More countries, similar results

A nonlinear programming approach to normalising test scores needed for growth regressions

MARTIN GUSTAFSSON

Stellenbosch Economic Working Papers: 12/12
JULY 2012

KEYWORDS: HUMAN CAPITAL, CROSS-COUNTRY GROWTH MODEL, TEST SCORE DATA, NONLINEAR PROGRAMMING, EDUCATION POLICY, PISA, SACMEQ, SERCE
JEL: C14, I28, O15

MARTIN GUSTAFSSON
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: MGUSTAFSSON@SUN.AC.ZA

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

More countries, similar results

A nonlinear programming approach to normalising test scores needed for growth regressions[1]

MARTIN GUSTAFSSON

ABSTRACT

Analysts such as Hanushek and Woessman have brought to the fore the deceptiveness of education enrolments, or years of schooling, in growth regressions and the need to consider educational quality. In this paper, a nonlinear programming solution is proposed as a way of normalising to a single scale country average test scores from various international testing programmes. This method, though less transparent and more dependent on certain subjective choices than the existing approach put forward by Hanushek and Woessman, allows for the inclusion of more countries, in particular more African and developing countries, into a growth regression. The regression produces the results one would expect, namely a strong conditional correlation between growth and educational quality. The utility of growth regressions with an educational quality variable for the education policymaker is discussed. A method for arriving at feasible annual improvements in educational quality and hence feasible country targets is presented

Keywords: human capital, cross-country growth model, test score data, nonlinear programming, education policy, PISA, SACMEQ, SERCE
JEL codes: C14, I28, O15

[1] Comments on an earlier draft of this paper from Servaas van der Berg are gratefully acknowledged, as are conversations with Luis Crouch some years ago on how to deal with the data problems discussed in this paper.

1. Introduction

Hanushek and Woessman (for instance 2007, 2009) have used country-level test score averages derived from international testing programmes in cross-country growth models in ways that have thrown new light on the role of the quality of schooling in economic development. This work has been highly influential in education policymaking circles and has been reiterated in a number of key reports directed at policymakers, including a 2010 OECD report[2] and the 2010 McKinsey education report[3], both underlining the widespread danger of forfeited economic growth resulting from education decision-making that is insufficiently focussed on what children learn.

Within economics, Hanushek and Woessman's work has improved our understanding of human capital theory. Whilst human capital as measured by years of schooling has featured as a strong predictor of economic growth in cross-country analysis since at least Mankiw, Romer and Weil (1992), the inclusion of test scores in this type of analysis has provided a stark illustration of how deceptive years of schooling can be as a measure of human capital. This realisation has contributed towards a shift in policymaking, from years of schooling alone, to years of schooling *and* what children actually learn. One manifestation of this shift has been the proliferation, in particular within developing countries in recent years, of standardised and often sample-based assessment systems to improve the monitoring of educational quality (UNESCO, 2006: 48).

Whilst the imperative of improving educational quality, in particular within developing countries, is an uncontroversial one amongst education policymakers, there are associated criticisms that warrant acknowledgement. Putting economic growth at the centre of national development can be controversial (Stiglitz, Sen and Fitoussi, 2009). Moreover, as Sen (1999: 5) has argued, educational improvement does not need to be linked to growth in order to emerge as a development priority. Better education comes with welfare improvements that are not necessarily linked to growth. Clearly, cross-country growth models that consider education do not offer policymakers the full picture. Yet they do describe an important part of it, in particular as far as policymakers in developing countries are concerned.

Hanushek and Woessman (2009) present an approach for transforming country scores from disparate tests to a single scale using the national assessment system of the United States as an anchor and thus making the United States the bridge country that joins the various tests. Whilst this approach has many strengths, a drawback is that it does not permit the use of country scores from testing programmes where the United States is not a participant. This limits the number of countries, in particular developing countries, that can be included in a cross-country economic growth analysis. A further limitation is that policymakers from the excluded countries are not able to see how well their countries perform relative to a range of countries from several programmes. Dealing with these limitations is one key concern of this paper.

A second key concern is to provide guidance with respect to the feasible magnitude of annual test score improvements, a matter of increasing concern for policymakers faced with the need to set realistic targets.

2. An alternative nonlinear programming approach

A specific limitation brought about by the requirement in Hanushek and Woessman's (2009) normalisation approach (see Appendix 2) that the United States be a participant in every test

[2] OECD, 2010a.
[3] Mourshed, Chijioke and Barber, 2010.

is that the regional programmes SACMEQ[4] in Africa and SERCE[5] in Latin America cannot be considered. These two regional programmes include 19 developing countries which do not participate in programmes where the US participates. These programmes are similar in many ways to the programmes considered by Hanushek and Woessman (2009): they focus on mathematics and reading and they use a scale where the mean across participating countries is 500 and the standard deviation across all countries at the pupil level is 100. Both programmes are limited to the primary level, however.

An approach to obtaining normalised country scores in a manner that permits the inclusion of programmes such as SACMEQ and SERCE is the focus of this section. As in the existing approach of Hanushek and Woessman (2009), in the new approach the relationship between the original score O and the transformed score T is considered linear, implying the solution must involve finding the intercept (α) and the slope (β) within each relationship. The assumption of a linear relationship, both in the existing and new approaches, implies an underlying assumption, namely that the shape of the distribution of country scores with respect to the same set of countries in different tests should be similar.

The viability of the approach proposed here depends on a requirement being fulfilled with respect to the pattern of bridge countries spanning programmes in the data used. The following table describes the data. Country scores from 20 tests, each represented by a year in Table 1, were used. All tests are from the period 2000 to 2009. In total 742 country scores from 113 countries were available. In the discussion that follows 'programme' means a group of tests within which scores are comparable. PISA[6] is one programme because in the period in question scores within a subject are comparable over time by design (OECD, 2009: 157 and OECD, 2010b: 21, 186). Moreover, PISA mathematics and reading scores are comparable insofar they follow the same scale of a mean of 500 and standard deviation of 100, anchored in a specific year, and involve the same countries in any one year. Similarly, PIRLS[7], SACMEQ and SERCE are each counted as a programme. TIMSS[8] Grade 4 and TIMSS Grade 8 are counted as two separate programmes, however, as different sets of countries participate. Around half of the TIMSS participants in 2003 and 2007 opted for testing in just one of the two grades[9].

