# TIMSS 2015 and PISA 2015
# How are they related on the country level ?

Eckhard Klieme

German Institute for International Educational Research (DIPF)

DIPF Working Paper published online, December 12 2016
© Deutsches Institut für Internationale Pädagogische Forschung, 2016

International Large Scale Assessments have been introduced by IEA in the 1960s. By running the Trends in International Mathematics and Science Study (TIMSS) every four years since 1995, IEA has successfully measured changes in students' achievement levels spanning two decades, with six waves of measurement.

Starting in 2000, the OECD followed by running their Programme of International Student Assessment (PISA) every third year. In 2015, PISA also was implemented for the sixth time. For the second time in the history of international Large Scale Assessments (first in 2003, now in 2015), researchers and policy makers are faced with two „parallel" studies providing country-level achievement data. For the first time ever, trends over twelve years (2003 – 2015) are available from both study programs. Especially as these data have been published very closely in time (Nov. 29, 2016, for TIMSS and Dec 6, 2016, for PISA) policy makers are asking: **Are the messages on student achievement in international comparison and change over the years we receive from TIMSS and PISA consistent ? If there are any discrepancies, how can these be explained ? The present note intends to provide answers by studying country-level TIMSS and PISA results over the period from 2003 to 2015**. Data are taken from the international reports (Mullis, Martin, Foy & Hooper, 2016; OECD, 2016b).

Several scholars carefully studied similar questions with regard to the 2003 assessment year (see especially Hutchison & Schagen, 2007, and Wu, 2010). Our intention is to carry their work further on, using more recent data - including changes in student achievement between 2003 and 2015 - and new concepts – namely, Opportunity-to-learn – to explain any differences between the two studies.

## 1. Similarities and differences in study goals, conceptualization, and design

While PISA assesses 15-year old students (irrespective of the grade they are in), TIMSS serves three populations: grade 4, grade 8, and upper secondary students. PISA covers three to four domains in each wave of measurement (Reading, Mathematics and Science Literacy, plus Problem Solving or other Cross-Curricular Competencies), while TIMSS covers Mathematics and Science Achievement. **The following analyses are focused on Lower Seconday (TIMSS- Grade 8, PISA: 15-year old students) Mathematics**. The main reason for this choice is both substantive and methodological: (1) The domain of mathematics is generally perceived to be more coherent and more canonically defined across countries than the domain(s) of Science. (2) PISA selects a „major (focal) domain" for each wave of measurement, and comparing trends backwards is possible only with respect to the

first time when the respective domain has been studied as the major domain – which in 2003 was Mathematics, while Science wasn't fully established before 2006.

The TIMSS 2015 Assessment of Grade 8 Mathematics Achievement and the PISA 2015 Mathematics Literacy Assessment are both covering a broad array of student knowledge and understanding in lower secondary school mathematics (Mullis & Martin, 2013; OECD, 2016a), and they both cover a wide range of context variables (Hooper, Mullis & Martin, 2013; Kuger, Klieme, Jude & Kaplan, 2016), but there are a number of differences in study conceptualization and design. (For more details, see Hutchison & Schagen, 2007; Wu, 2010; and the respective Technical Reports.)

**Curriculum vs. Literacy.** The TIMSS test is based on comprehensive analysis of mathematics curricula worldwide, and it is supposed to be curricular valid across countries, i.e. to cover mathematical ideas and tasks that students have seen in classrooms. The PISA test is based on a more general concept of „life skills" that students are supposed to need in order to be ready for further learning, starting a successful vocational or professional career, and becoming an informed citizen. However, it is also informed by concepts of mathematical competencies that are shared and used to guide mathematics education worldwide. As a consequence, PISA test items tend to be more often embedded into real-word contexts and to provide lengthier text than typical TIMSS items. TIMSS administers more mathematics items than PISA, especially more short, multiple-choice items, it more often addresses knowledge on facts and procedures, and has a larger proportion of items on Numbers and Algebra as compared to Data and Uncertainty.

