Globalised student achievement? A longitudinal and cross-country analysis of convergence in mathematics performance

Stefan Johansson a,b* and Rolf Strietholt a,b*

aDepartment of Education and Special Education, University of Gothenburg, Gothenburg, Sweden; bDortmund University of Technology, Center for Research on Education and School Development, Dortmund, Germany

ABSTRACT
With the aid of longitudinal country-level data from five IEA TIMSS assessments (1995–2011), the current study addresses the issue of the globalisation of curricula and achievement. To explore the hypothesis of global convergence, we study performance in four subdomains of mathematics. Using regression with fixed effects for countries, we consider whether the variation of subdomain scores decreases globally over time. Additionally, we explore qualitative differences in performance profiles using latent class analysis. Our results provide little evidence for a global harmonisation of student achievement. Rather, for regions with a similar language and culture, we observe similar strengths and weaknesses in mathematics content areas. Furthermore, these patterns remain stable over time. Directions for future research include the exploration of global trends in aspects of attained curricula for other subjects, and the use of information on school achievement.

KEYWORDS
International large-scale assessments; mathematics; comparative education; world curriculum; latent class analysis; TIMSS

World culture, regional convergence, and the isomorphism thesis

One of the central topics of the ‘world culture’ debate is the perceived convergence of national educational systems (e.g. Carney, Rappleye, and Silova 2012). The ways in which global forces affect education and national curricula have come under intense scrutiny from world culture theorists in recent decades (Dale 1999, 2000; Meyer, Kamens, and Benavot 1992; Meyer et al. 1997; Schriewer 2016). In this context, international assessments such as PISA and TIMSS have established indicators that are used to compare educational outcomes across countries. The perceived shortcomings of educational systems, as signalled for example by poor performance in such international rankings, can produce pressure for educational change, and a search for models from other countries which might solve the problems that are perceived in domestic systems (e.g. Baker and LeTendre 2005; Phillips and Ochs 2003; Spring 2008). High-achieving countries become role models for poorer-performing countries. However, ‘borrowed polices’ from high-achieving
countries are not easily implemented. Typically, the complex processes of decision-making, implementation and internalisation involve considerable transformation of the ‘borrowed’ policy (Carnoy and Rhoten 2002; Phillips and Ochs 2003; Steiner-Khamsi and Stolpe 2004). Furthermore, the potential drawbacks of ‘travelling’ policies can be that countries lose their own uniqueness, innovativeness, and creativity (Pettersson 2008). In such a scenario – and in spite of obvious national differences such as language, culture, and religion – it is suggested that countries tend to become increasingly similar over time, a process that in the literature is identified as isomorphism (Wiseman, Astiz, and Baker 2014).

In comparative education, there are different theories on the effects of globalisation. These theories aim to describe how educational systems become increasingly homogenous over time. To develop an understanding of diffusion and variation, a number of studies have used historical and functionalist world culture theories as a framing (see, for example, Benavot et al. 1991). In the historical tradition, research often rejects functionalist explanations about the emergence of institutions (Suárez and Bromley 2016). Applied to education, historic theories emphasise diversity in national curricula across countries or world regions, and a substantial degree of stability over time within countries or regions. According to functionalist theories, differences in national curricula stem from the different functions of education in different societies. Thus, while highly industrialised societies require more theoretical subjects in advanced mathematics and engineering, less industrialised societies emphasise instruction in vocational subjects and domestic science to a greater extent. As societies develop it is suggested that there will be an increased emphasis on modern subjects. Thus national curricula may vary in relation to the level of socioeconomic development; because socioeconomic differences are quite stable across countries, curricular differences would similarly reflect this stability (see, for example, Benavot et al. 1991).

However, these predictions have not found support in empirical work. In their pivotal study on school curricula, Benavot et al. (1991) did not find strong support for historic or functionalist theories. Furthermore, Meyer et al. (1997) developed the theoretical model of a ‘Common World Educational Culture’. They suggested that universal models of education and society, rather than distinct national factors, can explain the development of national education systems and curricula. National differences tend to be unstable and arise as a matter of chance in societies with differing political structures, cultures, and religious traditions. In the same vein, Dale (2000) proposed the idea of a ‘Globally Structured Agenda for Education’. This model, however, differs from ‘Common World Educational Culture’; because countries have a strong desire to compete in a global economy, Dale credits capitalism as the driving force behind curricular homogenisation. He argues further that countries do not converge globally but regionally, and that Europe, Asia and North America stand out as distinct regions. As an example, Dale and Robertson (2002) argue that regional organisations such as APEC (Asia-Pacific Economic Cooperation), EU (European Union) and NAFTA (North American Free Trade Agreement) have a major impact on education in their respective regions. Therefore, because of the differences between organisations’ activities, they contribute only marginally to a world-wide convergence. However they do make a contribution to an increasing globally-structured agenda for education within homogeneous regions (such as, Europe, Asia and North America). While previous research has attempted to empirically test these theories, it has
proved difficult to establish any firm evidence of the purported effects (see below). In this regard the tracing of patterns of diffusion and variation in achieved curricula – what students actually learn – is especially challenging.

In the current study, data from TIMSS on different facets of mathematics is used to test theories of harmonisation and convergence of education systems, the aim being to identify differences that might exist at the global and/or the regional level. Unlike PISA, the TIMSS data provides insight into learning in mathematics that can be linked to national curricula, as well as learning that is more generic in nature, and which is independent of particular curricula.

**How can the hypothesis of convergence be studied?**

There are a number of studies indicating that, with respect to what subject content is intended and implemented in school, countries tend to converge (e.g. Benavot et al. 1991; Bromley, Meyer, and Ramirez 2011). However, few studies shed light on how students’ patterns of knowledge actually develop over time. If curricula become increasingly similar, it is plausible that similar outcomes are produced across countries, and that similar performance strengths and weaknesses will appear. Here, the question involves which aspects of the intended curriculum (stipulated by a national education agency), and the implemented curriculum (taught by the teacher), are actually acquired by the student. Research covering this aspect is scarce, perhaps due to the fact that cross-national data is difficult to gather, especially when the focus is on long-term trends. However ILSA data may provide a possibility to explore the development of different school-systems. This is because, since the 1960s, comparative assessments have been regularly carried out, increasingly so after 1995 (Gustafsson 2008).