[4] Southern and Eastern African Consortium for Monitoring Educational Quality.
[5] Second Regional Comparative Study (from Spanish).
[6] Programme for International Student Assessment.
[7] Progress in International Reading Literacy Study.
[8] Trends in International Mathematics and Science Study.
[9] The process of ensuring that TIMSS scores are comparable across years is explained in Mullis *et al* (2008: 402).

Table 1: Summary of the test data

Test series	*Years*	*Coun-tries*	*2-point series*	*3-point series*	*4-point series*	*Total series*
PISA mathematics (age 15)	2000, 2003, 2006, 2009	72	14	10	35	59
PISA reading (age 15)	2000, 2003, 2006, 2009	72	14	11	34	59
PIRLS reading (Grade 4)	2001, 2006	47	28			28
TIMSS mathematics (Gr 4)	2003, 2007	41	22			22
TIMSS mathematics (Gr 8)	2003, 2007	61	36			36
SACMEQ mathematics (Gr 6)	2000, 2007	14	13			13
SACMEQ reading (Gr 6)	2000, 2007	14	13			13
SERCE mathematics (Gr 6)	2006	16				0
SERCE reading (Gr 6)	2006	16				0
Overall		113	140	21	69	230

Sources: PISA microdata from http://www.pisa.oecd.org, OECD, 2007, OECD, 2010b and Walker, 2011 (for PISA); Mullis et al, 2007 (for PIRLS); Mullis et al, 2004 and Mullis et al, 2008 (for TIMSS); Makuwa, 2010 (for SACMEQ); UNESCO, 2008 (for SERCE).
Note: A 2-point series would be a series of a specific country with two data points for two different years. There are no overlapping series in the count of series, meaning for instance that no 3-point series would also be counted as two 2-point series.

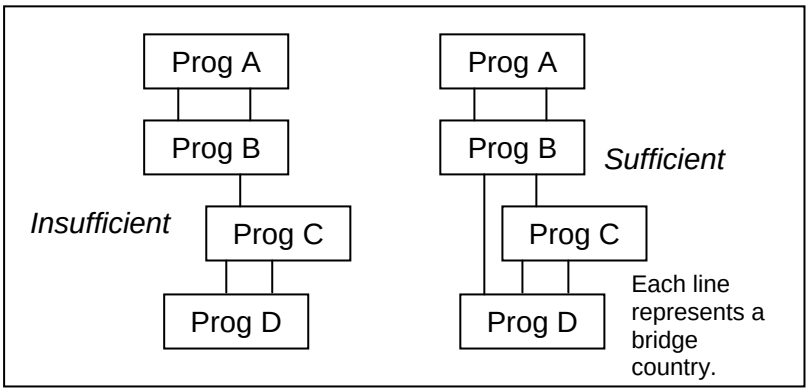
Table 2 indicates the number of bridge countries that join the six programmes to each other. For instance, two bridge countries join SACMEQ to TIMSS Grade 8, namely South Africa and Botswana. South Africa also joins SACMEQ and PIRLS. Compared to SACMEQ, SERCE has a greater number of bridge countries joining this programme to the large international programmes.

Table 2: Programmes and number of bridge countries

	PIRLS	PISA	SACMEQ	SERCE	TIMSS Gr 4	TIMSS Gr 8
PIRLS						
PISA	39					
SACMEQ	1	1				
SERCE	2	9	0			
TIMSS Gr 4	31	29	0	2		
TIMSS Gr 8	35	39	2	3	36	

The only requirement for the approach described below is that it should not be possible to divide the programmes into two sub-sets separated by just one bridge country. Any two sub-sets of programmes must always be joined by at least two bridge countries. If this condition is not met, the nonlinear programming solution explained below will not work as there would essentially be two separate optimisation processes instead of one. Figure 1 illustrates the requirement. Table 2 provides sufficient information to indicate that the requirement is met with respect to the data described in Table 1. Even if SACMEQ and TIMSS Grade 8 had not been joined by any bridge countries, the requirement would still have been met, because South Africa joins SACMEQ and PIRLS and Mauritius joins SACMEQ and PISA.

Figure 1: Programmes and bridge countries



In the approach presented here, one programme must be selected to provide the standard scale for the transformed country scores. For this, PISA was selected. For each country and programme, the mean of the existing country scores was calculated. For instance, one SACMEQ mean value for South Africa was calculated from the four original country scores for the two years and two subjects.

Transforming the country- and programme-specific values to comparable country scores involved solving a nonlinear programming problem whose objective function was the following:

$$\min z = f\left(\alpha_m, \beta_m, ..., \alpha_{m=n-1}, \beta_{m=n-1}\right) \tag{1}$$

The value z, which must be minimised, is a function of a number of pairs of intercepts, α, and slope coefficients, β, which together constitute the decision variables of the problem. The number of pairs is equal to the total number of programmes minus one, as one programme has been selected as the standard. There are n programmes in total, including the standard programme (n equals 6 in our case).

Function f can be described as follows:

$$z = \sum_{j=1}^{k} \left(\overline{D}_j W_j\right) \tag{2}$$

$$D_{ij} = \left(T_{i,m=c} - T_{i,m=d}\right)^2 \tag{3}$$

$$T_{im} = \alpha_m + \beta_m O_{im} \tag{4}$$

Value z is the sum of k values, k being the available bridges in the sense of non-zero values in Table 2. The variable k equals 13 in our case. D is a value attached to each country within a bridge. Equation (2) involves finding the mean for D across countries within each bridge (for instance the mean across 39 values within the PISA-PIRLS bridge) and multiplying this mean by a weight W (explained below). In equation (3), D for country i and bridge j is obtained by finding the squared difference between transformed country scores T for the two programmes c and d, the programmes joined by bridge j. In equation (4) the transformed country score for country i and programme m is a linear function of the original score for the country in the programme (as explained earlier, this original score can be the mean across several tests if more than one test exists).

