA results and increases in both the income of low-income countries and the levels of educational spending for low-spending countries. There is a weaker relationship between education spending and PISA results for high-spending countries and no detectable relationship between PISA results and increases in per-capita income for high-income countries.

Appendix: List of Countries and Codes

Country	Code
Albania	ALB
Algeria	DZA
Argentina	ARG
Australia	AUS
Austria	AUT
Azerbaijan	AZE
Belgium	BEL
Brazil	BRA
Bulgaria	BGR
Canada	CAN
Chile	CHL
Colombia	COL
Costa Rica	CRI
Croatia	HRV
Cyprus	CYP
Czech Republic	CZE
Denmark	DNK
Dominican Republic	DOM
Dubai-United Arab Emirates	ARE
Estonia	EST
Finland	FIN
France	FRA
Georgia	GEO
Germany	DEU
Greece	GRC
Hong Kong (China)	HKG
Hungary	HUN
Iceland	ISL
Indonesia	IDN
Ireland	IRL
Israel	ISR
Italy	ITA
Japan	JPN
Jordan	JOR
Kazakhstan	KAZ
Korea, Republic of	KOR
Kosovo	XKX
Kyrgyz Republic	KGZ

Country	Code
Latvia	LVA
Lebanon	LBN
Liechtenstein	LIE
Lithuania	LTU
Luxembourg	LUX
Macao (China)	MAC
Macedonia, Republic of	MKD
Malaysia	MYS
Malta	MLT
Mexico	MEX
Moldova, Republic of	MDA
Montenegro, Republic of	MNE
Netherlands	NLD
New Zealand	NZL
Norway	NOR
Panama	PAN
Peru	PER
Poland	POL
Portugal	PRT
Qatar	QAT
Romania	ROU
Russian Federation	RUS
Serbia, Republic of	SRB
Singapore	SGP
Slovak Republic	SVK
Slovenia	SVN
Spain	ESP
Sweden	SWE
Switzerland	CHE
Thailand	THA
Trinidad and Tobago	TTO
Tunisia	TUN
Turkey	TUR
United Kingdom	GBR
United States	USA
Uruguay	URY
Vietnam	VNM

References

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany (2017). *Maimonides Rule Redux*. NBER working paper 23486 (June). National Bureau of Economic Research.

Baumol, William J., and William G. Bowen (1966). *Performing Arts, the Economic Dilemma: A Study of Problems Common to Theater, Opera, Music and Dance*. MIT Press.

Choi, A., J. Calero, and J.O. Excardibul (2012). Private Tutoring and Academic Achievement in Korea: An Approach through PISA-2006. *KEDI Journal of Education Policy* 9, 2: 299–332.

Coulombe, S. and J. Tremblay (2006). Literacy and Growth. *Topics in Macroeconomics* 6, 2: 1–34.

Denny, K., and V. Oppedisano (2013). The Surprising Effect of Larger Class Sizes: Evidence Using Two Identification Strategies. *Labour Economics* 23: 57–65.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer (2015). The Next Generation of the Penn World Table. *American Economic Review* 105, 10: 3150–3182. <www.ggdc.net/pwt>.

Gracia-Perez, J.I., M. Hidalgo-Hidalgo, and J.A. Robles-Zurita (2014). Does Grade Retention Affect Students' Achievement? Some Evidence from Spain. *Applied Economics* 46, 12: 1373–1392.

Groningen Growth and Development Centre (2019). *The Database: Penn World Table version 9.1*. Faculty of Economics and Business, University of Groningen. <https://www.rug.nl/ggdc/productivity/pwt/>.

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

Hanushek, E.A, and L. Woessman (2006). Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries. *Economic Journal* 116: 63–76.

Jamison, E.A, D.T. Jamison, and E. Hanushek (2007). The Effects of Education Quality on Mortality Decline and Income Growth. *Journal of Economic Literature* 46, 3: 722–789.

Jensen, P., and A.W. Rasmussen (2011). The Effect of Immigrant Concentration in Schools on Native and Immigrant Children's Reading and Math Skills. *Economics of Education Review* 30, 6: 1503–1515.

Mazzeo, John, and Matthias von Davier (2013). Linking Scales in International Large-Scale Assessments. In L. Rutkowski, M. von Davier, and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (Chapman and Hall/CRC): 230–260.

National Center for Education Statistics (2017). *PISA IDE—International Data Explorer* (ver. 120, Aug. 15). <http://nces.ed.gov/surveys/pisa/idepisa>.

Organisation for Economic Co-operation and Development [OECD] (2000–2015). Program for International Student Assessment (PISA), 2000, 2003, 2006, 2009, 2012, and 2015 Mathematics, Reading, and Science Assessments.

Organisation for Economic Co-operation and Development [OECD] (2012). *Does Money Buy Strong Performance in PISA?* PISA in Focus #13 (February).

Organisation for Economic Co-operation and Development [OECD] (2018a). *Education Spending*. OECD Data. <https://data.oecd.org/eduresource/education-spending.htm#indicator-chart>.

Organisation for Economic Co-operation and Development [OECD] (2018b) *What is PISA?* <http://www.oecd.org/pisa/>.

