Chapter 2
International Large-Scale Assessments as Change Agents

Jo Ritzen

Introduction

Transparency to keep you honest: This notion, in a general sense, had been widely embraced in the national legislatures of many countries as an important change agent towards better societal outcomes. But in education, transparency in the form of educational outcomes—to be measured by (large-scale) assessments, among other methods—is a relatively new phenomenon as an intended agent of change.

This chapter is a survey of my personal experience in the recent past with international large-scale assessments (ILSAs) as change agents, with a focus on PISA, the Programme for International Student Assessment, which started in 1994[1] and has had a progressively larger impact on the educational policies of countries. Many countries felt PISA gave them an honest view of where they were in their aspirations to have the best possible talent development. It was not always a happy view. Sometimes it confirmed earlier fears that the country had fallen off track. Sometimes the PISA results were in sharp contrast to previously held beliefs in the quality of the country's education system. This PISA-shock has spurred a rapid change in country policies, with a likely unprecedented upward spiral in the quality of education.

My survey starts with a theoretical framework of transparency as a change agent. This model is based on assessments that allow for comparisons between institutions. It has to be slightly modified for an application to PISA, which provides a

The following is an adaptation of the keynote address for the International Large-Scale Assessment Conference at Educational Testing Service in Princeton, NJ, on March 16–18, 2011.

[1] While planning and preparation for this endeavor started in 1994, it took until 2000 for the first PISA assessment to be administered in OECD countries.

J. Ritzen (✉)
Empower European Universities, International Economics of Science, Technology and Higher Education, Maastricht School of Governance, UNU-MERIT, Kloosterweg 54 Bunde, Netherlands
e-mail: jo.ritzen@empowereu.org

M. von Davier et al. (eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research,*
DOI 10.1007/978-94-007-4629-9_2, © Springer Science+Business Media Dordrecht 2013

comparison among countries. Then I present the main findings of the PISA evaluation in terms of its policy impact within the framework of the theoretical model. Subsequently, I consider transmission mechanics (in particular, the role of the media) in translating the results of assessments into institutional change before providing a concluding section.

The main point is that assessments make comparisons in the accomplishments of different institutions or regions possible. They can take the role of signaling the quality of the educational institutions or the educational establishment of a region or a country. This signal can drive a healthy competition in which all participating partners profit.

Transparency as a Change Agent: A Think Model

Throughout my career both in government as a Minister for Education, Culture and Science in the Netherlands in the 1990s, as well in different capacities (among others as vice president for education, health and social protection) at the World Bank, I have emphasized the role of transparency as an important change agent towards better educational outcomes within countries and between countries.The theoretical model implicit in this emphasis is depicted in Fig. 2.1.

I consider the educational institution as the unit of performance for which educational outcomes are measured in a way that allows for comparisons with the educational outcomes of other institutions. Of course, one can also focus on the educational outcomes of a set of institutions, e.g., in a region, a state or a country.

The transmission of these measurements to "change" in institutions can take place in three different ways, as shown in Fig. 2.1.

1. The first is the direct line of "naming and shaming," which might lead schools to rethink their policies and become more performance-conscious.
2. The second refers to the reaction of consumers (students and their parents) and stakeholders (the community in which the school is located).
3. The third refers to the reaction of local, regional, or state governments to large-scale assessments.

Any sign that their school performs less well than others may have repercussions. I have found evidence for this in the external evaluation of policy impact of PISA (OECD 2008) in Basque country, where PISA results were known by school. Their parents and local stakeholders of some top-performing schools will make every attempt to have their school perform even better the next time around.

Needless to say, "iteration" (repeating the cycle of measurement) is a necessary condition for assessment to have an improvement function. "Consumer choice," whether direct (in an education system that allows choice even within the public system, or for a switch from public to private, or vice versa) or indirect (through migration to school districts with better schools), does have an impact on institutions.

