



# Measurement: Interdisciplinary Research and Perspectives

ISSN: 1536-6367 (Print) 1536-6359 (Online) Journal homepage: <http://www.tandfonline.com/loi/hmes20>

## Linking Large-Scale Reading Assessments: Comment

Eric A. Hanushek

To cite this article: Eric A. Hanushek (2016) Linking Large-Scale Reading Assessments: Comment, Measurement: Interdisciplinary Research and Perspectives, 14:1, 27-29, DOI: [10.1080/15366367.2015.1115298](https://doi.org/10.1080/15366367.2015.1115298)

To link to this article: <http://dx.doi.org/10.1080/15366367.2015.1115298>



Published online: 19 Feb 2016.



Submit your article to this journal [↗](#)



Article views: 18



View related articles [↗](#)



View Crossmark data [↗](#)

Full Terms & Conditions of access and use can be found at  
<http://www.tandfonline.com/action/journalInformation?journalCode=hmes20>

COMMENTARIES

## Linking Large-Scale Reading Assessments: Comment

Eric A. Hanushek

Hoover Institution, Stanford University

Applied researchers in education have only recently begun to appreciate the value of international assessments, even though now we have 50 years of experience with these. This slow start may reflect both a broadening of the set of countries that participate and the move to more regular cycles of testing. In any event, there has been a surge in analyses that focus on international comparative studies.<sup>1</sup>

Until recently, these assessments have been stand-alone surveys that have not been linked, and analysis has largely focused on cross-sectional uses of the international data. While there has been no move toward following individuals over time in the main surveys, both the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD) have recently begun equating scores across assessment.<sup>2</sup> This linking of examinations permits some longitudinal analysis of country changes, albeit on fairly short time dimensions.<sup>3</sup>

Striethold and Rosén consider a different approach whereby they can splice together reading assessments over a much longer period. They essentially attempt to mimic psychometric linking using a subset of items for each of the IEA reading tests since 1970. The purpose is to provide data on longer term trends that permit consideration of how educational policy affects student outcomes over extended periods of time.

As an economist, I am unprepared to discuss the details of their item response theory (IRT) scaling procedure. However, since Ludger Woessmann and I have been heavily invested in some parallel analyses,<sup>4</sup> it is useful to consider the outcomes of their procedure and how they compare to our approach.

Striethold and Rosén provide a picture of the primary school reading performance for four countries—Hungary, Italy, Sweden, and the United States—for the period 1970–2011. They provide estimates of the reading performance across 6 testing opportunities, including 2 parallel tests for 1991. These countries turn out to be the only ones that fully participated in all of the international reading assessments of the IEA.

What can we gain from such a linking of tests? Presumably, the new information allows comparing trends in student outcomes across countries. It may be possible, for example, to suggest that certain institutional features of a country leads to the corresponding trends. But with 5 observations for the 4 countries, researchers will likely find it difficult to convince anybody that they have identified any causal influence of institutional factors except in rather special circumstances.

One kind of analysis that might be convincing would be finding some major change in a country's policies during the time period so that the intertemporal pattern of achievement in that country could be directly contrasted with 1 or more of the other countries that had stable policies over the time period. Thus the stable countries would be the control group for the one with policy change, and it would be possible to do a simple difference-in-differences analysis; for example, the results of the 2-decade experiment of Sweden with deregulated schooling markets might be contrasted with one of the other 3 countries.