The weight W attached to each of the k bridges is calculated in such a way that a weight of 1.0 for each country is distributed across those bridges that join the programmes in which the country participates. The sum of all the k values of W is thus equal to the number of countries, 113 in our case. The weighting system therefore follows the principle generally followed in cross-country analysis, which is to assign an equal weight to every country. Without the weighting system, in other words if every bridge carried an equal weight, countries which are found repeatedly across many programmes would implicitly tend to carry too much importance in the nonlinear programming solution. A bridge is weighted more if (1) there are more countries in the programmes being bridged, (2) the countries in the programmes being bridged have a lower presence in other programmes and (3) the programmes being bridged have fewer bridges serving them. The following two equations describe how W is calculated.

$$F_m = \sum_{i=1}^{e} \frac{1}{P_i} \tag{5}$$

$$W_j = \left(\frac{F}{B}\right)_{m=c} + \left(\frac{F}{B}\right)_{m=d} \tag{6}$$

P_i is the number of programmes that country i participates in. The value F for programme m is the sum of all the reciprocals of P for the e countries participating in programme m. Weight W for bridge j is the sum of two values, one for each of the two programmes being bridged. The programme-specific value is F over B, where B is the total number of bridges that serve programme c (or d).

Returning to equation (1), what the nonlinear programming solution does is adjust the coefficients of the five linear relationships between T and O in a way that minimises the differences between the values T for the same country across different programmes. Because in effect $T=O$ in the case of PISA (or $\alpha=0$ and $\beta=1$), all country scores T gravitate towards the scale used by PISA. If the differences between the country scores of different programmes were only a matter of different programme-specific means and standard deviations with respect to the scoring process, then the difference in equation (3) would in all instances be zero and z in equation (1) would be reduced to zero.

The nonlinear programming problem is subject to just one constraint, which is that the transformed country scores T cannot be negative.

The results of the application of the new approach, in terms of programme-specific values T and the overall mean for each country, are provided in the Appendix 1[10]. The 113 overall mean values can be considered comparable country scores.

In exploring alternatives, a quadratic as opposed to a linear form was attempted for equation (4). This did not appear to come with any noticeable benefits, for instance in terms of the performance of the country scores in the growth models described in the next section. The use of the weight W did not make a large difference to the results. The largest difference was in the case of Zambia, where the transformed country score would have been 265 instead of the 261 reported in Appendix 1 had unweighted bridges been used. Importantly, though, the goodness of fit, in terms of R^2, of the regression models discussed below was marginally better with country scores produced using weighted bridges, compared to unweighted bridges.

What are key differences between the nonlinear programming approach described above and the approach used by Hanushek and Woessman (2009)? Two key differences stand out, both of which point to weaknesses in the nonlinear programming approach, or the cost of having

[10] The Excel file used for the calculations can be obtained from the author.

more countries in one's set of normalised country scores. One is that the approach presented here is applicable at a high level in the sense that it subsumes details relating to year, subject and institutional level within the relatively untransparent decision variables of equation (1). The approach is thus not amenable to an analysis of trends over time in the way the Hanushek and Woessman approach (2009) is. A second key difference is the dependence of the approach presented here on a weighting system, the design of which is at least partly a matter of subjective judgement.

3. The performance of the new normalised scores in a growth model

When Hanushek and Woessman's (2009: Table 1, Model 2) most basic growth regression involving educational quality and initial income as explanatory variables is reproduced using the overall country scores appearing in Appendix 1 below, the regression results are similar. The adjusted R^2 value is around 0.65, using the same 46 countries that were available for both models. However, the inclusion of more countries, something permitted by the new set of country scores, reduces R^2 to 0.56. It was possible to raise the number of countries from the 50 of the Hanushek and Woessman (2009) analysis to 66 (this excludes China, whose new score refers only to Shanghai). Including more of the 113 countries from the appendix was not possible due to data limitations relating to non-education variables, in particular initial GDP per capita in 1960.

To obtain a goodness of fit comparable to that of the Hanushek and Woessman (2009) models for the dataset of 66 countries requires standard transformations of two of the explanatory variables, specifically the natural log of initial GDP per capita (I) as well as the natural log of the gross enrolment ratio at the secondary level (E). As an indicator of enrolment, school enrolment at the secondary level over the period 1980 to 2005 was chosen here and not the overall average years of schooling used by Hanushek and Woessman, partly because secondary enrolment displays a higher correlation with growth and partly because average years of schooling is affected by tertiary level enrolments whilst the intention is to focus on the effect of basic education. A secondary school enrolment variable was also the preferred quantity of education variable in the growth analysis of Mankiw, Romer and Weil (1992: 419). Details relating to the data used here can be found in Table 5.

To attain a reasonable goodness of fit it was moreover necessary to exclude one outlier, Kenya, from the analysis. To identify which outliers to drop, the DFITS statistic was used (Baum, 2006: 128). Kenya had a country score which predicted much higher GDP growth than what was actually experienced.

Table 3: Growth models with new quality of education values

Variable	*1*	*2*	*3*
Constant	4.21***	3.28***	3.87***
	(3.60)	(3.56)	(4.65)
T [Learning outcomes]		1.88***	1.54***
		(9.18)	(7.74)
ln(*E*) [Enrolment]	1.53***		0.87***
	(5.66)		(4.11)
ln(*I*) [Initial income]	-0.96***	-1.08***	-1.40***
	(-4.59)	(-6.90)	(-8.78)
N	63	63	63
R^2	0.350	0.585	0.678
Adjusted R^2	0.329	0.572	0.661

*Note: Dependent variable is average annual GDP per capita growth 1960-2004. *** indicates that the estimate is significant at the 1% level of significance. Values in brackets are t statistics. For this analysis and and that in Table 4 the new country scores are divided by 100 to provide slope coefficients comparable to those of Hanushek and Woessman (2009).*

The Table 3 results permit similar conclusions about education and growth to those drawn by Hanushek and Woessman (2009), though there is a noteworthy difference. Models 1 and 2 confirm how much more valuable educational quality, as opposed to education enrolment, is if we want to explain why certain countries experience higher economic growth than others. In Model 3 both educational quality and enrolment are highly statistically significant, whilst in the comparable Hanushek and Woessman (2009: Table 1, Model 3) model only educational quality was found to be significant. It is likely that this is due to the use of a different measure of enrolment, one that excludes the post-school level, as well as the inclusion of more developing countries, or countries with greater variation with respect to secondary level enrolments. The coefficients associated with initial income remain negative and statistically significant across all models, something that Hanushek and Woessman (2009: 9) remind us points towards conditional convergence between countries.

