PISA – Models and the Reality

Anabela Serrão[1], Carlos Pinto-Ferreira[2]

[1]*Instituto Universitário de Lisboa – ISCTE,*
[2]*Instituto Superior Técnico – IST*
(Portugal)
[1]*aserraopt@gmail.com,* [2]*cpf@isr.ist.utl.pt*

Abstract

The Programme for International Student Assessment – PISA – is the most ambitious endeavour of large-scale education systems evaluation ever implemented. The Organization for Economic Co-operation and Development – OECD – launched this exercise for the first time in 2000, and in the 2012 edition 65 education systems were assessed. According to OECD, the programme "[…] is a triennial international survey which aims to evaluate education systems worldwide by testing skills and knowledge of 15-year-old students." And, "[…] tests are designed to assess to what extent students at the end of compulsory education, can apply their knowledge to real-life situations and be equipped for full participation in society."

Albeit being a prestigious programme, entrenched in sound theoretical grounds, and notwithstanding all the efforts made by PISA experts to mitigate shortcomings, the PISA is not exempt from criticisms of various kinds. When analysing the quotes mentioned above, and taking into consideration the applied methodologies, several questions can be raised and some concerns should be pointed out.

The first question arising in the process of evaluation is that any measurement always affects, direct or indirectly, the system itself, disturbing its inner workings. This fact is particularly relevant when social systems are at stake.

A second difficulty results when students from very different countries in what regards culture, tradition, and beliefs are subjected to the same test. Although all items are always carefully analysed by panels of experts in order to detect cultural bias or offending interpretations, there is no complete guarantee that the final set of items is adequate to evaluate all students.

Another question regarding the fairness of PISA results is the fact that a paper-and-pencil (or computer) test, limited to three disciplinary domains, cannot encompass the possibly rich, diverse, and unsuspected knowledge and skills of 15-year-old students.

There are also technical criticisms regarding the adopted approaches and methodologies, from the utilization of the Rasch model to negative remarks about the way data are collected and questions are coded.

Some of what could be considered advantages of PISA – the literacy based instead of a curriculum based approach, the assessment of 15-year-old students instead of a particular school year pupils, and the definition of a large set of indicators, as is the case of ESCS – have been also severely criticised.

Finally, some of the criticisms reside, not in the PISA methods and characteristics themselves but on an excessive focus on country rankings, primarily promoted by media, and consequently followed by political leaders.

The main objective of this research is to reframe difficulties and artefacts together with virtuous results of PISA, putting in perspective praises and criticisms to foster a better understanding of this important programme.

1. Introduction

PISA – Programme for International Student Assessment was launched by the OECD in the year 2000 with the participation of 43 countries/economies. The number of participating countries has grown steadily since then (in 2015, more than 70 countries/economies will participate), which demonstrates the increasing importance of the programme. The worldwide nature of the PISA survey organization, involving a complex and demanding process of student performance evaluation in very diverse countries has earned a remarkable credit, particularly as an invaluable indicator of educational policies adopted in the participating countries. According to OECD, PISA provides international comparisons of performance of education systems through valid and culturally cross-cutting tools for the evaluation of

basic skills for adult life, allowing know how effectively schools prepare students for life after formal education [1].

Since its inception, PISA has been subject to criticisms directed especially to the way country rankings are obtained, and to causal explanations presumably linking reasons to results [2][3]. Not surprisingly, as these two aspects are particularly delicate issues: the former because it touches country rivalry and competition, and the latter as explanation of causal links lies in the core of public policy making.

2. Criticisms and appreciations

The analysis of PISA survey as an assessment instrument performed in this research should encompass all the main aspects of this complex and diverse model. However, taking into consideration the preliminary character of this phase only the most relevant questions are presented and analysed. In this paper, criticisms are categorized as *epistemological, methodological, cultural,* and *operational.*

2.1. Epistemological questions

Epistemological issues are related to the question of evaluation validity: *"is it possible to obtain meaningful and sound knowledge about the performance of an arbitrary number of education systems, expressed by scores, statistically ranked?",*

The observer effect.