Only few studies have explored the relative strengths and weaknesses of different countries’ performances. Lie and Roe (2003) used PISA 2000 reading achievement data to investigate achievement patterns for the Nordic countries. They analysed the percentages of correct responses for each item in the dataset in order to discover whether or not the Nordic students had similar patterns of strengths and weaknesses in different aspects of the test. They found that Denmark, Norway, and Sweden shared a similar knowledge profile, thus having strengths and weaknesses on similar items. However, Finland was only weakly linked to the other Nordic countries. One reason put forward for this was that Finland performed especially well on more difficult items that required that students could read between the lines. Because the data used was from one measurement point only, Lie and Roe (2003) did not have the opportunity to address the issue of converging country performances over time. Furthermore, Kjærnsli and Lie (2008) carried out similar analyses on TIMSS 2003 data, when they investigated relative strengths and weaknesses in science achievement across a large group of countries. Kjarnsli and Lie used the residuals for the percentage corrects for science items in a cluster analysis, and found groups of countries with similar patterns. Countries that clustered within a group were, for example, English-speaking, East Asian, South-East Europe and Arabic countries. Further, the researchers identified characteristics of the different groups, and found that English-speaking countries tended to perform relatively better on the open constructed response items, in comparison with the other groups of countries. South-European countries, on the other hand, performed better on multiple-choice items.
Although, groups of countries tend to share similar languages and have similar cultures, Kjarnsli and Lie concluded that linguistic factors alone did not play a crucial role in linking the countries together.

While in mathematics at least there appear to be regional groupings of countries with respect to their various strengths and weaknesses, the cross-sectional studies summarised in the previous section do not provide insights about development over time. In particular, it might be asked whether fewer clusters might emerge over time.

Rutkowski and Rutkowski (2009) used similar analytical methods to the previously cited authors. However, by considering the extent to which students’ responses on TIMSS became more similar over time, they shifted focus to address the hypothesis of temporal convergence. In their study, they examined student responses to test items for a wide range of countries participating in TIMSS 95, 99, and 03. The underlying question was whether there could be any support for the theories of world education harmonisation (e.g. Dale 1999, 2000; Meyer et al. 1997). Similarly to previous studies, Rutkowski and Rutkowski (2009) identified similarities on a regional level (e.g. English-speaking, East Asian or East European) that were stable over time. However, their findings did not indicate an influence of global forces. Further, the timespan that their study covers might be too limited to capture global patterns of change and convergence. Furthermore, studying curricula change at item-level, instead of broader content areas, may not be appropriate where the aim is to capture curricula change (curricula do not cover test items, but refer to more general constructs such as Pythagoras’ theorem or geometry). Likewise, the approach with the p-value residuals used in the studies by the Lie and colleagues may be overly detailed, thus making it difficult to find curricular changes at the global level. This strategy might however be adequate to identify a group of countries where students are particularly good at items with certain characteristics, such as multiple choice or open response questions, or another group where students are good at problem-solving. However, in order to approach global changes in patterns of knowledge, we believe that it is necessary to examine larger streams of knowledge, such as subject content domains in mathematics (e.g. geometry, algebra). We also contend that, since curricular reforms do not happen overnight, it is of vital importance to extend the time period of study, so as to be able to discover evidence of convergence/differentiation.

The present study: tracing mathematical achievement over time

The present study specifically focuses on aspects of attained mathematics curricula, namely trends in student achievement across 60 educational systems. We aim to do this with the aid of five TIMSS cycles, covering a time period between 1995 and 2011. Compared to OECD studies such as PISA which attempt to capture competencies regarded as important for adult life and life-long learning, TIMSS focuses on curriculum-defined knowledge and skills (Wu 2010). Studies show that the TIMSS framework and national curricula are well aligned (e.g. The Swedish National Agency for Education 2008). Consequently, aspects of curricular development may be empirically testable using data from TIMSS.

The main purpose of this study is to investigate whether there has been a harmonisation of countries’ performances, resulting in either a global (i.e. ‘Common World Educational Culture’ hypothesis) or a regional (i.e. ‘Globally Structured Agenda for Education’ hypothesis) convergence. The basic idea is to investigate the performance
profiles in four mathematic content areas, *algebra, data and chance, geometry, and numbers*, and to track changes in countries’ relative strengths and weaknesses over time. To test the hypothesis of a harmonisation of school systems and curricula, we investigate whether there is any global or regional convergence between countries with respect to performance patterns in mathematics.

**Methods**

**International TIMSS data from 1995 to 2011**

We use secondary school data from five assessment cycles of the Trends in International Mathematics and Science Study (TIMSS). The data were collected in 1995 (136,973 students from 40 countries), 1999 (237,833 students from 35 countries), 2003 (237,833 students from 51 countries), 2007 (245,553 students from 57 countries) and 2011 (281,995 students from 49 countries). The samples refer to 87 countries. Of these, 15 countries took part in all five study cycles, 10 countries took part in four cycles, 21 countries in three cycles, 13 in two cycles, and 28 in only one of the cycles. Because the aim is to analyse change over time, we only consider data from the 59 countries that participated in at least two cycles. We base our analyses on a pooled dataset containing 204 country-by-year observations. TIMSS samples entire classes within schools and data sets cover sampling weights to compensate for unequal sampling probabilities. These weights were used to generalise the findings from the samples to the target populations in the respective countries. Some countries tested different grades, and this limits comparability over time. For this reason, we restrict our analyses to the data from the grade where most students were assessed in each respective country (which in most countries was grade 8).

**Achievement in different areas of mathematics**

The subject of mathematics consists of varying content areas that together form the concept of mathematics. In the present study, we follow delineations from the latest TIMSS cycles, and which enable four content areas to be distinguished: *algebra, data and chance, geometry, and numbers*. Table 1 shows that the earlier study cycles reported slightly different content areas. However, it should be noted that some of the differences

| Table 1. Descriptions of the mathematics tests. |
|-----------------------------------------------|
| **Items** | TIMSS95 | TIMSS99 | TIMSS03 | TIMSS07 | TIMSS11 |
|-----------|---------|---------|---------|---------|---------|
| Time      | Items   | Items   | Items   | Items   | Items   |
| 45 min    | 151 (125 MC, 26 CR) | 162 (125 MC, 37 CR) | 194 (128 MC, 66 CR) | 215 (117 MC, 98 CR) | 217 (118 MC, 99 CR) |
| Content   | Fractions & number sense (51) | Fractions & number sense (61) | Fractions & number sense (61) | 45 min Number (57) | 45 min Number (63) |
| areas     | Algebra (27) | Algebra (35) | Algebra (47) | Algebra (64) | Algebra (70) |
|          | Data representation, analysis, & probability (21) | Data representation, analysis, & probability (21) | Data representation, analysis, & probability (21) | Data (28) | Data & chance (41) |
|          | Geometry (23) | Geometry (21) | Geometry (31) | Geometry (47) | Geometry (43) |
|          | Measurement (8) | Measurement (24) | Measurement (31) | Measurement (47) | Measurement (43) |
|          | Proportionality (11) | | | | |

Note: MC: multiple choice; CR: constructed response (Open Items).
are not substantive, but rather terminological. The area *numbers* was initially labelled *fractions and number sense*. The terms *algebra* and *geometry* were consistently used across all cycles. The content area *data and chance* were labelled *data or data representation, analyses, and probability* in earlier study cycles (see also, Martin, Gregory, and Stenler 2000; Martin and Kelly 1998; Martin and Mullis 2012; Martin, Mullis, and Chrostowski 2004; Olson, Martin, and Mullis 2008). In addition to the terminological changes, two areas that were reported separately in early study cycles – *proportionality* and *measurement* – were not separately reported in the more recent cycles.