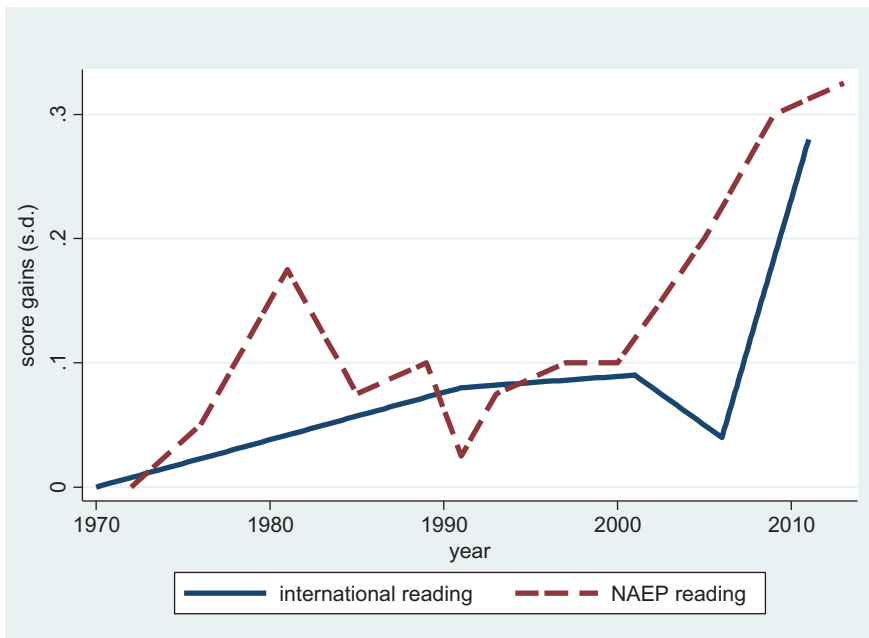


Figure 1. Comparison of international time series with NAEP time series.

For such a procedure, however, one must be quite confident of the reliability of the time pattern of the test linkages that Striethold and Rosén provide. One possible way to check on this is to compare their results to actual time-series data on achievement. Such a comparison is possible for the United States, because it has given its own longitudinally equated test to primary students since 1971. This test—the National Assessment of Educational Progress (NAEP)—gives a benchmark to compare the IRT results from the international reading tests.

Figure 1 shows time series from NAEP and from the Striethold and Rosén estimation.<sup>5</sup> (The solid line for international reading results comes from the IRT scaling of Striethold and Rosén.) Without conducting any formal tests, it is clear that the end points are very similar but that the time pattern in the intervening years is noticeably different; but it is just that intervening time pattern that would be crucial for the kind of difference-in-difference analysis suggested.

Woessmann and I have taken a very different tack in trying to use the full set of international assessments in our analysis of economic growth (Hanushek and Woessmann, 2015). We relied on the fact that the United States had participated in all of the IEA and OECD math, science, and reading assessments and that the NAEP testing programs provided longitudinal linkages of performance for U.S. students. We then could compare the performance of all international participants for each subject-grade-year test to that of suitably normalized U.S. performance according to NAEP. This procedure uses simple calibration instead of IRT linkages, but it does provide a much broader set of comparison countries than is possible under the Striethold and Rosén procedure, since comparison countries do not have to participate in all of the available tests.

Interestingly, when we decompose the variation in country scores, we find that 73% of the variation in measured test scores reflects overall country differences as opposed to measurement errors in the tests or to changes in country performance over time. We do, nonetheless, find that changes in test scores over time for the 15 countries observed from before 1984 through 2003 are an important determinant of changes in national economic growth rates. Thus, the time-series information is separately important.

While I agree with the motivation of Striethold and Rosén, I am left with some concerns about the reliability of their approach for international empirical analysis. Moreover, the limited samples they produce do not lend themselves to a wide range of analyses. On the other hand, they could perhaps expand their samples if they relax the requirement that countries participate in all of the separate tests.

## Notes

1. For a review of economic related studies, see Hanushek and Woessmann (2011).
2. Some countries (e.g., Canada and Australia) have moved toward expanding their national samples in the PISA (Programme for International Student Assessment) and using that to develop panel data for individuals, but this is not general across the participating countries.
3. Examples of analyses that use the country-level panel features include Brunello and Rocco (2013), who investigate the impact of immigration, and Hanushek, Link, and Woessmann (2013), who study school autonomy.
4. Our principal focus has been on how achievement is related to the economic growth of nations, although we also consider the impact on achievement of a variety of institutional differences in school systems (Hanushek & Woessmann, 2015).
5. The NAEP data are for 9 year olds, or roughly grade 3 students. Thus they are shifted 1 year forward to give the year when they would be in grade 4 (the international testing grade). Both series are normalized to give individual-level standard deviation away from the initial observation that is set equal to 0. NAEP scores can be found at <http://nces.ed.gov/nationsreportcard/>.

## References

- Brunello, G., & Rocco, L. (2013). The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores. *Economics of Education Review*, 32(1), 234–246. doi:10.1016/j.econedurev.2012.10.006
- Hanushek, E. A., Link, S., & Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics*, 104(September), 212–232. doi:10.1016/j.jdeveco.2012.08.002
- Hanushek, E. A., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Amsterdam, Netherlands: North Holland.
- Hanushek, E. A., & Woessmann, L. (2015). *The knowledge capital of nations: Education and the economics of growth*. Cambridge, MA: MIT Press.