In a model with a wider range of explanatory variables, roughly following Hanushek and Woessman (2009: Table 1, Model 8), educational quality as reflected by the new country scores remains statistically significant. The adjusted R^2 value in Table 4 of 0.79, in a model that includes 62 countries, compares to an adjusted R^2 value of 0.80 in the corresponding Hanushek and Woessman model with 45 countries. The most significant regional dummies were those of Sub-Saharan Africa (SSA) and Middle East and North Africa (MENA), both of which produced strongly negative coefficients. Below, the two are combined into one dummy variable. Following Burger and Du Plessis (2011), it is likely that this dummy variable masks omitted variables relating to, for instance, differences in the dependency ratio.

Table 4: Model with a large range of explanatory variables

Variable	
Constant	9.92*** (6.53)
T [Learning outcomes]	0.70*** (3.02)
ln(*E*) [Enrolment]	0.85*** (4.65)
Openness	0.01*** (3.31)
Rule of law	0.43** (2.34)
Total fertility rate	-0.13 (-1.49)
SSA/MENA dummy	-0.53** (-2.15)
ln(*I*) [Income]	-1.62*** (-10.36)
N	62
R^2	0.794
Adjusted R^2	0.768

*Note: Dependent variable is average annual growth 1960-2004. *** indicates that the estimate is significant at the 1% level of significance, ** at the 5% level. Values in brackets are t statistics.*

An obvious question is the degree to which endogeneity in the form of simultaneity exists in the growth regressions presented above. Does better economic growth somehow lead to better educational outcomes? This matter is not analysed here, though Hanushek and Woessman (2009) offer a convincing analysis, using instrumental variables relating to education policy choices, that points to the main direction of causality indeed being one from better educational outcomes to better economic growth.

Table 5: Data sources for growth models

Variable	*Source*	*N*	*Mean*	*Min.*	*Max.*
Real annual growth in GDP per capita (in PPP USD of 2006), average 1960-2004	Heston, Summers and Aten, 2006.	107	2.49	-2.00	8.36
GDP per capita (in real PPP USD of 2006), 1960	As above.	67	4753	412	15253
Gross enrolment ratio in secondary schooling, average 1980-2005	UNESCO: UIS, 2009.	101	67	4	118
Average country score	See section 3 above.	113	427	260	578
Economy openness in 2000, following methodology of Barro and Sala-i-Martin (2003: 529).	Heston, Summers and Aten, 2006; UN Statistics Division, 2009.	110	5.7	-68.3	244.9
Rule of law indicator for 2000	World Bank, 2010.	107	0.30	-1.25	1.92
Total fertility rate, 1960	UN Statistics Division, 2009.	108	4.88	1.82	8.40

4. Realistic improvement targets

Normalised educational quality values at the country level have a variety of practical uses for education policymakers. Insofar as they allow for the magnitude of the impact of educational quality improvements on economic growth to be known, they provide policymakers with a powerful argument for a better focus on learning outcomes by administrators, schools, parents and teacher unions. The OECD (2010a) has produced a booklet on the matter. Growth models can in fact be used to demonstrate to teachers how improvements in their earnings in the long run are in fact in their own hands. If teaching improves, economic growth and tax revenue improve and hence the possibility of paying teachers more. Normalised country scores can also assist policymakers in deciding what annual improvements in national average test scores to aim for on the basis of what has occurred across a wide range of countries at a similar level of performance. This is the focus of this section.

Setting system-wide test score improvement targets has become common practice amongst governments wishing to improve educational performance. In the United States, as part of the No Child Left Behind policy, every state must achieve certain targets. Hawaii, for example, produced a set of targets for the percentage of pupils attaining proficiency in mathematics, from a baseline of 10% in 2001 to 100% in 2014[11]. In other words, Hawaii's targets focussed on a percentage of pupils, not an overall average score. Brazil serves as an example of a country that has set targets based on average scores. In Brazil such targets at the federal, state, municipality and school level have been set using a central database of baseline scores and a formula that assigns higher targets where the baseline scores are higher[12]. Brazil's targets are based more or less but not exactly on average scores. The targets are in fact index values that include the average score and a grade promotion ratio. Combining these two variables within the index is intended to minimise the risk that schools will game the system by forcing weaker pupils to repeat their grades and hence delay (or even prevent, through dropping out) their participation in the test[13]. South Africa aims to increase the percentage of pupils performing at a basically acceptable level in languages and mathematics from a range of between 20% and 50% (depending on grade and subject) in 2010 to 60% in 2014, with monitoring occurring through the country's new Annual National Assessments programme[14].

System-wide test score targets, like many other educational targets, are susceptible to what Inbar (1996) refers to as ritualistic planning, or targets that are more a reflection of current political and popular aspirations than of evidence-driven analysis of what is possible. Guidance to policymakers with respect to test score targets is scarce and sometimes difficult to interpret. Hanushek and Woessman (2007: 44) provide a broad indication when they say that in a programme such as TIMSS, an improvement of half a standard deviation (meaning 50 points in TIMSS) is possible within a period of 10 to 30 years. From a base of 300 points, about the average score for Qatar, the worst performing country in TIMSS at the Grade 8 level in 2007, this represents a range in annual growth of 0.6% to 1.6%. This is not incompatible with the patterns identified below. Some guidance is also provided in the 2010 McKinsey report (Mourshed, Chijioke and Barber, 2010: 16), which points to best historical trends of around 0.2% per annum in 'PISA 2000 units', assuming a baseline of 500. This is considerably lower than the desired level of improvement put forward in Hanushek and Woessman (2007: 44), even if one takes into account the different points of departure and the different scales (PISA against TIMSS).

Before examining the best historical trends, which provide guidance as to how optimistic one can be about future improvements, it is instructive to examine the average trends in the recent past. Table 1 indicates that the data used for this analysis include 230 country- and test-specific time series. Just 90 of these series consist of more than two points in time – 21 have three points and 69 have four points. The mean annual improvement across all 230 series, with least squared trendlines used for the 3- and 4-point series, is just 0.23%.[15] If values are converted to the PISA scale using the coefficients described in a previous section, the figure is

[11] Accountability Resource Centre Hawaii, 2003.
[12] See http://sistemasideb.inep.gov.br/resultado/.
[13] Fernandes, 2007.
[14] South Africa, 2010.
[15] The use of percentages in changes in scores which have been transformed, as is the case in most international educational evaluations, is problematic. If the transformation was to a mean of 500 and a standard deviation of 100, a ten percent improvement from the average would be half a standard deviation. Similarly, if the score had been normalized to a mean of 1000 and a standard deviation of 100, such an increase would have been a full standard deviation. For this reason one should interpret percentage growth in the subsequent analysis cautiously as simply indicative; most international scores have been converted to broadly similar levels, so a 1% improvement is interpreted in roughly the same way whether it is applied to PISA scores or SACMEQ scores, for instance. Ideally it would be more correct to work with increases in points on a scale where all scores have already been transformed as in this paper.

similar, at 0.22%. The last figure becomes -0.03% if only OECD countries are considered and 0.34% for non-OECD countries, suggesting that there is educational quality convergence between developing and developed countries. However, if the diminishing marginal rate of improvement is taken into account a simulation would indicate that it would take exceedingly long for developing countries to catch up. South Africa, for instance, would require 160 years to reach the levels found in non-Asian OECD countries. For Brazil the duration would be around 130 years. Clearly the mean rates of improvement seen in the data are discouragingly low.

