International Journal of Educational Development 45 (2015) 77–88

Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

Reading achievement progress across countries

Maciej Jakubowski [a], Artur Pokropek [b,*]

[a] Faculty of Economic Sciences, Warsaw University, Poland
[b] Educational Research Institute (IBE), Poland

ARTICLE INFO

Article history:
Received 5 July 2014
Received in revised form 12 September 2015
Accepted 22 September 2015

Keywords:
PISA
PIRLS
Value added
Human capital
Early selection
Tracking

ABSTRACT

This paper discusses a method to compare progress in reading achievement across countries. The method uses measures of achievement in primary schools from PIRLS and compares them to secondary school results from PISA. Results describe an interval in which estimates of progress can lie, depending on the comparability of these two assessments. Progress estimates are also adjusted for age, gender and other characteristics that differ between countries and surveys. The paper also demonstrates how useful these estimates are for policy analysis by comparing achievement progress across early and late tracking countries.

1. Introduction

International surveys of students, such as the Progress in International Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA), assess representative samples of students from different countries to provide estimates of their average level of skills and knowledge related to reading competencies (von Davier et al., 2013). Country rankings produced by these surveys usually attract considerable attention, while more in-depth analyses of the factors that may influence these results are discussed less often (Grek, 2009). Although countries can compare their students' skills levels to those of other participating countries, cross-sectional surveys like PIRLS or PISA provide limited guidance to policy makers. Average country performance is only partly affected by teaching quality; between-country differences in such factors as parents' education, a country's economic and social development, or school enrolment levels usually play important roles in defining student outcomes (Fuchs and Wößmann, 2008).

To assess country's performance at the secondary level it seems useful to take into account primary school performance. Country-specific factors, like economic prosperity, teacher's professionalism, culture and social capital, parents' educational attainment or school resources tend to be similar and similarly affect students at both primary and secondary levels. Other factors, for example tracking policies or changes in curriculum, might affect students at different ages differently. Thus, by adjusting secondary school results for primary school achievement, or by looking at the achievement progress between primary and secondary level, one can see how the factors that vary between primary and secondary school affect performance. These factors will be more driven by policy changes or the effectiveness of secondary schools than by country-specific characteristics that can be rarely changed by policies, at least in the short-term. This idea is close to difference-in-differences method often used in cross-country analyses (see Hanushek and Wößmann, 2006, for an application in education which is extended in this paper). It is also similar to longitudinal studies that compare student performance over time for some organizational units, for example, across US states or across the same schools (see Card and Payne, 2002, for an analysis of US states, and Hoxby, 2000, for an analysis of schools). In each case, the idea is to exclude factors that affect both data points similarly and to relate outcome changes to factors and policies that vary between the two data points. Although such approach has its limitations, which we discuss below, there is no doubt that it has a potential to provide more useful secondary schools policy indicators than average scores collected at one moment in time.

To enable this type of comparison, it is first necessary to calculate internationally comparable reading achievement scores for primary

* Corresponding author.
E-mail addresses: mjakubowski@uw.edu.pl (M. Jakubowski), artur.pokropek@gmail.com (A. Pokropek).

http://dx.doi.org/10.1016/j.ijedudev.2015.09.011
0738-0593/

and secondary schools and then to report point and interval estimates of reading achievement progress or development across countries, which we attempt to do in the first part of the paper. These data enable answering different types of research questions, and in this paper we focus on three. First we compare how achievement progress of boys and girls varies across countries. Second we compare how performance gaps develop differently. Finally, we present an example of analysis comparing achievement progress between countries that track students very early (usually by segregating them into academic and vocational tracks around the age of 10 or 12) and those who have the same curriculum for all students until the age of 15 or 16. Our approach combines several known statistical and psychometric methods but to our knowledge is the first time that they are applied to international comparisons. Thus, the final goal of this paper is to present this novel methodology.

Although some international comparisons of achievement between different grades are already available for mathematics and science, e.g. 4th and 8th grade results in TIMSS (Trends in International Mathematics and Science Study), our study differs in three aspects. First, we present reliable international comparisons of achievement progress in reading literacy, which were previously unavailable. Second, we calculate differences for a longer period of time, between the age of 10 and 15, or between the 4th and 9th grades. This opens a way, for example, to the analysis of tracking policy, which we present in this paper. Finally, we propose a way to correct point estimates and a novel approach to estimate complex standard errors that account for all sources of errors when comparing results from two international student surveys.