To achieve consistency over time, we limit our analyses to the four content areas *algebra, data and chance, geometry, and numbers*. The items that refer to *proportionality* and *measurement* were redistributed to the four areas. A subject matter expert in mathematics ensured that a correct assignment of the items in these two areas was achieved. Also, some of the items in *proportionality* and *measurement* categories were assigned to other domains by the TIMSS administration. This was because these items recurred again in later assessments. For example, items from category *measurement* in 2003 were found in category *geometry* in 2011.

To measure student proficiency in mathematics, in each cycle the TIMSS tests include between 159 and 217 test items. Each item relates to one of the four content areas of *numbers, algebra, geometry, and data and chance*. The testing time was 45 min and, in each cycle, each content area was assessed by at least 19 test items. To study trends, we used overlaps in the assessment material of the respective study cycles to equate them. The common-item nonequivalent group design, and the item response theory (IRT) models that were used to link all tests onto the same metric are described in the Technical Appendix.

**Relative strengths and weaknesses in mathematics: centred scores**

To quantify relative strengths and weaknesses for each of the 204 country-by-year observations, we first used the IRT scores to compute the mean scores in algebra, data and chance, geometry, and numbers. These scores reflect not only relative strengths and weaknesses in the four content areas but also differences in the general performance levels in the respective country-by-year observations. For example, industrialised countries tend to perform higher on all scores, while developing countries tend to perform lower. To remove such general differences in performance level, we computed centred scores which were obtained by subtracting the mean of the four scores from each of these scores. This was done separately for each country-by-year observation. Thus the centred scores reflect the strength or weakness of each observation in a particular subdomain. By implication, the sum of the four centred scores for each country-by-year observation is zero.

|                  | Centred scores | Homogeneity Score |
|------------------|----------------|-------------------|
|                  | Algebra | Data & chance | Geometry | Numbers |                 |
| Mean             | 0.000   | 0.000          | 0.000    | 0.000   | 0.701           |
| SD               | 0.241   | 0.281          | 0.205    | 0.148   | 0.333           |
| Min              | -0.663  | -1.175         | -0.520   | -0.367  | 0.077           |
| Max              | 0.722   | 0.648          | 0.687    | 0.350   | 2.355           |
| country-by-year observations | 204     | 204            | 204      | 204      | 204              |
Table 2 shows that all centred scores have a mean that is zero, but that they vary across countries.

To illustrate the data at hand, Figure 1 plots the performance profiles of the Russian Federation and Jordan for each time point. Data for Jordan is missing in 1995 because the country did not participate in that TIMSS cycle. For example, as compared to the mean achievement observed in the Russian sample in 1995, their score was very high in algebra (0.435), moderately high in geometry (0.140), very low in data and chance (−0.370), and moderately low in numbers (−0.204). It should be noted that the mean of the four scores is zero. A flatter example of a performance profile is Jordan in 1999, with a small positive value in geometry (0.119) and small negative values in algebra (−0.013), data and chance (−0.101) and numbers (−0.005). Again, by implication, the mean of the four scores is zero.

**Homogeneity score: quantifying the balance of across mathematics areas**

Some countries’ performance profiles have pronounced strengths and weaknesses in particular content areas, while others are mostly flat. To describe the balance across the four centred scores, we constructed a measure that summarises the differences among the scores in a single homogeneity score. The homogeneity score is simply defined as the sum of the absolute values (i.e. the value of x without regard to its sign; it is denoted as |x|) of the four centred scores. A homogeneity score close to zero indicates a flat performance profile, while large values mark uneven profiles with pronounced strengths and weaknesses in particular content areas. While the average score for all countries is about 0.7, the standard deviation indicates substantial variability across countries (see Table 2). For example, the homogeneity score for the aforementioned Russian sample from 1995 is 1.148 (|0.435|+|−0.370|+|0.140|+|−0.204|) but for Jordan in 1999 it is only 0.239 (|−0.013|+|−0.101|+|0.119|+|−0.005|). The homogeneity scores are also depicted below each graph in Figure 1, this overview illustrating that flat profiles have lower scores than uneven profiles.

**Analytical strategies**

Two analytical strategies are used to investigate global and regional convergence in mathematics performance. First, to test the hypothesis of a global convergence, investigations of change in homogeneity scores over time were conducted. In an extension to this analysis, latent class analysis (LCA), was applied. This is a more explorative strategy. Here, the purpose was to identify groups of countries with similar performance profiles in mathematics, and to investigate regional differences and how they change over time.

**Change in the homogeneity score as an indicator of global convergence**

To test global convergence in mathematics, we investigated whether the performance in the four content areas of mathematics became more even over time. For this purpose, we analysed change in the homogeneity score using the pooled data from different years (assessment cycles). Here, the key issue to take into account is that the number and composition of countries that participated in the respective study cycles has changed over the years. Our approach to this issue can be most easily seen from a simple linear model with fixed effects for countries. We regress the homogeneity score $h$ in country $c$ at time $t$ on
Figure 1. Performance profiles in mathematics content areas for Russian Federation and Jordan for each time point.

Note: The symbols used in this graph match the symbols used in Figures 2 and 3.
the year of the assessment cycles \( t \). Furthermore, we added a set of dummies for countries, \( \mu_c \), to control for changes in the composition of countries:

\[
h_{ct} = \alpha + \beta t + \mu_c + \epsilon_{ct}
\]  

(1)

We are mainly interested in estimating \( \beta \), the change in the homogeneity score, holding change in the composition of countries constant. By implication, the estimation of \( \beta \) is based upon variation over time, since changes in the composition of countries are absorbed into the country fixed effects. The relevant variation with which we identify \( \beta \) is within-country variation of the homogeneity score.

**Tracking qualitative change in performance profiles**

While the unidimensional homogeneity score is a useful measure to summarise how flat or uneven countries’ performance profiles are, it is, however, unable to capture qualitative differences in the profile (i.e. strengths and weaknesses in particular content areas). Having different performance profiles in terms of strengths and weaknesses in the four content areas could for example depend on aspects such as how national curricula emphasise certain content areas. The homogeneity score does not reveal such differences. To investigate signs of harmonisation, we also considered qualitative differences in performance profiles. For this purpose, we used LCA. Using an LCA, we could identify classes of country-by-year observations with distinct performance profiles in mathematics. The four centred scores in *algebra, data and chance, geometry, and numbers* define the latent class variable.\(^2\) As a reference group, we constrained one latent class to have a perfectly balanced flat performance profile, where each centred score is zero. In all further classes, the performance profiles were estimated freely. Based on this classification, we explore stability and change in the performance profiles across countries and over time.