"Choice" is an important transmission mechanism of quality measures towards the energy and dedication to change. The size of the effects of "choice" on

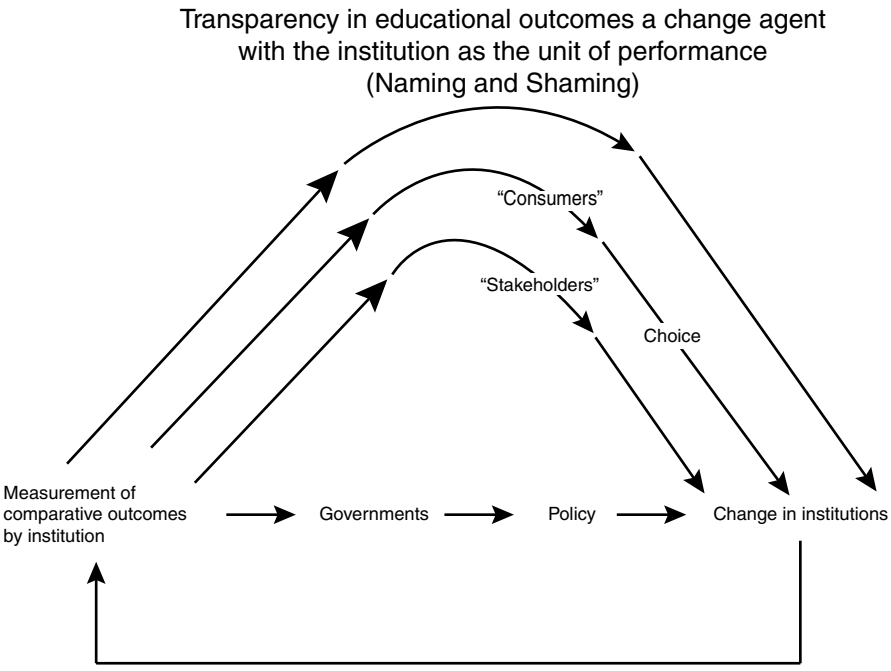

Fig. 2.1 Transparency in educational outcomes as a change agent with the institution as the unit of performance (Naming and Shaming)

institutional behavior depends on a great number of factors. First, the measures will—in education—only partially reveal the main features that consumers (students and/or their parents) have sought in the school. As a result, consumers will only partially let their choice depend on such measures. Second, the institutions may not want to react to changes in choice (decreased student number) on the basis of published performance measures, because they believe that those measures do not represent their strategic aims in full. Some schools may prefer to have a smaller student body along with a different educational mission than what is captured in the performance measures.

Some examples of the Netherlands in the 1990s are the higher education "Choice Guide" and the inspectorate assessment of schools.

In the 1990s, the Ministry of Education organized and financed the Choice Guide. In this guide, all degree courses of all higher education institutions were compared annually in groups of similar courses (e.g., medicine, economics, law, etc.), using student evaluations on a large number of items (approximately 30). The evaluations were summarized by grade. The effect on choice was and still is limited. However, the impact on institutions has been impressive. Some institutions have made it their mission (advertised in public) to be among the best evaluated and work hard to correct low evaluation scores. Most institutions do their best to avoid low scores. Those with consistently low scores for degree courses over a number of years are bashed in the press. Although they still may not have lost too many students (presumably because of the geographic limitations of choice), they are under considerable pressure to improve.

Although no detailed analysis exists, it is hard to believe that the annual publication has *not* led to a substantial improvement of higher education (in order of decreasing importance: Ph.D., master's degree, university bachelor's degree, professional education bachelor's degree, community college bachelor's degree), albeit only through choice, when serious alternatives were an issue.

The assessments of secondary schools of the Dutch inspectorate were published in one of the national morning newspapers (*Trouw*) for the first time on a Saturday in 1996. The demand was so overwhelming that the paper was sold out at 8 a.m. The publications did and do affect choice, and schools will do almost everything to ensure they are well evaluated on their performance by the inspectorate.