**Sampling: Grade-based selection of classrooms vs. age-based selection of students in schools.** TIMSS is designed to represent the population of students attending mathematics classrooms after 8 years of regular schooling. Randomly, lower secondary schools are sampled, and entire grade 8 mathematics classes are sampled within schools. Student age mostly ranges between 13 and 15 years. PISA is designed to represent the population of 15 year old boys and girls attending school. Randomly, lower secondary schools are sampled, and individual students are sampled within schools. Most PISA students are attending $8^{th}$, $9^{th}$ or $10^{th}$ grade. As a consequence, (a) PISA students are on average older than TIMSS students, (b) the difference between mean age in TIMSS and mean age in PISA varies depending on a country's rules for school entry and grade retention.

**Participating countries.** TIMSS serves a broad range of countries from all continents, including quite a few countries from Africa, Asia and the Middle East. PISA started as a survey for affluent, industrialized countries that are members of the OECD, but also attracts a number of participating countries and economies. In 2015, out of 35 OECD member States, all were participating in PISA, and 16 were participating in TIMSS-Grade 8 as well. All in all, 27 countries or systems participated in both assessments[1]. Back in 2003, the overlap was just 17 countries or systems.

**Mode of assessment and scaling.** In 2015, PISA was for the first time administered on computer in all but a few countries. TIMSS is going to introduce computer-based assessment in 2019. There are further technical differences such as details of the Item-Response-Theory approach used (e.g., TIMSS includes a „guessing" parameter for Multiple-Choice items which PISA doesn't, and PISA 2015 introduced a more comprehensive approach for linking scales to previous waves of the assessment).

[1] In addition, the United Kingdom participated in PISA 2015, but only England participated in TIMSS 2015. As we do not consider sub-national entities in our analysis, United Kingdom is not included here. Norwegian TIMSS data are reported for grade 8 throughout, although Norway reported grade 9 data as well in TIMSS 2015 (Mullis, Martin, Foy & Hooper, 2016).

These differences reflect partially different foci and goals of the studies: Both studies intend to inform policy makers and the public on the achieved level of student outcomes in mathematics. This is exactly why it makes sense to compare their findings on the country level. However, when researchers and policy makers wish to address more specific issues of educational policy and professional practice, the studies are rather complementary: As TIMSS is assessing entire classes and asking their teachers about curriculum and instruction, it is well prepared to study classroom-level, curriculum-based teaching and learning. PISA has a stronger focus on cross-curricular dispositions and skills, and school policies.

## 2. How are TIMSS 2015 and PISA 2015 related on the country level?

Arguably the most important finding from this report is: TIMSS and PISA provide similar pictures of student-achievement on the country level. There is a **close alignment between country mean scores from both studies** (Figure 1). The coefficient of correlation is .923, indicating that 85 % of the between-country-variation in PISA Mathematics Literacy can be explained by TIMSS, and vice versa.

It is worth noting that Science scores are equally well aligned on the country level: The coefficient of correlation is .926, accounting for 86 % of between-country variance.