An important step in latent class analysis is to determine the number of latent classes models in the data. As we did not have compelling reasons to decide on the number of
performance profiles in advance, we carried out estimations for eight candidate models, each with one to eight latent classes. Making selections among these models is driven by both empirical and substantive considerations. We used the likelihood-based Bayesian information criterion (BIC; Schwarz 1978) where smaller values indicate better fitting models. Further, the parametric bootstrapped likelihood ratio test provides a p-value for the comparison of a model that has \( k \) classes with an alternative model that includes \( k - 1 \) classes (BLRT; McLachlan 1987; McLachlan and Peel 2000). In addition, the entropy

| Country              | 1995 | 1999 | 2001 | 2007 | 2011 |
|----------------------|------|------|------|------|------|
| Netherlands          | △    | △    | △    | △    | △    |
| Scotland             | △    | △    | △    | △    | △    |
| Chile                | △    | △    | △    | △    | △    |
| New Zealand          | △    | △    | △    | △    | △    |
| Norway               | △    | △    | △    | △    | △    |
| Sweden               | △    | △    | △    | △    | △    |
| England              | △    | △    | △    | △    | △    |
| United States        | △    | △    | △    | △    | △    |
| Australia            | △    | △    | △    | △    | △    |
| Slovenia             | △    | △    | △    | △    | △    |
| Lithuania            | △    | △    | △    | △    | △    |
| Belgium Remish       | △    | △    | △    | △    | △    |
| Botswana             | △    | △    | △    | △    | △    |
| South Africa         | △    | △    | △    | △    | △    |
| Thailand             | △    | △    | △    | △    | △    |
| Turkey               | △    | △    | △    | △    | △    |
| Canada               | △    | △    | △    | △    | △    |
| Colombia             | △    | △    | △    | △    | △    |
| Hungary              | △    | △    | △    | △    | △    |
| Italy                | △    | △    | △    | △    | △    |
| Indonesia            | △    | △    | △    | △    | △    |
| Malaysia             | △    | △    | △    | △    | △    |
| Tunisia              | △    | △    | △    | △    | △    |
| Morocco              | △    | △    | △    | △    | △    |
| Saudi Arabia         | △    | △    | △    | △    | △    |
| Bahrain              | △    | △    | △    | △    | △    |
| Ukraine              | △    | △    | △    | △    | △    |
| Dubai, UAE           | △    | △    | △    | △    | △    |
| Qatar                | △    | △    | △    | △    | △    |
| Spain, Basque        | △    | △    | △    | △    | △    |
| Latvia               | △    | △    | △    | △    | △    |
| Kuwait               | △    | △    | △    | △    | △    |
| Iran                 | △    | △    | △    | △    | △    |
| Japan                | △    | △    | △    | △    | △    |
| Israel               | △    | △    | △    | △    | △    |
| Cyprus               | △    | △    | △    | △    | △    |
| Slovak Republic      | △    | △    | △    | △    | △    |
| Czech Republic       | △    | △    | △    | △    | △    |
| Hong Kong            | △    | △    | △    | △    | △    |
| Korea, Rep. of       | △    | △    | △    | △    | △    |
| Singapore            | △    | △    | △    | △    | △    |
| Oman                 | △    | △    | △    | △    | △    |
| Palestinian Auth.     | △    | △    | △    | △    | △    |
| Egypt                | △    | △    | △    | △    | △    |
| Ghana                | △    | △    | △    | △    | △    |
| Syrian Arab Republic | △    | △    | △    | △    | △    |
| Jordan               | △    | △    | △    | △    | △    |
| Chinese Taipe (TWN)  | △    | △    | △    | △    | △    |
| Romania              | △    | △    | △    | △    | △    |
| Russian Federation   | △    | △    | △    | △    | △    |
| Macedonia, Rep. of   | △    | △    | △    | △    | △    |
| Moldova              | △    | △    | △    | △    | △    |
| Bulgaria             | △    | △    | △    | △    | △    |
| Serbia & Montenegro  | △    | △    | △    | △    | △    |
| Lebanon              | △    | △    | △    | △    | △    |
| Armenia              | △    | △    | △    | △    | △    |
| Georgia              | △    | △    | △    | △    | △    |

| relative frequency class 1 △ | 0.34 | 0.29 | 0.24 | 0.15 | 0.20 |
|-------------------------------|------|------|------|------|------|
| relative frequency class 2 △ | 0.34 | 0.29 | 0.24 | 0.15 | 0.20 |
| relative frequency class 3 △ | 0.41 | 0.41 | 0.41 | 0.41 | 0.41 |

Figure 3. Country performance profiles from 1995 to 2011.
Note: The symbols used in this graph match the symbols used in Figures 1 and 2.
quantifies the precision to which each country-by-year observation can be placed into classes (Ramaswamy et al. 1993). We also paid particular attention to the estimated frequency of country-by-year observations in each class. This is because classes with just a few observations indicate an overextraction of classes, rather than a class with a substantive meaning (Masyn, Henderson, and Greenbaum 2010).

Results

Global convergence

As a baseline for our longitudinal analyses (presented below), we pooled the country-level data from the five assessment cycles, and regressed the homogeneity score on the year of the assessment. The results for this analysis suggest that the homogeneity score has not changed over time. This would thus indicate that the differences between mathematics subdomains are quite similar over time. Results are reported in column 1 in Table 3. However, this analysis estimates the correlation between the homogeneity score and the year of the assessment without taking into account that the composition of the countries had changed over the years. If for a moment we assume that there was a global convergence over time, what we observe may be that countries with more unbalanced performance profiles participated in more recent studies. Following this line of argument, the observed zero correlation is due to the reverse effects of time and the composition of countries at the respective time points.

Because it exploits longitudinal variation on a country-level, the main model with country fixed effects effectively controls for the fact that the composition of countries that participate in TIMSS has changed over time. The results for this longitudinal analysis show a negative effect of time on the homogeneity score that is statistically significant at the 10% level (column 2). As the time-variable was coded in such a way that the unstandardised regression parameter stands for the linear change between 1995 and 2011, the observed effect of $-0.074$ corresponds with a decrease in the homogeneity score that amounts to 0.222 standard deviations ($-0.074/0.333$) over a period of 16 years. This implies that countries’ scores have become more balanced over time, meaning that the differences between mathematics subdomains became smaller.

Table 3. Testing convergence in the homogeneity of math performance.

|                      | Cross-sectional | Longitudinal |
|----------------------|-----------------|--------------|
|                      | 1               | 2            | 3            |
| Time (continuous)$^a$| $-0.005 (0.070)$| $-0.074^* (0.042)$|               |
| Time = 2007          |                 | 0.009 (0.038) |               |
| Time = 2003          |                 | 0.003 (0.039) |               |
| Time = 1999          |                 | 0.000 (0.043) |               |
| Time = 1995          | $0.110^{**}$ (0.046) |               |
| Country fixed effects| No              | Yes          | Yes          |
| Country-by-year observations | 204       | 204        | 204        |
| Countries            | 59              | 59          | 59          |
| $R^2$                | 0.000           | 0.716       | 0.723       |

Notes: Each column presents results of a separate regression. Dependent variable is the homogeneity score. Standard errors are in parentheses.

$^a$In columns 1 and 2, time of the assessment was as a continuous variable with the values $0.00 = 1995$, $0.25 = 1999$, $0.50 = 2003$, $0.75 = 2007$, and $1.00 = 2011$.