Any process of evaluation presupposes measurement, and any measurement impacts the variables under observation. This effect is particularly acute in social science contexts where subjects, institutions and systems tend to modify their behaviour in order to obtain what are considered "good results". In what regards educational evaluation processes, and from the evaluators' point of view, there are *intended* and *unintended* observation effects.

From the PISA evaluators' point of view, the introduction of education systems reforms improving equity, for instance, is an intended consequence of the programme, whereas the specific preparation of the students chosen for PISA evaluation is an unintended effect of the programme.

At a higher level, it is known that certain countries have changed their education systems in such a way that they can be considered "PISA-driven". From the standpoint of PISA co-ordinators, is this a good (intended) or bad (unintended) effect of the programme?

Model adherence to reality.

Models are entities engendered to allow the representation of some aspect of reality aiming at, among other things, description (what is it?), prescription (how should it be?), prediction (what is the future behaviour?), control (how to govern it?), analysis (how is it composed?), and synthesis (how to compose parts?). "Re-*presentation*" (meaning something to be *presented* in a new or different form) always corresponds to the result of an abstraction effort (in order to reduce, to compress, to simplify, and/or to make tractable a complex aspect of reality). As a consequence, models are never perfect entities, and as it was humorously stated by George Box, "essentially, all models are wrong, but some are useful" [4]. In fact, the crucial question here is how to achieve model epistemic soundness: to what extent the model *adequately* represents the relevant aspects of the reality under consideration. Therefore, models are either good or bad, not in absolute terms but only relatively to their adherence to a specific aspect of reality under representation.

When modelling the performance of education systems, PISA makes use of the Rasch model, a psychometric instrument aimed at measuring latent traits of individuals, by exposing them to a suitable number of items, providing stimuli and collecting the corresponding answers. The theoretical framework of this approach is based on the principle of *invariant comparison.* In a nutshell, trait level estimates of students do not depend on the set of items applied, and item parameters do not depend on the particular group of students under evaluation. This principle enables the estimation of data about item difficulty indices, plausible student scores and education system indicators, which should be consistent with the model. It should be pointed out that, in the Rasch model it is not the model that should be modified to accommodate the data, but the other way around.

When the same model is applied, in each PISA cycle, to an ever-growing domain, demanding the adaptation of its characteristics on the fly, it can be questioned whether the model validity limits have been exceeded.

Education as competence acquisition

A collection of items on reading, mathematics and science, based on paper and pencil (or computer) tests is not supposed to encompass all the rich body of knowledge and skills learned through an education process. Education, as a continuous process of learning, involves the acquirement not only of knowledge, skills, beliefs, and principles, but also, attitudes, habits, behaviours, and metacognition

strategies. It comprises ethical and aesthetic values transmitted from one generation to the next, involving all the community intervention – which includes, but is not limited to the school. And finally, paper and pencil (or computer) tests do not allow oral expression and understanding of the spoken word evaluation. Also, test structure limits the creativity of students.

Albeit being very relevant, the desideratum of knowing whether students "can apply their knowledge to real-life situations" and are "equipped for full participation in society", cannot obviously be understood as the evaluation of education systems in their entirety.

A natural apprehension regarding the fact that certain relevant knowledge areas are not covered by the PISA test is related with the observer effect above mentioned: only what is evaluated is well taught and learnt and all the rest tends to be forgotten, and as a consequence, education systems could lose their rich and distinct properties, by a process of regrettable homogenization.

2.2. Methodological questions

Methodological issues are related to the question *"Which methods should be chosen and requirements ought to be met to obtain good model performance?".*

The PISA survey has introduced a very interesting and unique approach to circumvent the difficulties inherent to country education system comparisons by adopting a competence oriented, instead of a curriculum-based approach to evaluation. Although a competence evaluation always presupposes that students possess knowledge and skills, the focus on *how* they solve real-life problems mitigates the difficulties arising from curriculum differences among countries. (Notice that real-life problems are less knowledge-intensive and more based on critical reasoning and common sense). As a competence-oriented survey, PISA elects as its target population, students of a specific age – 15-year-olds – instead of a particular school grade. Given the heterogeneity of the students in what regards grade enrolment, how to ensure that student and country comparability is not compromised. For instance, a 15-year-old student enrolled in grade seven is not "in the end of compulsory education" and certainly is not equipped with the necessary knowledge for "full participation in society" [1].