In Fig. 2.1, I put "stakeholders" next to "choice." In those countries where choice is not used as an allocation mechanism, school boards and other local stakeholders can be the parties that put pressure on institutions to improve in terms of their performance measures.

The third line of change is the one of government. Government may choose to alter policies or even close down schools that are performing badly. The "No Child Left Behind" policy in the United States allows for such an approach in combination with the possibility of revamping the school as a charter school.

Both the United Kingdom and the Netherlands (as examples) have legislation that allows for the termination of public finance for low-performing schools. Needless to say, this refers only to a limited number of cases each year. Policies are not so likely to be altered if the results of assessments are not known by institution.

The direct impact ("naming and shaming"), the impact of stakeholders, and the impact of government agencies are strongly related to the way the public interest is taken up by the media. In the subsequent section, "The Policy Impact of PISA," I will discuss their role more in detail.

To this point, I have taken a "technical" position on assessments and change without asking the more normative question whether such change is socially desirable. This question uncovers the "value-added" debate. Educational accomplishments indeed should be considered in a value-added fashion. Unfortunately, few countries have followed the lead of Poland in showing a willingness to pursue this necessary direction.

When assessments are taken to the country level—as with PISA—and used for cross-country comparisons, the value-added question is less intrusive, assuming that the base distribution of learning achievements of, say, 5-year-olds is not so different among countries, at least those with similar per capita incomes.

The resistance to participation in ILSAs—although substantially different among countries—often comes from the education community. Many education leaders express the fear that assessments drive them "to teach to the test" or that the assessments "label" students. Yet, high quality performance tests that are not used to promote or select students are unlikely to have this effect. The "labeling" of students indeed can take place, because the education environment can be adapted based on the test to serve the student better, as an intended effect.

In the case of the Dutch Choice Guide for higher education, there was originally a strong resentment from the side of the higher education institutions that was

expressed as unease with the types of measures used. But one cannot escape the impression that the resentment was based more on unwillingness to be transparent than on concerns about the types of measures used.

The Policy Impact of PISA

PISA is an internationally standardized assessment that monitors the quality of education systems in terms of student outcomes. PISA assesses the ability of 15-year-olds to apply their knowledge in reading, mathematics, and science to real-life problems, rather than the acquisition of specific curriculum content. Assessments take place every three years and use a framework that is jointly developed by Organisation for Economic Co-operation and Development (OECD) countries. Contextual data are collected through background questionnaires for students and schools, with between 5,000 and 10,000 students typically tested in each country.

The first survey was conducted in 2000. It focused on reading literacy and measured students' "capacity to understand, use and reflect on written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society." The survey was completed by students in 43 countries (29 OECD member and 14 nonmember countries and economies; for 11 of the 43 countries and economies, data was collected in a follow-up study, PISA-PLUS, in 2002).

The second survey was conducted in 2003. It assessed students in mathematical literacy and examined young adults' "capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned, and reflective citizen" (see page 12 in OECD 2006). A total of 41 countries (30 OECD member and 11 nonmember countries and economies) participated in the 2003 assessment cycle.

The third survey was conducted in 2006. It had science literacy as its focus and assessed the capacity of students' "scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues, understanding of the characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen." A total of 57 countries (30 OECD member and 27 nonmember countries and economies) participated in this survey.

The fourth survey was conducted in 2009 (see OECD 2010).

In 2007 the OECD commissioned a group of three individuals, of which I was one, to do an evaluation of the impact of PISA on policy. This group produced a report one year later.

The research design consisted of the following two parts:

- A *quantitative* strand: A total of 905 questionnaires were distributed to policymakers, local government officials, school principals, parents, academics and research-

ers, and media representatives in 43 countries and economies (of which 24 were OECD member countries) via email. Of these, 548 questionnaires were returned. This corresponds to an overall response rate of 61 %. Furthermore, responses were obtained from 42 representatives at the PISA Governing Board, 33 members of the business community, and 36 representatives of teacher organizations.