*$10%$ significance level (two-tailed).

**$5%$ significance level (two-tailed).
To explore non-linearity in the change of the homogeneity score, we replaced the continuous time-variable by dummies for each year, using 2011 as the reference year. Although this analysis confirms that the homogeneity score has decreased over time, it is important to note that the change occurred only between 1995 and 1999. The change between 1995 and 1999 corresponds roughly with the effect size observed in the previous analysis, where time was modelled as a continuous variable. Since 1999, however, we observe hardly any evidence for a global convergence in mathematics performance. Indeed, since 1999, the variation between the four content areas is remarkably stable.

**Performance profiles across countries and over time**

As an extension to the results on change in the homogeneity score previously presented, we will next explore performance profiles across the four content areas: algebra, data and chance, geometry, and numbers. To study these qualitative differences, the four observed centred scores were used to specify latent class models with one to eight latent classes. For sake of simplicity, we show model fit for one to four classes only, as the fit decreases after four classes. **Table 4** shows that the model with four classes has the best BIC value, and the highest entropy. Additionally, the BLRT rejects the three-class model over the four-class model. However, the four-class model produced a very small class (only 1% of the country-by-year observations) with parameters that were probably not reliably estimated. The other three classes of the four-class model have a similar frequency distribution in the three-class model. This suggests that the three-class model provides a more parsimonious way of describing the latent structure. For this reason, we can make the tentative inference that the latent profiles in mathematics, as measured by the TIMSS tests and represented in our data, derive from three latent subpopulations.

As illustrated in **Figure 2**, the three subpopulations have distinct performance profiles in the four mathematics content areas. We constrained the performance profile in the first class to be perfectly balanced across all four content areas, so that the mean performance in each subdomain is zero. In contrast to this flat profile, the second class is characterised by relative strengths in numbers and data and chance, and by relative weaknesses in algebra and geometry. The third class is a reverse image of the second class. It should be noted that we use different symbols to denote the performance profiles. Before we examine the full sample of countries, it is worthwhile taking a second look at **Figure 1**.

**Table 4.** Fit indices for models with one to four latent classes.

|       | 1        | 2        | 3        | 4        |
|-------|----------|----------|----------|----------|
| Log Likelihood | 102.235  | 130.576  | 193.338  | 222.996  |
| Estimated free parameters | 4        | 8        | 12       | 16       |
| BIC   | −183.197 | −218.608 | −322.858 | −360.901 |
| BLRT (p) | −        | <0.01    | <0.01    | <0.01    |
| Entropy | 1        | 0.735    | 0.730    | 0.809    |
| f (rf) | 204 (1.00) | 172 (0.84) | 107 (0.53) | 94 (0.46) |
| f (rf) | 32 (0.16) | 49 (0.24) | 56 (0.28) |
| f (rf) | 48 (0.23) | 51 (0.25) |
| f (rf) | 3 (0.01)  |

Notes: BIC: Bayesian information criterion; BLRT: bootstrapped likelihood ratio test (p-value); f: frequency; rf: relative frequency.
The figure shows that, in all five years, the Russian Federation is assigned to the class with relative strengths in algebra and geometry, and relative weaknesses in data and chance and numbers. The profiles of Jordan, on the other hand, are less stable. In 1999 and 2003 they are most similar to the balanced profile. However, in 2007 and 2011, they switch to the class with relative strengths in algebra and geometry, and relative weaknesses in data and chance and numbers.

Moving on, we now use the three-class model classification of the 204 country-by-year observations to explore stability and change in performance profiles over time. The most likely latent class membership of each country-by-year observation is plotted in Figure 3. In order to simplify the interpretation of the results, rather than in alphabetic order, the countries are listed by class membership. As an extension to the results on change in the homogeneity score that we previously presented, the within-country change in class membership offers a more nuanced picture of qualitative differences in countries’ performance patterns. At the same time, the analyses are by nature more explorative. Once again, the LCA provides little evidence for a global convergence towards a particular performance profile. On the contrary, pronounced strengths or weaknesses exist and persist over time. Although some countries change from the flat profile (class 1) to a class with more pronounced strengths and weaknesses (or vice versa), we do not find a clear trend that certain performance patterns replace others. Rather, transitions between classes appear stochastic.

Interestingly, there are quite striking and persisting similarities with respect to culture, region, and language in some of the classes. The composition of countries with flat performance profiles, class 1, is notably diverse. Consequently, we draw the tentative conclusion that the two other classes are composed of countries that share a set of different commonalities. The English-speaking countries in the sample, New Zealand, England, Australia, Scotland, and the United States, as well as the two Nordic countries, Norway and Sweden, have relative strengths in data and chance and numbers, and relative weaknesses in algebra and geometry (class 3). In contrast, most Post-Soviet states in our sample, including Armenia, Georgia the Russian Federation, Moldova, and several of the Balkan countries (Bulgaria, Republic of Macedonia, Romania, and Serbia and Montenegro) all show relative weaknesses in data and chance and numbers, and relative strengths in algebra and geometry (class 2). Apparently, similarities in terms of culture, region, and language tend here to be manifest in common achieved curricula.

Discussion

The overall aim of the current study has been to examine if countries converge with respect to their patterns of knowledge in mathematics subdomains. If countries’ performance patterns were to harmonise over time, this could be taken as a sign of a development towards a global curriculum. However, the results showed little evidence for a convergence at a global level. Rather, there is compelling evidence that tradition and culture are strong forces when it comes to students’ content knowledge. Similarities in culture and language seem to have a substantial impact on what students learn and their knowledge patterns, with countries within the same region/culture/language being clustered together on many occasions. The relatively stable pattern that characterises certain clusters of countries provides evidence that would seem to refute the concepts and models
suggested in world culture theory (cf. Schriewer 2016). Rather, historical the patterns we found may be better explained by institutionalism, which, because of its focus on phenomena such as ‘path-dependency’ and ‘self-reinforcing feedback mechanisms’ in processes of institutional elaboration, seems to offer a more compelling account for our findings. A main point of this theory is that social changes evolve quite slowly, meaning that studies over shorter time-frames may not trace any institutional change.

Regional patterns of knowledge

Having the slow evolvement of institutions in mind, the patterns of strengths and weaknesses in subdomains of mathematics may be worth highlighting. We observe that the subdomains of numbers and data are the strengths for one group of countries, while algebra and geometry are strengths in another of the groups. This pattern of differences can be traced to countries that share similar features, such as language, culture and tradition, while the clusters per se are rather diverse with respect to these factors. The cluster with relative strengths in algebra and geometry comprises primarily the post-Soviet states. Going into more detail, we can note that theories of mathematics education, and consequently the creation of curricula, are quite different between the two clusters of countries (Figure 2, class 2 and 3). Whereas in Russia, for example, algebra is traditionally taught from primary school upwards, in some Western countries’ algebra is not introduced until the eighth grade. In Russia, the study of algebra also precedes the study of arithmetic, this since the curriculum develops knowledge of algebraic structures through relationships between quantities such as length, area, and weight (where a letter denotes a specific quantitative property of an object), rather than from real numbers (see, for example, Schmittau and Morris 2004). In fact, this is quite the opposite of how mathematics teaching is structured in many of the Western countries. While there has been some interest in adapting to strategies across national borders, the differences seem nevertheless to be still well-manifested in students’ test scores. The Iron Curtain separated East and West until 1991, and even in the 2000s a mathematics educator in the West may only have had a vague understanding of what was taking place in Russian mathematics education, and vice versa (Karp 2006).