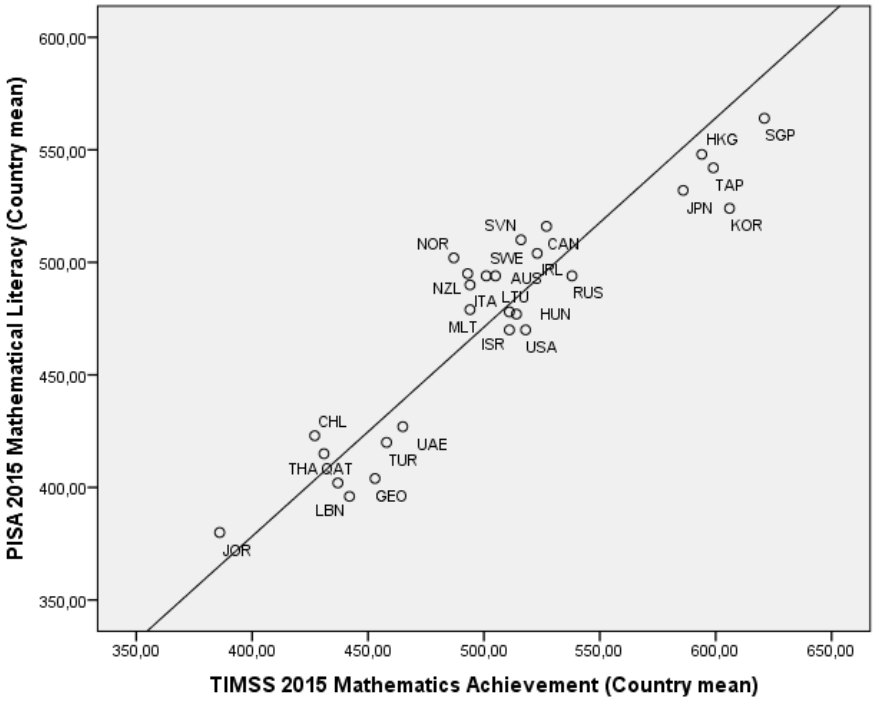


Figure 1: Relationship between country mean scores for TIMSS 2015 (Mathematical Achievement) and PISA 2015 (Mathematics Literacy). The straight line illustrates the linear regression.

The two-dimensional layout of Figure 1 helps identify a pattern that would not be perceived as easily using just one of the studies: East Asian countries (including OECD-members Japan and Korea) on the upper end, countries from yet developing regions like Near and Middle East (including OECD-members Turkey and Chile) on the lower end are forming clusters with similar profiles of student achievement in TIMSS and PISA, while European OECD-members, English-speaking countries, Russia and Lithuania belong to the central cluster. This pattern would of course look different if an even more diverse set of countries would implement both TIMSS and PISA, but basically this pattern can

be found in many international Large Scale Assessments. This includes some minor, but typical **deviations** from the overall linear relationship: The top-achieving East Asian systems seem to do a little better in TIMSS Mathematics than you would expect from their PISA results, while some Nordic and English-speaking countries (Norway, Sweden, Australia, Canada, Ireland, and New Zealand) are doing a little better in PISA. This is in line with the pattern that previous research has found in 2003.

## 3. How has the relationship between TIMSS and PISA evolved since 2003?

In 2003, the TIMSS grade 8 mathematics scores and the PISA Mathematics Literacy scores were also highly correlated. Based on the 17 countries which administered both tests, the coefficient was .867. As can be seen in Figure 2, the correlation back than was partly driven by two outliers, Indonesia and Tunisia. However, if these are dropped, the coefficient is still .717, statistically significant ($p<.01$). It is worth noticing that the pattern of deviations from the linear regression line is much like in 2015: East Asian countries doing a little better in TIMSS, Nordic and English-speaking countries doing a little better in PISA.

We also compared two pairs of TIMSS/PISA assessments which were administered one year apart: Mathematics scores correlate at .931 for TIMSS 2007 and PISA 2006 (25 countries), and .944 for TIMSS 2011 and PISA 2012 (28 countries). **Obviously, the close alignment between TIMSS and PISA on the country level is not a new phenomenon. Their messages on overall student attainment across educational systems are quite similar.** From 2003 to 2006/07 and 2011/12 the correlation even increased, while in 2015 it dropped slightly[2]. Nevertheless, there are some discrepancies, and the next section will discuss how to explain them.

