Review of: "Facing the Facts About Test Score Gaps"

Leslie Rutkowski[1]

1 Indiana University

Potential competing interests: No potential competing interests to declare.

By way of introduction, I am a professor of quantitative methodology at Indiana University. My research is in the area of international large-scale assessment (ILSA; e.g., cross-national assessments of learning). I study the models, methods, and designs used to make cross-cultural comparisons of learning and other learning-related constructs. I am chair of the PISA Technical Advisory Group and chair of the NAEP Questionnaire Standing Committee. I limit my comments on this article to the authors claims about ILSAs serving as proxies for IQ.

This article is at best a misunderstanding of the evidence and at worst it is scientific racism. Tests such as PISA, TIMSS, and PIRLS have not been validated to be used as proxies of intelligence. Further, their use as proxies of intelligence is problematic on a number of grounds. First, these assessments are low-stakes for students. This means that their performance on an assessment such as PISA has absolutely no bearing on their lives. As a result, empirical evidence (https://www.oecd-ilibrary.org/education/how-is-students-motivation-related-to-their-performance-and-anxiety_d7c28431-en; https://eric.ed.gov/?id=EJ1293962) has shown that motivation is highly heterogeneous across populations, with students in some countries being highly motivated, and other much less so. This is also evidenced by the so-called “effort thermometer” in PISA where 69% of students in OECD countries reported that they spent less effort on PISA than a graded/marked test. This empirical evidence calls into serious question the comparability of these measures. Second, inspection of trends over time in these assessments show substantial variation in the Flynn effect, with some countries exhibiting a strong Flynn effect and others much weaker or a decline in performance over time. A common response is “the Flynn effect is dead in highly economically developed, industrialized countries.” Ok, then explain why we see a meaningful Flynn effect in Hong Kong in PIRLS or in Japan and US in TIMSS? Clearly, ILSA results are a blend of skill and will (Eklöf, 2010).

So, what are we measuring? Notably, ILSAs are cross-sectional, observational, sample-based, low-stakes measures of either curriculum attainment (TIMSS/PIRLS) or skills necessary to function in a modern society (PISA). That's it. They are not, cannot, and should not be regarded as IQ proxies.

The Lynn and Becker data that the authors cite as further evidence is plagued by methodological issues. For instance, countries that did not participate in international assessments received imputed scores from neighboring countries. These means, for instance, that Albania received scores from Greece in PISA 2018 or that Austria received scores from Germany in TIMSS 2011. A cursory look at comparisons where they do exist suggest that - besides being prima facie obviously ridiculous - “giving” scores from neighboring countries can produce nonsense results. As just one example, Austria performed statistically significantly below the OECD average in PISA 2018 in reading (score = 484). By

comparison, Germany scored 14 points higher - not different than the OECD average. Scores are also corrected with poor or no explanation. For instance, they use a rural sample in Burkina Faso and test them on digit span (a single subset of measures from the WISC-IV). The original score was 90 and the “corrected” score was put at 73.80. Put another way, used as is, this adjusted score puts the population of Burkina Faso at just above intellectually disabled, according to the DSM-V, or what was referred to as “mentally retarded” in early versions of the DSM.

I do not evaluate other aspects of this manuscript; however, the evidence on educational test results is so profoundly flawed that I would strongly recommend this manuscript for rejection.