Educational change: national reforms or international trends?

As regards the policy impact of ILSAs such as PISA and TIMSS, the implications of the current study call for a more balanced discussion about how results might impact on educational outcomes. In this regard, it is important to consider whether educational reforms/changes operate at national, or a global scale. Furthermore, it is important to stress the distinction between educational reforms, and actual changes in educational outcomes. While there is evidence that the international organisations that conduct international assessment influence the global educational discourse (see, for example, Meyer, Strietholt, and Epstein 2018 on UNESCO, IEA and OECD), it is unclear as to whether ILSAs lead to similar or divergent reforms across countries. For example, Grek (2009, 35) has suggested that ‘PISA is a major governing resource for Europe’. In her in-depth comparison of how PISA 2000 results were perceived in Finland, Germany, and the UK, it is revealed that only Germany implemented major curricula reforms. Traditionally, it is
not the German federal government but the 16 states that have jurisdiction over education. In response to the PISA findings, however, a national conference of federal and state ministers was organised in 2002, and reforms such as the development and implementation of national educational standards and the introduction of large-scale assessment testing at the end of primary and secondary education were implemented in subsequent years. New projects like ‘Chemie im Kontext’ and ‘Physik im Kontext’ were launched in direct response to the PISA testing model. Obviously, these policy changes and reforms are specific to the German context, and there is no evidence that Finland and the UK implemented similar projects. Nevertheless, the hypothesis of an international harmonisation of performance profiles relies on the assumption that different countries do in fact implement similar reforms.

**TIMSS influence on national curricula**

Furthermore, there seems to be an emerging trend where many countries lack pronounced strengths or weaknesses in later TIMSS cycles; rather, they perform fairly similarly in all subdomains. This might explain why students in these countries had the opportunity to learn the tested content, and that the links between the intended, implemented and attained curricula are strong. However, the countries without relative strengths and weaknesses perform quite differently in absolute terms. While some countries are among the top performers, others are among the low performers. If the intended curricula overlap substantially across all these countries, there is likely to be something else that differs (e.g. teacher instruction, school resources, etc.). The contextual differences may be substantial since variation in achievement is large among these countries.

Another interpretation of the trend towards less pronounced strengths and weaknesses may be that certain countries are ‘teaching to the test’ to higher degree than others (see, for example, Biggs 1999; Koretz 2002). The tests in TIMSS should be aligned to curriculum goals but it seems anyway reasonable that the focus on the tests in ILSAs, such as TIMSS, varies across countries. Further, the ILSA results have higher stakes for certain countries (cf. Grek 2009) and may even drive curriculum changes. Indeed, the curriculum movements have been strong in many countries, perhaps particularly in East Asia (see, for example, Leung and Li 2010). We also observe some changes as regards the knowledge patterns for several East Asian countries, such as Singapore, Hong Kong, Korea and Taiwan. In the recent past, some countries’ performances, including Singapore’s, have become more focused on algebra and geometry, whereas they had previously performed evenly across all four subdomains. This might be due to the fact that they strive to follow the development of the tested content in TIMSS, where in the more recent assessments the number of algebra items has increased.

**Limitations and further research**

The present study is not without limitations. First, studying the harmonisation of performance patterns using international test score data is based on the assumption that the tests provide valid measures of educational outcomes for different countries. In this regard, it is important to emphasise that IEA studies such as TIMSS are curriculum-based, which means
that the tests are based on a review of the curricula of various countries. The OECD’s PISA, on the other hand, does not claim to assess student achievement based on actual curricula, but rather on what, in the view of the OECD experts, students need to know (Lockheed and Wagemaker 2013; Wu 2010). For this reason, we consider IEA’s TIMSS as a more eligible source for the present investigation. Although we would not venture as far as the naïve assumption that TIMSS captures all aspects of mathematics specified in national curriculum worldwide, we believe that the four content areas of algebra, numbers, data and chance, and geometry cover at least most of them. Furthermore, it is important to emphasise that a lack of content coverage would not necessarily conflict with the main findings of the present study. While we had the possibility to include a large number of countries in our study, it is worthwhile noting that a large majority of the countries have a relatively high human development index (HDI), and that countries in Africa and Latin America are relatively few. As made clear by Carney, Rappleye, and Silova (2012), researchers sometimes tend to ‘describe the world’ based on test data from a set of countries, which in fact may be far from complete. Thus, we recognise too that our sample is not fully representative of the global situation.

A second limitation when studying global forces in education is the necessity to achieve a high level of aggregation. At the end of the day, there may be some facets in the data that cannot be captured. For example, there is likely to exist some variability with respect to knowledge patterns within the three clusters we have identified. The strengths and weaknesses of each country within a cluster are not identical. Furthermore, differences exist within countries. In order to be able to observe the general patterns of the phenomena in focus at a collective level, it becomes necessary to look beyond the more specific and context-bound individual characteristics. Hence, for the purposes of this study, aggregation was essential. In this respect, we believe that we managed to trace larger streams of development over time. Our hope is that the patterns of achievement across countries that were revealed, as well as the results of individual countries, will serve as a useful basis for fruitful discussions on global processes in education in relation to ILSAs.

Third, as explained by Meyer et al. (1997), it may be difficult to bring curricular innovations into the classroom. While we might have benefitted from extending the time-period, we cannot with confidence state that we would be able to trace more obvious changes in students’ test scores. It should also be noted that some countries lack the resources to reform their school system, leading to a good deal educational reforms that are merely symbolic. Based on the discussion of the possible impact of ILSA’s on national policies, this may be especially relevant because, in their strivings to attain higher educational results, countries might be tempted to copy the approaches of the highest-performers.

Last but not least, a natural limitation of the present study is the focus on mathematics. We did not include other domains such as reading, science, and civic education, or for that matter universal educational aims such as moral education, citizenship and employability. Although we believe that the investigation of specific aspects of the curriculum is important and has methodological advantages (e.g. data for all subdomains is available), we acknowledge the need for further research. For a nuanced understanding of education, research on other educational aims is needed. In the same vein, the restricted focus on grade eight is a limitation of the present study. We hope that our study motivates further studies that replicate our approach, and that the findings from such studies may contribute to establishing a more holistic perspective of attained curricula worldwide.
Notes

1. The term country mostly refers to national states but sometimes also to regions within countries if their educational systems are highly autonomous (e.g. Flanders and Wallonia in Belgium).
2. The residual scores for numbers, algebra, geometry, and data and chance were constructed in such a way that they were not independent of each other because the sum of the scores equals zero for each case. To model this dependency, we constrained one mean within each class: \( \text{mean}_{\text{data & chance}} = - (\text{mean}_{\text{number}} + \text{mean}_{\text{algebra}} + \text{mean}_{\text{geometry}}) \). Furthermore, we constrained the variances of the residual scores to be equal across classes.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by grants from the Swedish Research Council (grant number 726-2013-296).