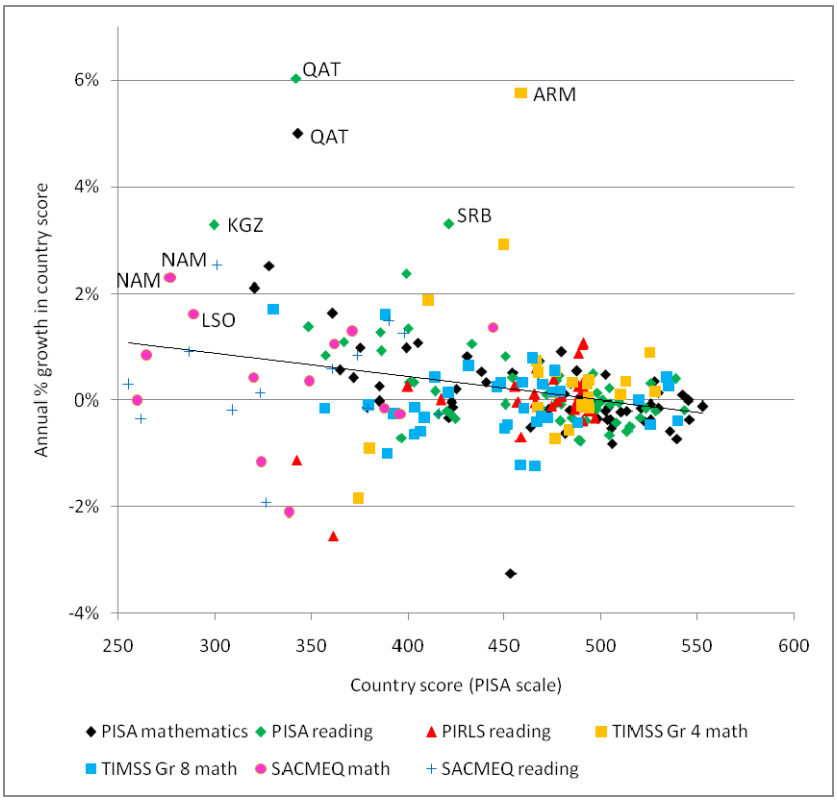
An analysis of the 90 series with more than two points suggests that there is considerable measurement error in the country scores. Of the 90 series, 15 display movements that are consistently upward, 15 display movements that are consistently downward and 60 display inconsistent trends consisting of both upward and downward movements. One might expect that of the 60, many would display small inconsistencies reflecting a situation of virtually no change. To test this, a margin of tolerance of 0.05 of the standard deviation of country scores per programme was used to determine when there was no change. If the change between two data points was within the margin of tolerance, the trend was considered a flat one with zero change. Using this margin still resulted in 47 series that displayed inconsistent trends moving both up and down. It seems unlikely that the quality of human capital would in reality display trends as erratic as what these figures suggest and reasonable to assume that the patterns seen in the data indicate measurement error. This measurement error should caution us above all against reading too much into the two-point series of individual countries. In fact, although these two-point series are used in the analysis that follows, the analysis was also done without these series. The exclusion of the two-point series did not make a difference that would substantially change the conclusions presented below.

The 15 consistently upward trends referred to above are found in 11 countries, none of which had any series, from the set of 90 series, with a consistently downward trend. Examining the education policies of these 11 countries should be of particular interest. The 11 countries, in descending order of size of their annual improvement, are: Brazil, Chile, Indonesia, Tunisia, Turkey, Poland, Germany, Israel, Uruguay, Hungary and Switzerland. The countries Chile, Tunisia, Turkey and Germany each had two series of more than two points with a consistently upward trend (for these four countries the average improvement across the two trends was used for the ranking provided here).

The following graph displays the annual percentage increases of each of the 230 series referred to in Table 1, using the PISA scale for the horizontal axis. What immediately stands out is the greater degree of vertical dispersion amongst countries with lower country scores. This pattern remains if two-point series are removed. Part of the explanation for the greater vertical dispersion on the left-hand side of graph is that in the case of the 78 countries with more than one series (for instance because they have different series for different subjects in the same programme), within a single country trends are more likely to vary from series to series if the country performs poorly. Specifically, the within-country gap between the worst improvement and the best improvement amongst the 78 countries in question was on average 1.1 percentage points if the lowest half of performers were considered and 0.5 percentage points if the top half of performers were considered. Possible explanations are a greater degree of measurement error amongst developing countries and that developing countries, though they may on average display greater annual improvements, improve inconsistently and in spurts, perhaps due to policy or leadership instability.

The trendline in the graph confirms the pattern, albeit weak (R^2 is only 0.08), of greater increases amongst developing countries in percentage terms (as will be seen below, this pattern also exists with respect to absolute increases).