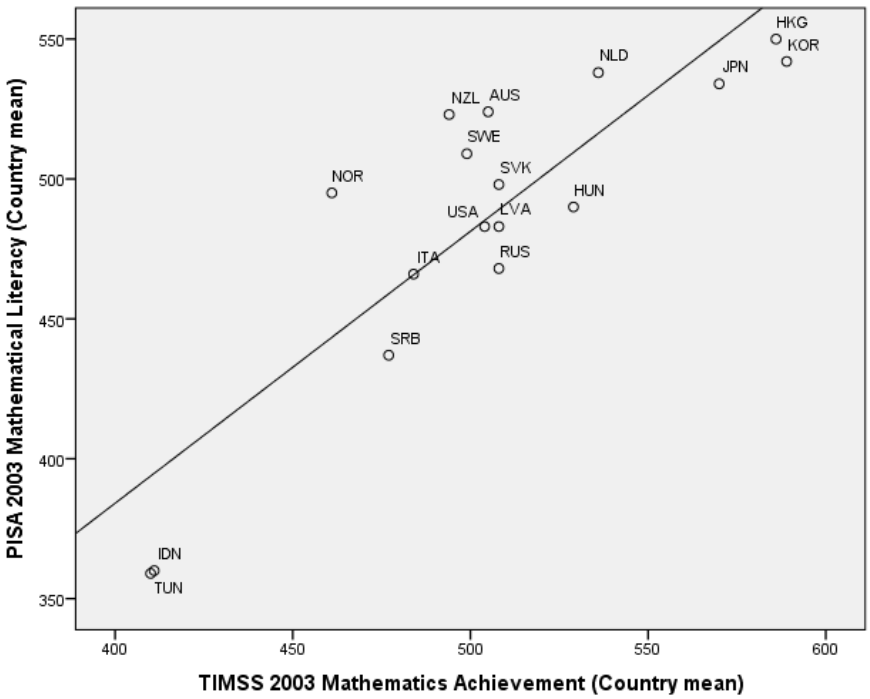


Figure 2: Relationship between country mean scores for TIMSS 2003 (Mathematical Achievement) and PISA 2003 (Mathematics Literacy). The straight line illustrates the linear regression.

[2] This is also true if correlations are computed across countries who participated in all four comparisons.

## 4. How can differences be explained?
## Opportunity to learn as a specific focus of TIMSS

In order to explain any discrepancies between TIMSS and PISA on the country level, two distinct approaches may be used: (a) Accounting for differences in assessment design. (b) Looking for features of educational systems that may explain the difference. While previous research has focused on the former approach, this report is targeting the latter, aiming at policy-relevant, non-technical explanations.

As explained in section 1 of this paper, differences in sampling (grade-based vs. age-based) and differences in facets of mathematics being covered are the most important factors distinguishing the TIMSS assessment design from the PISA assessment design. In order to take the first factor in account, Wu (2010) used the index of "mean student age" in TIMSS: the older TIMSS participants within a certain country are on average, the more similar they are to the PISA sample. In order to take the second factor into account, Wu (2010) developed an index of "content advantage" for each country. The index estimates how country results in TIMSS would change if content areas within mathematics (such as Number, Algebra, or Data and Uncertainty) would have contributed the same share of items in the TIMSS test as they did in PISA. Both indices – mean student age in TIMSS, and content advantage - were used to predict country-level PISA scores. While TIMSS 2003 scores alone accounted for 71 % of between-country variance in PISA 2003 math scores[3], adding the two indices allowed Wu (2010) to explain 93 % of the variance. The conclusion was: **Differences in student sampling, plus differences in test content account for most of the discrepancies between TIMSS and PISA math scores observed on the country level**.

However, from a policy point of view, the explanation of discrepancies by features of the assessment design is of little help. Isn't there any difference in what educational policy and research can learn from TIMSS and PISA mean scores apart from "technical" differences in study design?

In the following, we are studying a factor that is closer to the teaching and learning students experience in classrooms: **Opportunity to learn (OTL)**. OTL has been studied extensively in IEA studies and shown to be an important factor explaining differences in student outcomes (e.g., Burstein et al., 1993; Schmidt & Maier, 2009). The more (and deeper) content students are exposed to, the better their results in Large Scale Assessments. The hypothesis tested here is the following: **As TIMSS is focused on curricular content, it should convey information on OTL on top of the general level of mathematical competencies that is assessed in PISA.** Thus, it can be expected that the small discrepancies left between TIMSS grade 8 and PISA mathematics scores can at least partly be explained by students' opportunity to learn mathematical content.

In 2011/2012, both TIMSs and PISA (which at this time was focused on mathematics as its major domain) included measures of Opportunity to learn:

- TIMSS asked teachers to judge to what extent their students had been taught core curriculum elements. Altogether, the survey covered 19 topics from the areas of Numbers, Algebra, Geometry, Data and Chance. TIMSS 2011 reported country-level indicators for „Percentage of Students taught the TIMSS Mathematics topics".