Notes on contributors

Dr. Stefan Johansson is senior lecturer and researcher at the University of Gothenburg, Department of Education and Special Education. His research interests center on validity issues of educational assessments, in particular on consequential aspects of validity in international large-scale assessments (ILSA). In his previous publications, he has utilized ILSA data to examine different assessment forms, curricula development, and the role of teacher competence for student achievement. Stefan is a former recipient of the IEA Bruce H. Choppin Memorial Award.

Dr. Rolf Strietholt is a researcher at Technische Universität Dortmund. He is also affiliated with the University of Gothenburg. His current interests lie in the field of international comparisons of educational systems, so-called comparative education, and include educational effective research studies with a special focus on measuring and explaining inequalities in student performance. Rolf teaches courses on educational measurement and causal analysis. Rolf is a former recipient of the IEA Bruce H. Choppin Memorial Award.

ORCID

Stefan Johansson http://orcid.org/0000-0002-2051-7248
Rolf Strietholt http://orcid.org/0000-0001-6344-642X

References

Baker, David, and Gerald LeTendre. 2005. National Differences, Global Similarities: World Culture and the Future of Schooling. Stanford, CA: Stanford University Press.

Benavot, Aaron, Yun-Kyung Cha, David Kamens, John W. Meyer, and Suk-Ying Wong. 1991. “Knowledge for the Masses: World Models and National Curricula, 1920–1986.” American Sociological Review 56 (1): 85–100. doi:10.2307/2095675.

Biggs, John. 1999. “What the Student Does: Teaching for Enhanced Learning.” Higher Education Research & Development 18 (1): 57–75. doi:10.1080/0729436990180105.
Bromley, Patricia, John W. Meyer, and Francisco O. Ramirez. 2011. “Student-Centeredness in Social Science Textbooks, 1970–2008: A Cross-National Study.” Social Forces 90 (2): 547–570. doi:10.1093/sf/sor004.

Carney, Stephen, Jeremy Rappleye, and Iveta Silova. 2012. “Between Faith and Science: World Culture Theory and Comparative Education.” Comparative Education Review 56 (3): 366–393. doi:10.1086/665708.

Carnoy, Martin, and Diana Rhoten. 2002. “The Varying Effects of Regional Organizations as Subjects of Globalization of Education.” Comparative Education Review 46 (1): 1–9. doi:10.1086/324053.

Dale, Roger. 1999. “Specifying Globalization Effects on National Policy: A Focus on the Mechanisms.” Journal of Education Policy 14 (1): 1–17. doi:10.1080/026809399286468.

Dale, Roger. 2000. “Globalization and Education: Demonstrating a ‘Common World Educational Culture’ or Locating a ‘Globally Structured Educational Agenda’?” Educational Theory 50 (4): 427–448. doi:10.1111/j.1741-5446.2000.00427.x.

Dale, Roger, and Susan L. Robertson. 2002. “The Varying Effects of Regional Organizations as Subjects of Globalization of Education.” Comparative Education Review 46 (1): 10–36. doi:10.1086/324052.

Grek, Sotiria. 2009. “Governing by Numbers: The PISA ‘Effect’ in Europe.” Journal of Education Policy 24 (1): 23–37. doi:10.1080/02680930802412669.

Gustafsson, Jan-Eric. 2008. “Effects of International Comparative Studies on Educational Quality on the Quality of Educational Research.” European Educational Research Journal 7 (1): 1–17. doi:10.2304/eerj.2008.7.1.1.

Gustafsson, Jan-Eric, and Monica Rosen. 2006. “The Dimensional Structure of Reading Assessment Tasks in the IEA Reading Literacy Study 1991 and the Progress in International Reading Literacy Study 2001.” Educational Research and Evaluation 12 (5): 445–468.

Karp, Alexander. 2006. “‘Universal Responsiveness’ or ‘Splendid Isolation?’ Episodes from the History of Mathematics Education in Russia.” Paedagogica Historica 42 (4–5): 615–628. doi:10.1080/00309230600806914.

Kiefer, Thomas, Alexander Robitzsch, and Margret Wu. 2015. TAM: Test Analysis Modules (R package version 1.10–0). http://cran.r-project.org/web/packages/TAM.

Kim, Seock-Ho, and Allan S. Cohen. 1998. “A Comparison of Linking and Concurrent Calibration under Item Response Theory.” Applied Psychological Measurement 22 (2): 131–143. doi:10.1177/01466216980222003.

Kim, Seock-Ho, and Allan S. Cohen. 2002. “A Comparison of Linking and Concurrent Calibration under the Graded Response Model.” Applied Psychological Measurement 26 (1): 25–41. doi:10.1177/0146621602026001002.

Kjaernsli, Marit, and Svein Lie. 2008. “Country Profiles of Scientific Competence in TIMSS 2003.” Educational Research and Evaluation 14 (1): 73–85. doi:10.1080/13803610801896588.

Kolen, Michael J., and Robert L. Brennan. 2004. Test Equating, Scaling, and Linking: Methods and Practices. 2nd ed. New York, NY: Springer.

Koretz, Daniel M. 2002. “Limitations in the Use of Achievement Tests as Measures of Educators’ Productivity.” The Journal of Human Resources 37 (4): 752–777. doi:10.2307/3069616.

Leung, Frederick K. S., and Yeping Li. 2010. Reforms and Issues in School Mathematics in East Asia: Sharing and Understanding Mathematics Education Policies and Practices. Boston, MA: Sense Publishers.

Lie, Svein, and Astrid Roe. 2003. “Unity and Diversity of Reading Literacy Profiles.” In Northern Lights on PISA. Unity and Diversity in the Nordic Countries in PISA 2000, edited by Svein Lie, Pirjo Linnakylä, and Astrid Roe, 147–158. Oslo: University of Oslo, Department of Teacher Education and School Development.

Lockheed, Marlaine E., and Hans Wagemaker. 2013. “International Large-Scale Assessments: Thermometers, Whips or Useful Policy Tools?” Research in Comparative and International Education 8 (3): 296–306. doi:10.2304/rcie.2013.8.3.296.

Martin, Michael, Kelvin D. Gregory, and Steven E. Stemler, eds. 2000. TIMSS 1999 Technical Report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, Michael, and Dana Kelly, eds. 1998. TIMSS Technical Report Volume II: Implementation and Analysis (Primary and Middle School Years). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Martin, Michael O., and Ina V. S. Mullis, eds. 2012. *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Meyer, John W., John Boli, George M. Thomas, and Francisco O. Ramirez. 1997. "World Society and the Nation-State." *American Journal of Sociology* 103 (1): 144–181. doi:10.1086/231174.