Figure 2: Annual country score increases by level



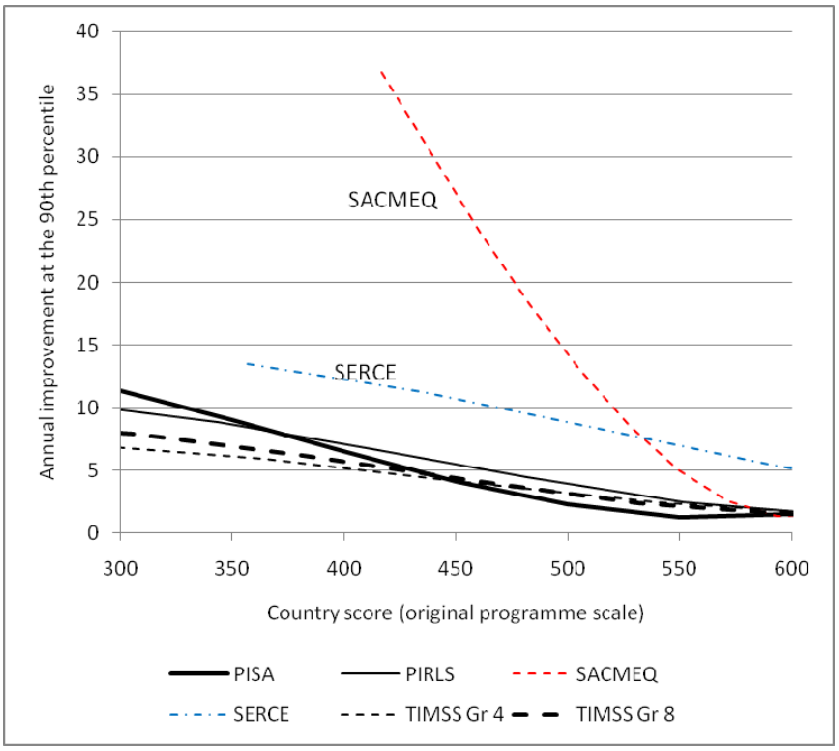
Sources: See Table 1.
Note: In order to position each point, which represents a time series, along the horizontal axis, original country scores had to be converted to the PISA scale. This was done by using the five sets of coefficients obtained through the equation (1) minimisation process. For each series, the mean of the converted country score values was used to plot the point on the horizontal.

Figure 2 is able to provide policymakers with a rough idea of the maximum possible improvements that should inform system-wide targets. Clearly annual improvements of as high as 5% appear to be possible, though rare. At the same time it is important to acknowledge that of the 11 consistent improvers referred to earlier, the best average annual improvements within a single series were only 1.6%, in the case of Brazil, followed by 1.3%, in the case of Tunisia.

Stochastic frontier analysis is conceivably an appropriate technique for identifying a systematic efficiency frontier, or a feasible outer limit for expected improvements, in the data. However, the dataset being considered here is too limited in terms of the number of efficient units, or countries, for this technique. Quantile regression analysis can assist, however, in identifying a performance frontier with a margin of safety to deal with the possibility of outliers. The next graph displays the annual improvement in programme-specific scores as a quadratic function of the country score at the 90th percentile (pseudo R^2 was 0.23). The regression analysis was performed across all 230 series using scores on the PISA scale. Thereafter scores were converted back to their original scales as these are what policymakers would be familiar with. South Africa, with a SACMEQ country score in mathematics of 486 in 2007, might aim for an annual improvement of around 17 SACMEQ points according to Figure 3. Brazil, with a PISA mathematics score of 386 in 2009, might aim for an annual improvement of around 7 PISA points (not much higher than its actual annual rate of improvement in PISA mathematics of 1.6% referred to above). The SACMEQ curve is situated well above the PISA curve in Figure 4 because, for instance, a score of 450 SACMEQ points has a lower cognitive value and represents a lower transformed score than

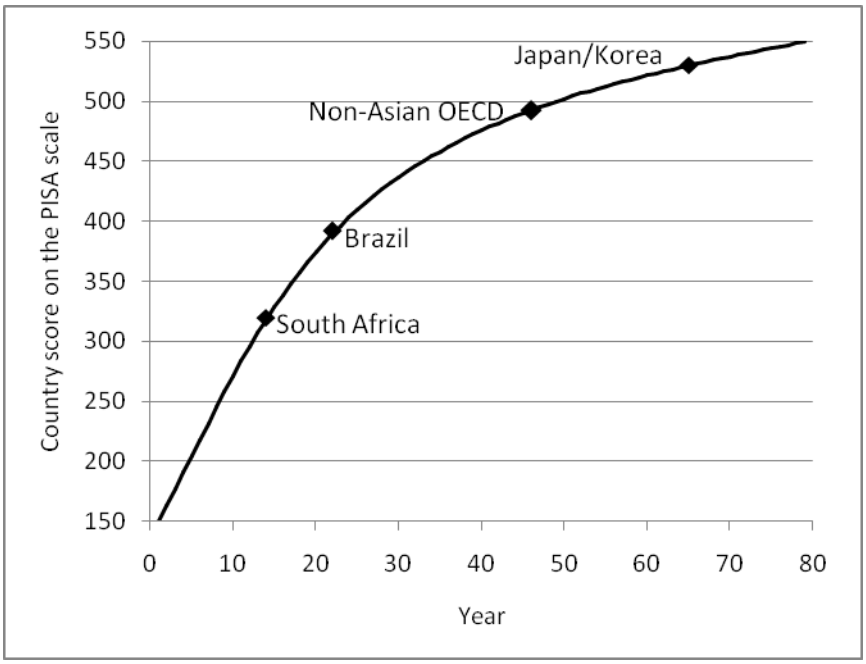
450 PISA points and expected improvements are greater at lower values. It is important to emphasise that although the curves in Figure 4 are programme-specific, the shape and position of each curve is influenced by patterns seen across all programmes.

Figure 3: Programme-specific feasible improvement targets (absolutes)



The PISA curve from the previous graph produces the minimum years to improve curve of the next graph. For year 1, a low level of performance of 150 PISA points was chosen and in subsequent years improvements occur in line with the PISA curve from Figure 3. The four points in the graph reflect actual recent levels of performance, in terms of PISA points. If optimal education policies are implemented, South Africa might attain the level of performance of non-Asian OECD countries (average of 491 on the PISA scale) after 34 years (52 minus 18 on the horizontal axis). It would take Brazil about 25 years to achieve the same. These are clearly optimistic scenarios, but not impossible ones.

Figure 4: Minimum years needed to improve



5. Conclusion

A nonlinear programming approach to transforming the average scores countries achieve in various international testing programmes so that scores become comparable across programmes has been explained. This approach is similar to the existing Hanushek and Woessman (2009) approach, for instance insofar as it assumes a linear relationship between original and transformed country scores. However, its benefits, being above all its ability to bring together a greater variety of testing programmes, comes at a cost, namely decreased transparency and intuitiveness, as well as a degree of subjectivity in the determination of a weighting system. Yet the results of the new approach (provided in Appendix 1) perform as one would expect in a cross-country growth model.