- A *qualitative* strand: Five case-study countries and economies were selected, taking into account variations in terms of the levels of impact PISA has achieved, performance in PISA, and equity and government structure (centralized/decentralized/federal/regional). Geographical balance was also taken into consideration. The case-study countries and economies were Canada, Hong Kong-China, Norway, Poland, and Spain.

I personally visited Spain and Poland as part of this intensive, in-depth qualitative review.

Let me first focus on the question of whether PISA acted as a change agent, and if so, how. Our model of Fig. 2.1 needs some adaptation because, in this scenario, consumer choice is irrelevant (generally, students will not choose to follow their education in other countries).

The overall level of policy impact of PISA in each country was estimated by combining the respondents' assessment of the extent to which PISA influenced policymaking at the national/federal and local levels in all three PISA assessments. The categories were constructed based on the distribution of answers to the questions of countries that returned more than four questionnaires. Early on, countries asked for and received the assurance that individual country data would not be published by us. Rather, deciles were generated from the distribution of respondents, who judged that PISA was extremely or very influential in informing policy. PISA was considered to have a comparatively low impact in countries falling into the range from the lowest to the third decile. The policy impact of PISA was considered to be medium in countries from the fourth to the seventh decile, and high in countries in the deciles above. This resulted in the following classification:

- Group A
  - *Countries where PISA achieved relatively low levels of impact on policy formation*: Czech Republic, Ireland, Italy, the Netherlands, the Slovak Republic, Turkey, the United Kingdom, Bulgaria, Croatia, Hong Kong-China, Latvia, Lithuania, Romania, the Republic of Serbia, and Uruguay.
- Group B
  - *Countries where PISA achieved relatively medium levels of impact on policy formation*: Australia, Austria, Belgium, Canada, Finland, Greece, Hungary, Iceland, Switzerland, Chile, Chinese Taipei, Colombia, and Qatar.
- Group C
  - *Countries where PISA achieved relatively high levels of impact on policy formation*: Denmark, Germany, Japan, Mexico, Norway, Spain, Sweden, Israel, the Kyrgyz Republic, Macao-China, Slovenia, and Thailand.

Note that this survey took place in 2007. If we had done this after the 2009 PISA results were published in 2010, the picture would have been quite different: The United States and the United Kingdom reacted quite strongly to the PISA observation that their countries continued to belong to the low achievers among the richer countries. All in all, it seems that it takes a while for the PISA message to sink in into the policy domain in those countries that are not top PISA performers.

What then is the framework in which these reactions can be placed? Surely the reactions are rooted in the discrepancy between expectation and realized results. But how do countries come up with expectations regarding PISA results? Several scenarios seem likely:

- Expectations are based on "comparison" (neighboring?) countries (applicable to 2000 PISA).
- Expectations are based on expected changes over time due to national improvement efforts (applicable to 2003 and 2006).
- Expectations are based on a combination of both.
- Expectations are based on the wish to belong to the "world top."

The relation, on the one hand, among the above indicators derived from the PISA scores for 2003 and 2006 and, on the other, the measure of the reaction on PISA, was statistically analyzed without any significant result. Neither "neighboring" country score, average score, nor world top fits the bill as "comparison/benchmarking" for all countries. For Spain and Poland (the countries that I could study in depth), their reaction can only be explained by the ambition to belong to the world top: The benchmark in terms of expectation are to be the very best PISA countries in the world, even if neighboring countries seem to take a more "relaxed" attitude. Digging one spade deeper, I tried to ascertain the level of ambition with respect to the educational accomplishments of 15-year-olds in the country by looking at general government documents (such as annual budgets).

It appears indeed that countries expressing great general ambitions ("belonging to the top 10 in competitiveness"—a goal that at least some 50 countries in the world endorsed in the survey!)—also are more ambitious with respect to education and seem to react more strongly to the PISA results than others.