[3] In addition to the 17 countries covered in our analysis (see Figure 2), Wu included 5 sub-national regions; therefore her figures slightly differ from ours.

- PISA asked students to judge their familiarity with mathematical concepts. There was a list of 13 mathematical terms like „exponential function“ or „arithmetic mean“. An overall score of familiarity with mathematical concepts was developed and aggregated on country-level.[4]

We used these two measures of country level Opportunity to learn (OTL), in addition to country level PISA 2012 math scores, to explain country-level TIMSS 2011 math scores. As Table 1 shows, adding a measure of OTL allows for a small, but significant increase in explanatory power: Instead of 88 % of the between-country-variation in TIMSS-Scores, we are able to explain 3 % more by adding „Familiarity with math concepts“, and nearly 6% more adding „TIMSS mathematics topics taught“.

| | Model using „Percentage of Students taught the TIMSS mathematics topics“ as the measure of OTL | Model using „Familiarity with math concepts“ as the measure of OTL |
|---|---|---|
| Effect of PISA 2012 score | .946 *** | .768 *** |
| Effect of OTL measure | .155 ** | .237* |
| $R^2$ (variance accounted for) | .938 | .925 |
| $R^2$ (prediction by PISA score only) | .891 | |

Table 1: Explaining country-level math scores in TIMSS 2011

Assuming Opportunity to learn is quite stable on the country level, we may use these measures to explain TIMSS 2003 and TIMM 2015 results as well. In fact, using either OTL measure as a predictor in addition to the respective PISA score allows for significantly better explanation of TIMSS scores in all waves of the studies. For 2015, both OTL measures can be combined in a single analysis, where each of them has a significant contribution. Thus, the proportion of TIMSS between-country-variance accounted for is increasing from 85 % (using PISA scores as the only predictor) to 96 % (using both OTL measures on top).

If we add mean age of students participating in TIMSS 2015 as a fourth predictor – accounting for the differences in sampling -, all four predictors significantly contribute to explaining country-level TIMSS scores, and overall they account for 97.4 % of the variance – even more than Wu (2010) reports in her analysis of the 2003 data.

It is important to note that in our analysis the bulk of additional explanatory power (explaining what PISA scores cannot account for) comes from measures of OTL rather than “technical” features of the assessment design. These analyses clearly show that **TIMSS scores – although being closely related to PISA scores on the country level – carry additional information related to the quality of the mathematics curriculum implemented in classrooms**.

[4] Three „foils“, i.e. concepts that in fact are not established in mathematics, were added to the list. These were used to correct for guessing and response bias (Kyllonen & Bertling, 2014).

## 5. Do both studies agree on the amount of change between 2003 and 2015 ?

Finally, we would like to know if TIMSS and PISA provide coherent views on the change of mathematical achievement levels between 2003 and 2015. To this end, we calculated the difference between TIMSS 2015 scores and TIMSS 2003 scores for each country participating in both studies. Similarly, we calculated the change in country level means for the 2015 and 2003 PISA Mathematics Literacy assessments.

Both TIMSS and PISA change scores are available for 11 countries (see Figure 3). The change scores correlate substantially (r=.612, p<.05, explaining 37 % of the variance). This result cross-validates the trend analyses from both studies: **The assessment of changes in achievement level based on TIMSS is supported by similar findings from PISA, and vice versa.** However, the alignment is weaker compared to the cross-sectional findings reported above (sections 2 and 3). Clearly, change scores are less reliable than cross-sectional country means, as there is error from both waves plus the so-called linking error (Mazzeo & von Davier, 2014).