A different methodological concern regards model requirements. Considering that the Rasch model is the best approach to perform a worldwide evaluation process successfully, it should be ensured that all necessary requirements are met. The mentioned demand of internal consistency hypothesis of the Rash approach entails a critical aspect of model application: item should not behave in a dissimilar way across countries. This requirement forces the withdrawal of items whose behaviour is inappropriate, potentially weakening the adherence of the model to reality.

2.3. Cultural questions

In this kind of issues, the essential question to be answered can be formulated as follows: *"What mechanisms and measures should be utilized in order to ensure that all items are culturally neutral?"*.

Taking into consideration that the vast majority of PISA items were developed by English speaking professionals it is undeniable its culturally distinct character. It should be noted, however, that efforts have been made in order to develop in non-English speaking countries the expertise to create new items.

Since the first cycle of PISA in the year 2000, the programme has evaluated a growing number of countries/economies, encompassing more regions with an ever-expanding set of cultures. In these regions, different cultural landscapes, differing beliefs and religions, dissimilar day-by-day experiences, and diverse languages imply a very difficult exercise of unbiased test development. As a consequence, the translation effort becomes more and more difficult because of the: (i) high number of different languages, (ii) regional differences with respect to the same language, (iii) difficult cultural adaptation of items, and (iv) need for the attainment of a phraseology adequate to 15-year-old students. In fact, after translation, texts tend to become longer possibly introducing significant differences in reading time. Some national versions of released items show poorly accurate translation revealing the use of a too sophisticated or improper wording *(traddutore-traditore)*.

2.4. Operational questions

Operational issues are related to the question *"Which questions, at implementation level can contaminate the evaluation results?".*

The motivation and commitment at student, school and country levels to PISA is an important driver to guarantee that the evaluation results mirror the real value of the country education system. Despite all the rules and criteria established by PISA and that must be carefully followed in every country, the way national centres organize the exercise, not only influences response rates but also impacts the

evaluation results. In a voluntary assessment, motivated students are also a critical factor in the competence and knowledge performance exhibition.

Another implementation question is school participation – regarding refusal and replacement. One of the aspects carefully pursued by PISA is to ensure proper representation of all school characteristics in each country (dimension, location, private/public, school levels, etc.).

A different issue regarding implementation is the process of open item coding. Notwithstanding all rules and procedures, which ought to be strictly followed, there is always an open interpretation of coding criteria influencing the results. For instance, bad calligraphy, orthographic errors, incomplete answers, and/or poor wording when coded by a very demanding coder could result in a punitive attitude, negatively influencing results.

3. Conclusion

Given the ambitious dimension of the programme, considering the set of aforementioned difficulties, PISA co-ordinators have been compelled to foster innovative initiatives and continuous improvement of the survey to best fit a changing reality. In particular, the introduction of computer-based assessment, the evaluation of problem-solving competences, the launching of financial literacy tests, the adaptation of specific booklets to measure lower end competences, to name a few, are examples of constant progress concerns.

Despite all the ameliorations on the survey, as the model is not perfect, results should be taken not in absolute terms – particularly rankings and causal relationships should be read *cum grano salis*. This fact should reframe the discussion from rankings to deeper intra-country analyses, profiting from the richness of PISA databases. Recall that this programme provides the best database (student competence and student, family and school contextual indicators) ever produced and available for national and international research.

References

[1] OECD (1999). Measuring Student Knowledge and Skills: A New Framework for Assessment, Paris: OECD.
[2] Prais, S. (2003). Cautions on OECD's Recent Educational Survey (PISA). Oxford Review of Education. 29-2003-2, 139-163.
[3] Goldstein, H. (2004), International Comparisons of Student Attainment: Some Issues Arising from the PISA Study. Assessment in Education – Principles, Policy, and Practice 11-2004-3, 319-330.
[4] Box, G. E. P., and Draper, N. R., (1987). Empirical Model Building and Response Surfaces. John Wiley & Sons. New York, NY.