PISA, of course, showed that educational outcomes differ remarkably among countries, including countries with similar levels of income. Figure 2.2 shows the performance distribution for five different countries and the OECD average.

Finland might be considered to have reached "the production possibility frontier" (as economists would call it), while other countries still have substantial room for improvement.

Figure 2.3 shows that a great majority of questionnaire respondents (85 %) regarded policymakers as the main *stakeholders* responsible for implementing policies in light of PISA, followed by school principals and local government officials. Professional teacher associates and academics and researchers were considered third and fourth. "Consumers" (parents) or stakeholders reflecting the "consumers" side (the business community) are not regarded as important in engendering change based on assessment. If you think that these results are brought about by the framing

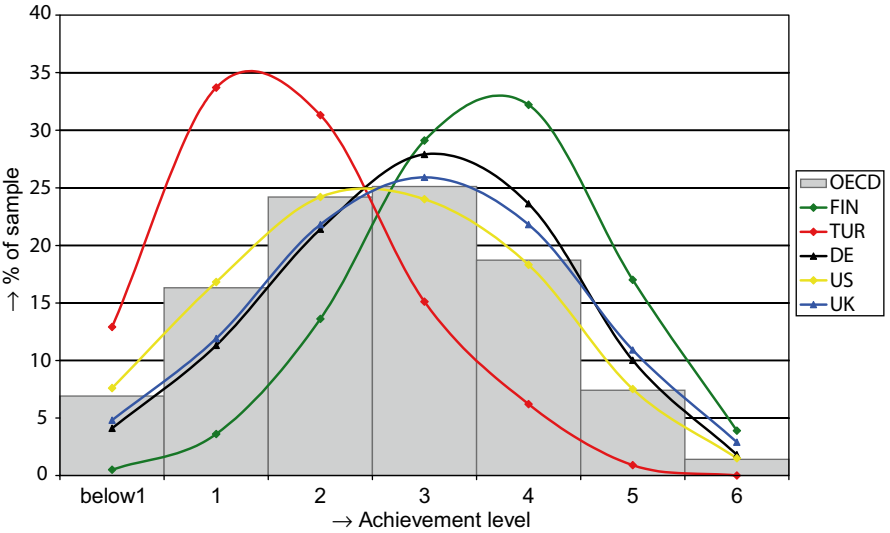

Fig. 2.2 The Academic Achievement Curve (selected countries), *OECD* Organisation for Economic Co-operation and Development countries, *FIN* Finland, *TUR* Turkey, *DE* Denmark, *US* United States, *UK* United Kingdom. (Source: Ritzen 2010, p. 177)

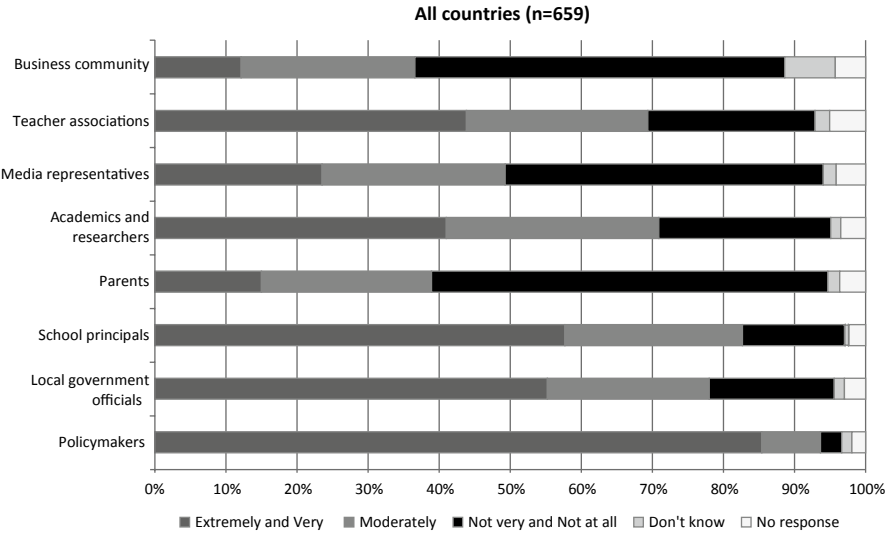

Fig. 2.3 Stakeholders responsible for implementing policies in light of the PISA results, Question: Who would you identify as being responsible for implementing policies in light of the PISA results in your country? Please indicate the degree of responsibility for each stakeholder and specify to which PISA assessment you refer