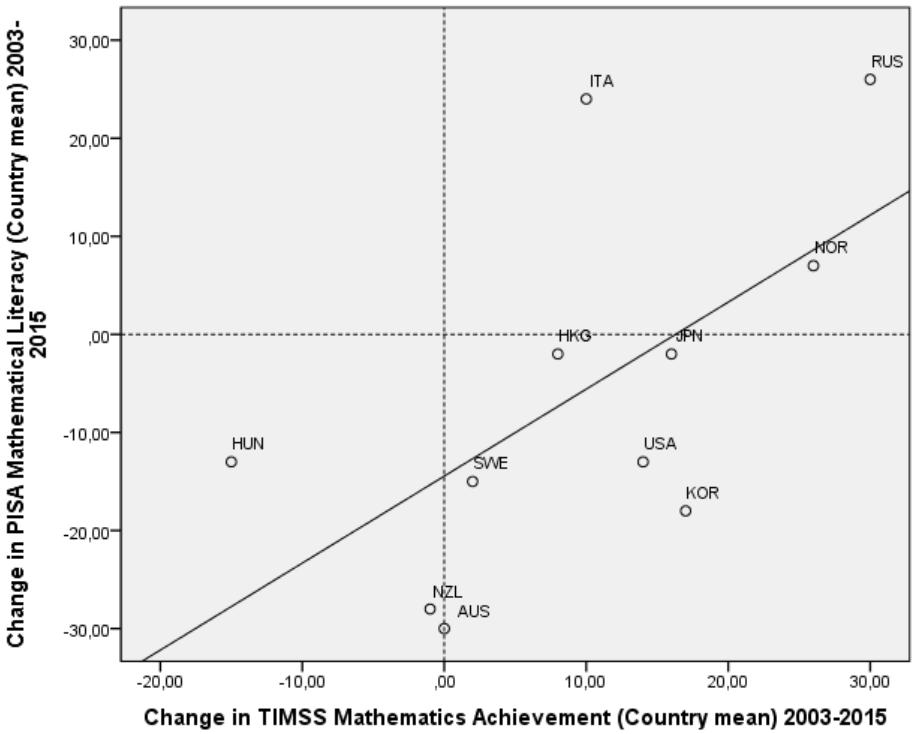


Figure 3: Relationship between change in country mean scores between 2003 and 2015 for TIMSS (Mathematics Achievement) and PISA (Mathematics Literacy). The straight line illustrates the linear regression.

Thus, countries doing **relatively** better on TIMSS tend to do so in PISA as well. Nevertheless, there may be discrepancies in **absolute change**. It is worth noticing that change scores illustrated in Figure 3 are **mostly positive for TIMSS, and mostly negative for PISA**. Until 2011/2012, both TIMSS- and PISA-scores showed a balanced mix of gains and losses. Recently however, i.e. for the time interval 2011/12 – 2015, change scores had opposite directions for quite a few countries, as visualized in the lower right quadrant of Figure 4. In Singapore as well as in the US, TIMSS-scores significantly increased while PISA-scores significantly decreased at the same time. Among 22 countries that implemented both studies in 2011/2012 as well as in 2015, only 2 had a significant negative change

score in TIMSS, but 7 had a negative change score in PISA. Since similar patterns exist for Science[5], it seems unlikely that the discrepancies can be attributed to profiles in OTL or curriculum reform. Rather, they might be related to the new mode of assessment in PISA 2015. Just one of these 22 countries, Jordan, kept testing on paper in PISA 2015, and this is exactly the outlier showing significant loss in TIMSS and no significant change in PISA. For one country not included here, Germany, a negative impact of computer-based assessment on PISA 2015 science and mathematics scores has been established (Robitzsch et al., 2016), while OECD (2016b) found no general mode effect on the international level.