Analysis of the transformed country scores was used to establish what countries at different levels of educational development might regard as feasible improvement targets. That targets and improvement strategies are necessary is illustrated by the fact that present trends imply that on average developing countries would require over a century to reach current developed country levels. The trends seen amongst the best improvers, however, point to the possibility of a much shorter trajectory of two or three decades.

References

Accountability Resource Centre Hawaii (2003). *Reading and mathematics AYP starting points, intermediate goals, annual measureable objectives.* Honolulu. Available from: <http://arch.k12.hi.us > [Accessed February 2011].

Barro, R.J. & Sala-i-Martin, X. (2003). *Economic growth*. Cambridge: MIT Press.

Baum, C.F. (2006). *An introduction to modern econometrics using Stata*. College Station: Stata.

Burger, R. & Du Plessis, S. (2011). Examining the robustness of competing explanations of slow growth in African countries. *Studies in Economics and Econometrics*, 35(3): 21-48.

Fernandes, R. (2007). *Índice de Desenvolvimento da Educação Básica (Ideb)*. Brasília: Ministério da Educação. Available from: <http://www.publicacoes.inep.gov.br> [Accessed November 2009].

Hanushek, E.A. & Woessman, L. (2007). *The role of school improvement in economic development.* Washington: National Bureau of Economic Research. Available from: <http://papers.nber.org> [Accessed June 2007].

Hanushek, E.A. & Woessman, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation*. Washington: National Bureau of Economic Research. Available from: <http://papers.nber.org> [Accessed March 2009].

Heston, A., Summers, R. & Aten, B. (2006). Penn World Table version 6.2 (dataset). Philadelphia: Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania. Available from: <http://pwt.econ.upenn.edu> [Accessed June 2007].

Inbar, D.E. (1996). *Planning for innovation in education*. Paris: IIEP. Available from: <http://www.unesco.org> [Accessed December 2006].

Makuwa, D.K. (2010). Mixed results in achievement. *IIEP Newsletter*, XXVIII(3). Available from: <http://www.iiep.unesco.org> [Accessed October 2010].

Mankiw, N.G., Romer, D. & Weil, D.N. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, 107(2): 407-443.

Mourshed, M., Chijioke, C. & Barber, M. (2010). *How the world's most improved school systems keep getting better.* New York: McKinsey & Company. Available from: <http://www.cid.harvard.edu> [Accessed February 2011].

Mullis, I.V.S., Martin, M.O. & Foy, P. (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed May 2009].

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. & Chrostowski, S.J. (2004). *TIMSS 2003 international mathematics report*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed January 2008].

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. & Foy, P. (2007). *PIRLS 2006 international report.* Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed April 2007].

OECD (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris. Available from: <http://www.oecd.org> [Accessed May 2008].

OECD (2010a). *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes*. Paris. Available from: <http://www.oecd.org > [Accessed May 2010].

OECD (2010b). *PISA 2009 results: What students know and can do*. Paris. Available from: <http://www.oecd.org > [Accessed January 2011].

Sen, A. (1999). *Development as freedom*. New York: Alfred A. Knopf.

South Africa (2010). *Call for comments on Action Plan to 2014: Towards the Realisation of Schooling 2025* [Government Notice 752 of 2010]. Pretoria. Available from: <http://www.info.gov.za> [Accessed October 2011].

Stiglitz, J.E., Sen, A. & Fitoussi, J.-P. (2009). *Report by the Comission on the Measurement of Economic Performance and Social Progress*. Paris.

UNESCO (2006). *Education for All global monitoring report 2007: Strong foundations*. Paris. Available from: <http://www.unesco.org> [Accessed November 2006].
UNESCO (2008). *Primer reporte: Los aprendizajes de los estudiantes de América Latina y el Caribe.* Santiago: OREALC/UNESCO. Available from: <http://www.unesco.org> [Accessed March 2010].
UNESCO: UIS (2009). Education statistics (internet-based data querying facility). Montreal. Available from: <http://www.uis.unesco.org>
United Nations Statistics Division (2009). National accounts estimates of main aggregates (internet-based data querying facility). New York. Available from: <http://data.un.org>
Walker, M. (2011). *PISA 2009 plus results*. Camberwell: Australian Council for Educational Research. Available from: <https://mypisa.acer.edu.au > [Accessed December 2011].
World Bank (2010). World governance indicators (dataset). Washington. Available from: <http://www.worldbank.org> [Accessed June 2010].

Appendix 1: Normalised country scores

Country	*ISO code*	*PIRLS*	*PISA*	*SAC-MEQ*	*SER-CE*	*TIMSS Gr4*	*TIMSS Gr8*	*Overall*
Albania	ALB		373					373
Algeria	DZA					401	395	398
Argentina	ARG	401	391		392			395
Armenia	ARM					459	464	462
Australia	AUS		522			477	472	490
Austria	AUT	485	497			475		486
Azerbaijan	AZE		406					406
Bahrain	BHR						404	404
Belgium	BEL	477	513			502	497	497
Belize	BLZ	336						336
Bosnia and Herzegovina	BIH						442	442
Botswana	BWA			355			380	367
Brazil	BRA		381		392			387
Bulgaria	BGR	492	422				452	455
Canada	CAN	491	529			477	484	495
Chile	CHL		419		407		395	407
China (Shanghai)	CHN		578					578
Colombia	COL	403	387		388	387	390	391
Costa Rica	CRI		426		425			425
Croatia	HRV		470					470
Cuba	CUB				467			467
Cyprus	CYP	454				478	446	459
Czech Republic	CZE	484	495			464	475	479
Denmark	DNK	491	503			486		493
Dominican Republic	DOM				328			328
Ecuador	ECU				352			352
Egypt	EGY						403	403
El Salvador	SLV				370	373	363	369
England	ENG	491				493	476	487
Estonia	EST		507				493	500
Finland	FIN		543					543
France	FRA	474	501					488
Georgia	GEO	437	377			436	411	415
Germany	DEU	488	497			487		491
Ghana	GHA						331	331
Greece	GRC	475	463					469
Guatemala	GTM				353			353
Hong Kong	HKG	490	539			525	526	520
Hungary	HUN	491	487			484	488	487
Iceland	ISL	466	503					485
Indonesia	IDN	391	379				407	392
Iran	IRN	400				411	409	406
Ireland	IRL		506					506
Israel	ISR	465	448				458	457
Italy	ITA	491	473			475	460	475
Japan	JPN		523			511	520	518
Jordan	JOR		394				421	408
Kazakhstan	KAZ		398			501		449
Kenya	KEN			388				388
Korea	KOR		542				535	539
Kuwait	KWT	361				365	373	366
Kyrgyzstan	KGZ		310					310
Latvia	LVA	488	478			494	478	484
Lebanon	LBN						432	432
Lesotho	LSO			288				288
Liechtenstein	LIE		516					516
Lithuania	LTU	486	475			491	475	482
Luxembourg	LUX	498	474					486
Macao	MAC		509					509
Macedonia	MKD	417	377				428	407
Malawi	MWI			260				260
Malaysia	MYS		409				466	437