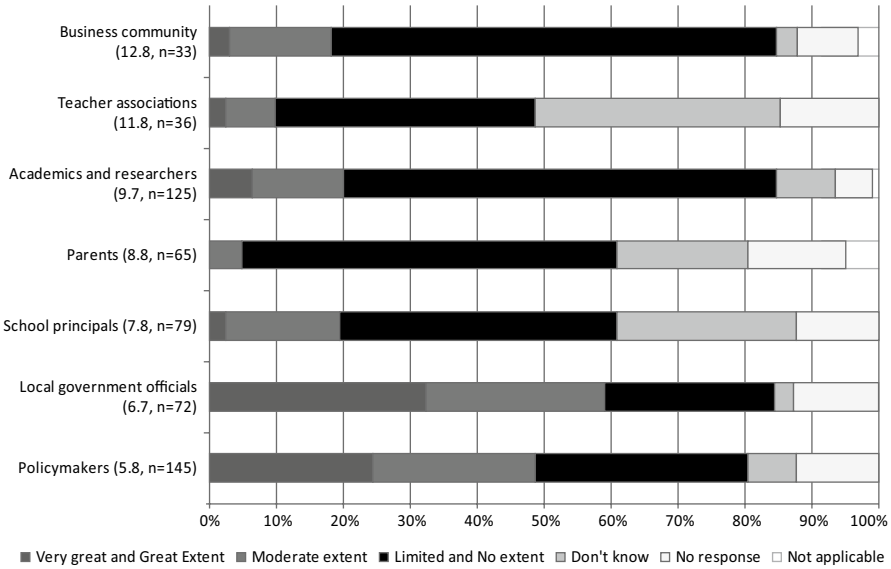

Fig. 2.4 Extent of key stakeholder's own responsibility for results, Question: To what extent do [members of own stakeholder group] feel responsible for your country's results in PISA?

of the questions, then it is good to know that all questions (including, "Who would you identify as the most significant stakeholders in PISA and its results in your country") led to the same result.

One would imagine that the responsibility for the PISA results would be claimed (in case of good results) by all parties (and definitely by the schools). Of course, one would expect all parties to dodge responsibility in the case of not-so-good results. However, the latter seems to be more generic as Fig. 2.4 shows.

All parties dodge responsibility. Thirty-two percent of local government officials and 24 % of policymakers claim responsibility, but *only 2 % of school principals and representatives of teacher organizations respond that they feel responsible.* Unfortunately no analysis was made along the lines of positive or negative responsibility answers in relation to higher or lower scores. For me, this response is beyond comprehension. How can a whole sector ignore its responsibility for the results, as these results would seem to imply? Note that we are not looking at an isolated reply from one country, but at a reasonably well spread group of respondents from a sizable number of different countries.

One might have expected that the link between PISA and "change" would be less if the stakeholders felt that PISA did not adequately address the core mission of education as viewed by education institutions, the government, or the "consumers," as we suggested in our model in the section on transparency as a change agent. This is, however, *not* the case. Overwhelmingly the stakeholders regarded student

performance in reading (78 %), mathematics (75 %) and science (71 %) as well as international comparisons and rankings (70 %) as extremely important or very important.

This was also the case when reporting on PISA regarding the relationship between home background and student performance (58 %), the relationship between school context and student performance (53 %), and student interests, motivation and attitudes (48 %).