Meyer, John W., David Kamens, and Aaron Benavot. 1992. *School Knowledge for the Masses: World Models and National Primary Curricular Categories in the Twentieth Century*. London: Falmer Press.

Meyer, H. D., R. Strietholt, and D. Y. Epstein. 2018. "Three Models of Global Education Quality and the Emerging Democratic Deficit in Global Education Governance." In *International Handbook of Teacher Quality and Policy*, edited by M. Akiba and G. K. LeTendre, 132–150. New York: Routledge.

Olson, John F., Michael O. Martin, and Ina V. S. Mullis, eds. 2008. *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Pettersson, Daniel. 2008. "Internationell kunskapsbedömning som inslag i nationell styrning av skolan." [International Knowledge Assessments: An Element of National Educational Steering] PhD diss., Uppsala University.

Phillips, David, and Kimberly Ochs. 2003. "Processes of Policy Borrowing in Education: Some Explanatory and Analytical Devices." *Comparative Education* 39 (4): 451–461. doi:10.1080/0305006032000162020.

Ramaswamy, Venkatram, Wayne S. Desarbo, David J. Reibstein, and William T. Robinson. 1993. "An Empirical Pooling Approach for Estimating Marketing Mix Elasticities with PIMS Data." *Marketing Science* 12 (1): 103–124.

Rutkowski, Leslie, Eugenio Gonzalez, Marc Joncas, and Matthias von Davier. 2010. "International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting." *Educational Researcher* 39 (2): 142–151.

Rutkowski, Leslie, and David Rutkowski. 2009. "Trends in TIMSS Responses Over Time: Evidence of Global Forces in Education?" *Educational Research and Evaluation* 15 (2): 137–152. doi:10.1080/13803610902784352.

Schmittau, Jean, and Anne Morris. 2004. "The Development of Algebra in the Elementary Mathematics Curriculum of V.V. Davydov." *The Mathematics Educator* 8 (1): 60–87.

Schriewer, Jürgen. 2016. *World Culture Re-contextualised Meaning Constellations and Path-Dependences in Comparative and International Education Research*. London: Routledge. 186.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6: 461–464. doi:10.1214/aos/1176344136.

Steiner-Khamsi, Gita, and Ines Stolpe. 2004. "Decentralization and Recentralization Reform in Mongolia: Tracing the Swing of the Pendulum." *Comparative Education* 40 (1): 29–53. doi:10.1080/0305006042000184872.

Suárez, David, and Patricia Bromley. 2016. "Institutional Theories and Levels of Analysis: History, Diffusion, and Translation." In *World Culture Re-contextualised Meaning Constellations and Path-Dependences in Comparative and International Education Research*, edited by Jürgen Schriewer, 139–160. London: Routledge.

The Swedish National Agency for Education. 2008. *Med fokus på matematik och naturvetenskap. En jämförelse mellan TIMSS 2007 för årskurs 4 och de nationella målen för årskurs 5* [Math and Science in Focus: A Comparison between TIMSS 2007 Grade 4 and National Standards for Grade 5]. Stockholm: The Swedish National Agency for Education.
Wiseman, Alexander W., M. Fernanda Astiz, and David P. Baker. 2014. “Comparative Education Research Framed by neo-Institutional Theory: A Review of Diverse Approaches and Conflicting Assumptions.” Compare: A Journal of Comparative and International Education 44 (5): 688–709. doi:10.1080/03057925.2013.800783.

Wu, Margaret. 2010. “Comparing the Similarities and Differences of PISA 2003 and TIMSS.” doi:10.1787/5km4psnm13nx-en.

Technical Appendix: Linking the Subdomain Achievement Scales of Five Cycles of TIMSS

Tests design
To measure student proficiency in mathematics, the TIMSS tests in each cycle cover between 159 (1995) and 217 (2011) test items that are related to one of the four content areas numbers, algebra, geometry, and data and chance. The items that refer to proportionality and measurement were redistributed to the four aforementioned areas. The four content areas were assessed by at least 19 test items in each study cycle. To minimise testing burden of individual student, the assessment material was distributed into booklets and each student did not work on all but only some booklets (Rutkowski et al. 2010). The testing time for mathematics was 45 min in all cycles. TIMSS also administers a 45-minute-long science tests but science is not of interests in the present investigation.

The different study cycles used different tests but overlaps in the assessment material provide links between them. For each study cycle some items were integrated in the two subsequent study cycles and such overlaps exist for all four content areas. In the area data and chance, for example, seven items from 1995 were integrated into the test from 1999 and three items into the test from 2003. In addition, seven test items from 1999 were integrated into the test from 2003 and three items into the test from 2007. Such overlaps exist for all content areas and they serve as bridges between the tests of the different years. In the same vein, overlaps exist for the subsequent study cycles for all content areas.

Missing data
It is useful to distinguish between three types of missing data in assessment data: not-administrated, omitted, and not-reached items. We treated not-administrated items as missing data in estimating item and student parameters. The omitted items were treated as incorrect responses, because we did not want to reward students for skipping an item. Sometimes, however, students run out of time and do not complete the items at the end of the test. The not-reached items are typically treated as if they were not administrated when estimating the item parameters. However, they may be treated either as incorrect or as if they were not administrated when student proficiency scores are generated. The first approach punishes slow students. In the present study, we adopted the second alternative because the proportion of students who did not finish the tests varied between the study cycles and such differences may indicate that processing speed is a systematic source of variance in the different tests. Such differences limit the comparability of the tests over time (see Gustafsson and Rosen 2006). We avoid this issue by treating the not-reached items at the end of the tests as if they were not administrated.

Scaling of achievement data
To study trends in student achievement, we equated the respective tests. For this purpose, we made use of the overlaps in the assessment material of the respective study cycles. This was done separately for the content areas numbers, algebra, geometry, and data and chance. The equating had a common-item nonequivalent group design in which no text passage appeared in all study cycles but a set of common items appeared in multiple cycles (Kolen and Brennan 2004, see Figure 1).
Item response theory (IRT) models were used to link all tests onto the same metric by means of a concurrent calibration of all item parameters (Kim and Cohen 1998, 2002). Within each content area, we used the individual raw data on the item level from all five study cycles and then employed a one-parameter logistic IRT model with an extension for partial credit items for constructed response items, and a three-parameter logistic IRT model for multiple-choice items to estimate all item parameters in a concurrent calibration (see Kim and Cohen 2002). The R package TAM (Test Analysis Modules; Kiefer, Robitzsch, and Wu 2015) was used for the multiple group IRT analyses. The program uses the expectation-maximisation algorithm to achieve marginal maximum likelihood estimates of the item parameters. Latent normal distributions of student proficiency are assumed for each country in the respective study cycle. Then, we used these item parameter estimates to compute EAP (Expected A Posteriori) students’ ability scores. To simplify interpretation, the scores were standardised to metric with a mean of zero and a standard deviation of one.