Country	ISO code	PIRLS	PISA	SAC-MEQ	SER-CE	TIMSS Gr4	TIMSS Gr8	Overall
Malta	MLT		453				464	458
Mauritius	MUS		414	417				415
Mexico	MEX		407		410			408
Moldova	MDA	455	393			475	445	442
Montenegro	MNE		401					401
Morocco	MAR	342				381	393	372
Mozambique	MOZ			332				332
Namibia	NAM			289				289
Netherlands	NLD	494	527			494	497	503
New Zealand	NZL	479	524			468	468	485
Nicaragua	NIC				361			361
Norway	NOR	457	497			450	448	463
Oman	OMN						385	385
Palestine	PSE						389	389
Panama	PAN		366		358			362
Paraguay	PRY				358			358
Peru	PER		339		373			356
Philippines	PHL					389	389	389
Poland	POL	472	492					482
Portugal	PRT		473					473
Qatar	QAT	354	343			353	341	348
Romania	ROU	458	419				450	443
Russia	RUS	491	462			494	479	481
Saudi Arabia	SAU						357	357
Scotland	SCO	477				468	467	471
Serbia	SRB		430				460	445
Seychelles	SYC			402				402
Singapore	SGP	488	544			529	539	525
Slovakia	SVK	475	483			470	478	476
Slovenia	SVN	466	496			467	470	475
South Africa	ZAF	318		322			311	317
Spain	ESP	467	480					473
Swaziland	SWZ			368				368
Sweden	SWE	497	506			474	469	486
Switzerland	CHE		514					514
Syria	SYR						388	388
Taiwan	TWN	483	521			513	534	513
Tanzania	TZA			385				385
Thailand	THA		422				432	427
Trinidad and Tobago	TTO	412	415					414
Tunisia	TUN		376			374	414	388
Turkey	TUR	422	441				426	430
Uganda	UGA			317				317
Ukraine	UKR					454	446	450
United Arab Emirates	ARE		426					426
United Kingdom	GBR		506					506
United States	USA	487	491			486	476	485
Uruguay	URY		425		427			426
Yemen	YEM					311		311
Zambia	ZMB			261				261
Zimbabwe	ZWE			345				345

Appendix 2: The Hanushek and Woessman (2009) approach

Hanushek and Woessman (2009: A14) use around 800 country scores from 77 countries, just over half of which are high income countries, to obtain normalised country scores that are comparable across all the 77 countries. The test data are from 32 different tests, where a test is specific to a year, an age (or grade level) and a subject. For instance, the 2003 TIMSS test in mathematics for Grade 8 would be a 'test'. The tests were part of the Programme for International Student Assessment (PISA) of the OECD or of the various assessment programmes of the International Association for the Evaluation of Educational Achievement

(IEA), such as TIMSS and PIRLS. They occurred in years during the period 1964 to 2003. Moreover, data from one national testing system, the National Assessment of Educational Progress (NAEP) of the United States, are used. There were test scores for three subjects, mathematics, science and reading, and for three institutional levels, primary, lower secondary and upper secondary.

The fact that correlations between country scores in the same subject but from different programmes were found to be high confirmed the viability of obtaining normalised scores following a single scale. A key end product of the normalisation process published by Hanushek and Woessman (2009: A15), and summarised here with some simplifications, is a general and comparable measure of cognitive ability per country, where this measure is an average across two subjects, mathematics and science, across all three institutional levels and spanning all the years for which the country had data.

The analysts take advantage of the fact that NAEP covers the three subjects previously mentioned and the three institutional levels, with testing occurring every two to four years, for the whole period in question. Added to this is the fact that the United States has participated in all the 32 international tests.

The approach used to obtain the normalised country scores can be summed up as follows. The aim of the approach is to transform all country scores to the scale followed by PISA in 2000. Each transformed country score *T* can be described as a linear function of the original country score *O.*

$$T = \alpha + \beta O$$

The slope coefficient β in the case of a specific test can be said to be obtained as follows. Here the original notation is used, but with the terms reorganised to display the linearity of the function.

$$\beta_{a,s,t} = \frac{SD^{OSG}_{s,PISA}}{SD^{OSG}_{a,s,t}}$$

The original country score *O*, for institutional level *a,* subject *s* and year *t* is converted to the PISA 2000 scale through multiplication by the ratio of the standard deviation across country scores in PISA in 2000 in subject *s* to the standard deviation across country scores with respect to *O. OSG* above refers to the 'OECD standardization group' or a group of 13 countries which displayed consistent participation across many tests.

The intercept α for the test covering institutional level *a,* subject *s* and year *t* is obtained as follows:

$$\alpha_{a,s,t} = O^{US}_{s,PISA} - O^{US}_{a,s,t}\left(\frac{SD^{OSG}_{s,PISA}}{SD^{OSG}_{a,s,t}}\right) + \left(NAEP^{US}_{a,s,t} - NAEP^{US}_{a,s,1999}\right)\frac{SD^{US,PISA}_{s}}{SD^{US,NAEP}_{a,s}}$$

The first term on the right-hand side is the US country score in PISA in 2000 in the subject in question, *s*. From this is subtracted the original US country score for the test in question converted to the PISA 2000 scale using the ratio referred to previously. The third term is the change over time in the US NAEP mean converted to the 2000 PISA scale using a ratio of standard deviations of scores across United States pupils. If the test in question is a 2000 PISA test then α reverts to zero (the difference between the two NAEP means would be zero as there was no NAEP test in 2000 and in its place the 1999 NAEP results would be considered).

Despite the time dimension within the approach described above, Hanushek and Woessman (2009) limit their use of the transformed country scores *T* to cross-sectional analysis. As was seen in this paper, there appears to be considerable noise in the trends over time, something which makes the interpretation of inter-temporal patterns difficult.