In the qualitative part of the research, for which I was able to visit schools for whom institutional PISA scores were available, I did find a confirmation of the hypothesis that schools that do well on PISA also feel more comfortable with the PISA outcome measurements as reflecting their mission, while schools that underperform in PISA indicate that they feel that civic education, socialization, and a broad development of talents of students are more important than the three domains as measured in PISA.

This was also found in a question in the quantitative part—to what extent PISA addresses the policy needs of participant countries and economies. The answers were less positive than those on the relevance of the measures used, maybe because of different objectives on the part of stakeholders than those captured in the actual measurement.

Of course, stakeholders might also be apprehensive of the accuracy of the measurement, but that did not turn out to be the case (OECD 2008, p. 24).

Transmission

Large-scale assessments can only play a role as a change agent if the information derived from these assessments reaches the stakeholder, and even more so if stakeholders are challenged because of the results of the assessments. The media play a tremendously important role in this process. However, this turns out to be an autonomous and rather unpredictable role. One would surmise that the PISA results would compare to the benchmark of other countries in the same league as a predictor, according to the same expectation model as was suggested earlier for the overall impact of PISA on policy (expectations based on the average, the neighboring countries, or the world top).

Again, although not considered statistically, the evidence suggests (as with the policy reaction) that there is no clear cut case for suggesting any general "expectation" model on the part of the media.

Media are part of the dissemination of PISA results. In general, the dissemination of results by the media played a substantial role in determining the policy impact. Some country governments were aware of this and organized dissemination through press conferences held for representatives of the media, as well as through conferences with stakeholders and schools (as Poland and Spain did). But other countries simply let PISA run its course and paid little to no attention to dissemination.

Here we recognize the dissemination effort as it relates to the government's interest in policy change. Dissemination (including media coverage) does not seem to be an exogenous, but rather an endogenous variable in the model, with the exception of that part of the media that plays more an "NGO (nongovernmental organization) role."

I found a compelling similarity between the reaction of Lang (2010) to the No Child Left Behind assessments and reactions to PISA at the school level: If you want large-scale assessment to have an impact on schools, then you should disseminate results and discuss them with teachers, parents, school boards, and so forth on the local level.

Also in the PISA evaluation, the local level—where the changes should take place—felt uninformed and uneasy as well.

Even if all were well informed, the question is, What do you do with the PISA results? How should a restructuring of schools take place so that better results can be achieved?

Conclusion

1. ILSAs can be important change agents, provided that the assessment addresses the primary concerns of stakeholders in education. The diversity in the objectives of education and differences in the priorities that different stakeholders place on distinguishable objectives (like mathematics versus citizenship) will generally reduce the impact of large-scale assessment as a change agent to some extent. It is important to include a variety of measures in the assessment that reflect principal components of the diversity in the missions schools have adopted.
2. Most large-scale assessments dodge the value-added question. This undermines their potential for change. The present shortcut to measure only outcomes and not to include value-added is unavoidable, but value-added measures should be conceived and implemented in future assessments.
3. The drawback of "teaching to the test" inherent in the impact of large-scale assessments on policy may be exaggerated, unless the survey results are used as high stakes tests.

For ILSAs such as PISA, policy change seems to depend mostly on the level of ambition of the country as expressed in the comparison/benchmarking. Some countries seem to be happy to follow the mean, or the mean of the scores of neighboring countries. Others aspire to become part of the world top performers.

Overall, we conclude that the competition between educational institutions or educational establishments of regions or countries can have a healthy, quality-improving effect on education, once a proper quality signal in the form of an assessment that allows for comparisons is developed.

References

Lang, Kevin. 2010. Measurement Matters. *Journal of Economic Perspectives,* 24(3, Summer): 167–182.

OECD. 2006. *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006.* Paris: OECD.

OECD. 2008. External evaluation of policy impact of PISA. Paris: OECD.

OECD. 2010. PISA 2009 results. Paris: OECD.

Ritzen, Jo. 2010. *A Chance for European Universities*. Amsterdam: Amsterdam University Press.