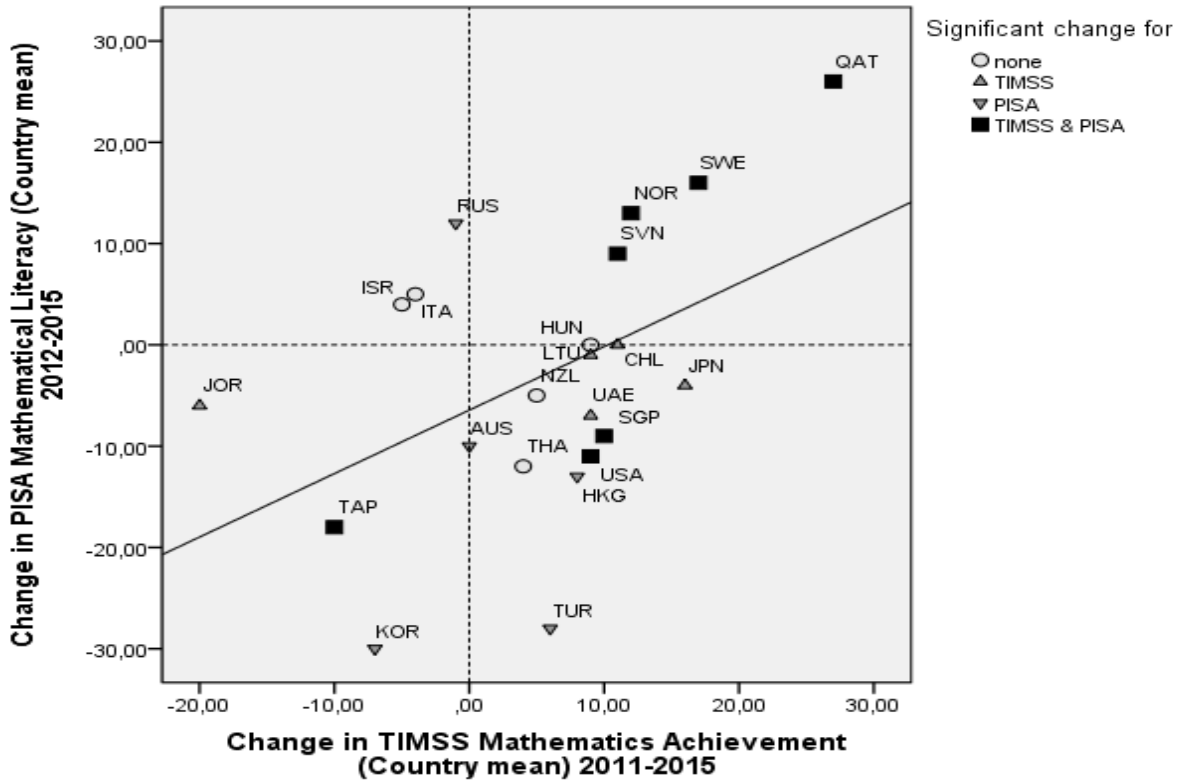


Figure 4: Relationship between change in country mean scores between 2011 and 2015 for TIMSS (Mathematics Achievement) and between 2012 and 2015 for PISA (Mathematics Literacy). The straight line illustrates the linear regression. Results of within-country significance testing are taken up from the respective international study report.

In sum, **change scores on the country level are less robust, and the reasons for these changes are still not well understood in research. Comparing change and trends measured by two different studies (TIMSS and PISA) provides additional insights, e.g. on potential mode effects.**

[5] For Science, out of these 22 countries, Hong Kong, Lithuania, Turkey and United Arab Emirates show significant losses in PISA and gains in TIMSS. Overall there are 9 significant losses in PISA and 2 in TIMSS.

## 6. Conclusions

Comparing findings from TIMSS and PISA on the country level, this report found support for the following statements:

**1) On the country level, mean scores from TIMSS-Grade 8 and PISA are closely related. This validates both indicators of overall achievement in mathematics.**

**2) In addition to the general level of mathematical competence, as measured by PISA, country-level mean scores in TIMSS reflect the quality of mathematics curricula implemented in classrooms. TIMSS-scores are sensitive to Opportunities-to-learn provided in a country.**

**3) Measuring change in student achievement on the country level is less robust than measuring student achievement in any single wave of assessment. More methodological and educational research is needed to understand trends on the country level, including the discrepancies between TIMSS and PISA and potential mode effects.**