Organisation for Economic Co-operation and Development [OECD] (2019). *Education Spending* (indicator). OECD iLibrary. <https://doi.org/10.1787/ca274bac-en>, as of July 16, 2019

Robitzsch, A., and O. Ludtke (2018). Linking Errors in International Large-Scale Assessments: Calculation of Standard Errors for Trend Estimation. *Assessment in Education: Principles, Policy and Practice*. <https://doi.org/10.1080/0969594X.2018.1433633>.

Schneeweis, N., and R. Winter-Ebmer (2008). Peer Effects in Austrian Schools. *Empirical Economics* 32: 387–409.

Vandenberghe, V., and S. Robin (2004). Evaluating the Effectiveness of Private Education across Countries: A Comparison of Methods. *Labour Economics* 11, 4: 487–506.

About the Author

John M. Krieg

John M. Krieg is a Professor of Economics at Western Washington University where he also directs the University's Office of Institutional Effectiveness. Professor Krieg received his Ph.D. from the University of Oregon and undergraduate degrees from Northwestern University. His research interests focus on the economics of education with specific application to teacher labor markets, policy evaluation, and employee unions. Among others, his research has appeared in the *American Education Research Journal*, *Economic Inquiry*, *Education Finance and Policy*, *Economics of Education Review*, *Industrial and Labor Relations Review*, and *Education Evaluation and Policy*. Currently, Professor Krieg serves as a co-director of the Teacher Education Learning Collaborative, which is funded by the Bill and Melinda Gates Foundation.

Acknowledgments

The author wishes to thank the Donner Canadian Foundation for funding this study. Any errors or oversights are the sole responsibility of the author. As the researcher has worked independently, the views and conclusions expressed in this study do not necessarily reflect those of the Board of Directors of the Fraser Institute, the staff, or supporters.

Publishing Information

Distribution

These publications are available from <http://www.fraserinstitute.org> in Portable Document Format (PDF) and can be read with Adobe Acrobat® or Adobe Reader®, versions 7 or later. Adobe Acrobat Reader® DC, the most recent version, is available free of charge from Adobe Systems Inc. at <get.adobe.com/reader/>. Readers having trouble viewing or printing our PDF files using applications from other manufacturers (e.g., Apple's Preview) should use Reader® or Acrobat®.

Ordering publications

To order printed publications from the Fraser Institute, please contact us via e-mail: sales@fraserinstitute.org; telephone: 604.688.0221, ext. 580 or, toll free, 1.800.665.3558, ext. 580; or fax: 604.688.8539.

Media

For media enquiries, please contact our communications department via e-mail: communications@fraserinstitute.org; telephone: 604.714.4582.

Copyright

ISBN

978-0-88975-567-3.

Citation

John M. Krieg (2019). *International Student Assessments: Performance and Spending*. Fraser Institute. <http://www.fraserinstitute.org>.

Supporting the Fraser Institute

To learn how to support the Fraser Institute, please contact us via post: Development Department, Fraser Institute, Fourth Floor, 1770 Burrard Street, Vancouver, British Columbia, V6J 3G7, Canada; telephone: toll-free to 1.800.665.3558, ext. 548; e-mail: development@fraserinstitute.org; or visit our webpage: <www.fraserinstitute.org/support-us/overview.aspx>.

Purpose, Funding, and Independence

The Fraser Institute provides a useful public service. We report objective information about the economic and social effects of current public policies, and we offer evidence-based research and education about policy options that can improve the quality of life.

The Institute is a non-profit organization. Our activities are funded by charitable donations, unrestricted grants, ticket sales, and sponsorships from events, the licensing of products for public distribution, and the sale of publications.

All research is subject to rigorous review by external experts, and is conducted and published separately from the Institute's Board of Directors and its donors.

The opinions expressed by authors are their own, and do not necessarily reflect those of the Institute, its Board of Directors, its donors and supporters, or its staff. This publication in no way implies that the Fraser Institute, its directors, or staff are in favour of, or oppose the passage of, any bill; or that they support or oppose any particular political party or candidate.

As a healthy part of public discussion among fellow citizens who desire to improve the lives of people through better public policy, the Institute welcomes evidence-focused scrutiny of the research we publish, including verification of data sources, replication of analytical methods, and intelligent debate about the practical effects of policy recommendations.

About the Fraser Institute

Our mission is to improve the quality of life for Canadians, their families and future generations by studying, measuring and broadly communicating the effects of government policies, entrepreneurship and choice on their well-being.

Notre mission consiste à améliorer la qualité de vie des Canadiens et des générations à venir en étudiant, en mesurant et en diffusant les effets des politiques gouvernementales, de l'entrepreneuriat et des choix sur leur bien-être.

Peer review—validating the accuracy of our research

The Fraser Institute maintains a rigorous peer review process for its research. New research, major research projects, and substantively modified research conducted by the Fraser Institute are reviewed by experts with a recognized expertise in the topic area being addressed. Whenever possible, external review is a blind process. Updates to previously reviewed research or new editions of previously reviewed research are not reviewed unless the update includes substantive or material changes in the methodology.

The review process is overseen by the directors of the Institute's research departments who are responsible for ensuring all research published by the Institute passes through the appropriate peer review. If a dispute about the recommendations of the reviewers should arise during the Institute's peer review process, the Institute has an Editorial Advisory Board, a panel of scholars from Canada, the United States, and Europe to whom it can turn for help in resolving the dispute.