Previous studies used two strategies in the attempt to increase comparability. The first was to adjust and compare country-level statistics. For example, Brown et al. (2007) report a similarity in country rankings obtained from studies like PISA, PIRLS, but also TIMSS and IALS (The International Adult Literacy Survey). They note, however, that although measures of central tendency provide a relatively consistent picture, that is not the case with measures of dispersion, suggesting that differences between surveys, such as different scaling models, might limit possible comparisons. Rindermann (2007) and Jakubowski (2010) compare country-level statistics and both suggest that results of PIRLS and PISA are more consistent after adjusting for differences in average student age. In the second approach, student results are compared within groups of students that share similar characteristics. For example, Carnoy and Rothstein (2013) compare PISA and TIMSS results within student groups defined by social classes. This enables them to adjust for the large share of disadvantaged students in the U.S. and obtain a more comparable picture of performance across countries.

We re-estimate student performance using micro-data from both PIRLS and PISA assessments. To our knowledge, this is the first attempt to apply the same psychometric approach to data from different studies in order to increase their comparability. Although authors comparing PIRLS or TIMSS with PISA usually mention differences in measurement models which might affect the comparisons, these issues were never fully addressed. In a typical approach authors simply re-standardize country averages or other statistics to put them on the same scale without properly addressing complex issues of scale comparability (see for example Hanushek and Wößmann, 2006). This motivates our approach to re-estimate student performance from the micro-data using the same measurement model for both PIRLS and PISA and to estimate errors that arise when comparing different achievement surveys.

It should be noted, however, that we can only account for errors that can be corrected or estimated with the available data. For example, our results might be affected by survey response, drop-out or enrolment rates. In fact, studies reported that these factors might mildly affect comparisons, while they do not change country rankings or analytical results based on larger groups of countries.[1] Anyhow, the results for countries with low response, high exclusion, or high drop-out rates should be taken with more caution. Performance comparisons between primary and secondary schools might be even more sensitive to these factors as enrolment is usually nearly universal in primary schools while it might decrease at the secondary level for some countries. We provide relevant data in Table A6 in the Appendix to inform readers about the coverage of samples in each country. In general, we do not find any correlation between the population coverage of 15-year-olds as reported by PISA and our measures of achievement progress. Thus, while some caution is in order when interpreting results for countries with smaller coverage rates, we do not find evidence that these factors affect our results. Our analysis also suggests that more effort is necessary to provide data that allow reliable comparisons across countries, especially for developing countries where the participation rates in large-scale assessments are usually lower.

We compare reading achievement in primary school, as measured by PIRLS, to reading achievement of 15-year-olds in the PISA survey. The results of PIRLS 2001 are compared with results from PISA 2000, while the results from PIRLS 2006 are compared to results from PISA 2009. While PIRLS is entirely devoted to assessing reading achievement, only PISA 2000 and PISA 2009 focused on reading and provide the most reliable comparison with PIRLS. Thus, to increase reliability we decided to compare surveys that focus on reading achievement. We also compare surveys that were administered more or less at the same time. This is due to three reasons. The first one we already discussed and is related to the lack of reliable reading assessments that would follow the same cohort of students over time.[2] Secondly, taking surveys from different moments in time would increase the probability that outcomes are affected by factors external to the education system, for example, economic crisis, migration or political changes that might happen over a longer period of time. Finally, comparing student outcomes adjacent in time allows us to exclude the effect of policies similarly affecting students at different grades. Otherwise, it would be difficult to separate policies that differently affect primary and secondary school students from those that affect all students similarly but could change between assessments. Thus, our results do not represent progress of individual students over time but rather learning opportunities between primary and secondary level based on outcomes of students attending schools at the same time.

Achievement is compared using random draws of test items from both surveys giving a range of results possible to obtain under differently constructed tests. Thus, our results describe an interval in which estimates of progress can lie, depending on the comparability of the two assessments. We also adjust progress estimates for differences in the distribution of student background characteristics and for differences in testing age across countries and surveys. The results are precise enough to compare changes in achievement across countries, even after taking into account the fact that different combinations of test items would give different estimates of progress. Results are provided for all students and for subpopulations defined by gender, immigrant status, and

[1] Hanushek and Wößmann (2011) reported correlation between response and enrolment rates and outcomes of various international surveys of achievement. On the other hand, Micklewright and Schnepf (2012) analyse non-response for PISA in the UK to report that the relationship between performance bias and response rates is not straightforward with both positive and negative biases possible depending on the particular response pattern in a country.

[2] For example, one could compare PIRLS 2001 to PISA 2006 which covered similar cohort of students, but as PISA 2006 has focused on science rather than reading the comparisons would be much less reliable.