Finally, It needs to be understood that the correlations reported here are based on maximum 27 "observations" (countries), each representing highly aggregated test data from thousands of students. It is a well known phenomenon in educational research that correlations become stronger the more they are aggregated. If the correlation would have been estimated on the student level – allowing the same students to take both tests and comparing their scores, which to our knowledge no researcher has done so far[6] -, the correlation between TIMSS and PISA would most probably be lower, because these assessments do in fact measure different facets of mathematical competence, as explained in section 1. When aggregated on the country level, the (average) test scores no longer represent any student's ability to solve specific kinds of math problems. Rather, the aggregated score is an indicator of a country's overall efficiency in promoting mathematical competence among its children and youth. Thus, the high correlation on country level is **strong evidence for the validity of both studies as indicators of country performance in mathematics,** but this should **not be mixed up with the validity of either test** when, e.g., evaluating the impact of teaching reforms on learning outcomes, or evaluating the impact of socio-economic background on student literacy. Most probably, TIMSS would be better answering the first question, while PISA would be better answering the second question. Both studies have specific merits, and countries can profit from implementing both of them.

## Authors Address

Professor Dr. Eckhard Klieme
Deutsches Institut für Internationale Pädagogische Forschung
Schloßstraße 29, D - 60486 Frankfurt/Main
klieme@dipf.de

DIPF
Educational Research and Educational Information

[6] Klieme, Neubrand & Lüdtke (2001), enhancing the PISA 2000 design for Germany, had students work on selected TIMSS items in addition to the PISA mathematics test. Both PISA and TIMSS items fit into a unidimensional Rasch scale, suggesting that PISA and TIMSS measure similar competencies on the individual level as well. However, neither PISA maths literacy nor the TIMSS mathematics test were fully represented.

## References


Burstein,L. (Ed.) (1993). *The IEA Study of Mathematics III: Student Growth and Classroom Processes.* Oxford: Pergamin Press.

Hooper,, M., Mullis, I. & Martin, M. (2013). TIMSS 2015 Context Questionnaire Framework. In Mullis & Martin (2013), pp. 61-82.

Hutchison, D. & Schagen, L. (2007). Comparisons between PISA and TIMSS – Are We the Man with Two watches?. In T. Loveless (Ed.), Lessons Learned. What International Assessments tell us about Math Achievement. Washington: The Brookings Institution, pp. 227-262.

Klieme, E., Neubrand, M. & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider,. . . M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 139–190). Opladen: Leske + Budrich.

Kuger, S. , Klieme, E., Jude,N. & Kaplan, D. (Eds.) (2016). *Assessing Contexts of Learning.* Dordrecht: Springer.

Kyllonen,P. & Bertling, J. (2014). Innovative Qusetionnaire Assessment Methods to Increase Cross-Country Comparability. In L. Rutkowski, M.v. Davier & D. Rutkowski (Eds.), *Handbook of International Large Scale Assessment.* Baco Raton, FL: CRC Press, pp. 277-286.

Mazzeo, J. & von Davier, M. (2014). Linking Scales in International Large-Scale Assessments. In L. Rutkowski, M.v. Davier & D. Rutkowski (Eds.), *Handbook of International Large Scale Assessment.* Baco Raton, FL: CRC Press, pp. 229-257.

Mullis, I.V.S. & Martin, M.O. (Eds.). (2013). *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* Retrieved from Boston College, TIMSS & PIRLS International Student Center website: http://timssandpirls.bc.edu/timss2015/international-results.

OECD (2014). *PISA 2012 Technical Report*. Paris: OECD.

OECD (2016a). *PISA 2015 Assessment and Analytical Framework*. Paris: OECD.

OECD (2016b). *PISA 2015 Results (Volume I). Excellence and Equity in Education*. Paris: OECD.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F. & Heine, J.-H. (2016). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica*, 63, Advance online publication. DOI: 10.1026/0012-1924/a000177

Schmidt, W.H. and A. Maier (2009). Opportunity to Learn. In G. Sykes, B. Schneider and D.N. Plank (Eds.), *Handbook of Education Policy Research.* New York: Routledge, pp. 541-559.

Wu, M. (2010). Comparing the Similarities and Differences of PISA 2003 and TIMSS. *OECD Education Working Papers*, No. 32. Paris: OECD. http://dx.doi.org/10.1787/5km4psnm13nx-en