Large-Scale Research and the Question of Educational Experience: Recovering the Heart of the Matter

Padraig Hogan[1]

Abstract. The international rise of large-scale evaluations of quality in education in recent decades has been accompanied, ironically, by a major neglect of the issue of quality itself. This is mainly because of fundamental shortcomings in the design of most large-scale evaluation programmes. In programmes like PISA, TALIS and SABER-Teachers, the central but intricate questions of quality become defined as questions of indexed quantity, thus deflecting the former to the margins, or out of the picture. Beginning with a review of the points just mentioned, this contribution then proceeds to identify and investigate some key inadequacies in the conceptions of evidence employed by large-scale evaluation programmes. Deficiencies in their gathering of evidence are likewise examined. A comparative perspective is also included to reveal overlooked exclusions in the supposedly neutral PISA instruments. The manifold character of what an adequate research exploration of quality in educational experience would look like is then investigated. Finally, the case is made for advancing the kind of evaluation programme that includes an adequate understanding of quality in education and that does justice to this in its research design.
Key words: evidence, resonance, experience, quality, research design.

Introduction – A Question of Quality?

The idea of evidence-based policy has become widespread internationally in recent decades. It has given a new and strategic importance to research, not least to educational research. On the face of it, this is good news for both researchers and policy-makers in education. It represents a dramatic shift from the indifferent and sometimes hostile attitudes toward research that originally accompanied the rise of neoliberalism in policy-making following the Reagan-Thatcher decade of the 1980s (Gillard 2018, Ch.15; Bassey 1992). Since the turn of the century, international bodies like OECD and the World Bank have been to the fore in advancing the notion of evidence-based policy. This advance has also given large-scale evaluation exercises a more central place in educational research. Prominent among International al Large-Scale Assessments (ILSAs) are the following five: (1) PISA (Programme for International Student Assessment) and (2) TALIS (Teaching and Learning International Survey), both directed by the OECD; (3) the SABER-Teachers programme (System Approach for Better Education Results), directed by the World Bank; (4) the TIMSS programme (Trends in International Mathematics and Science Study) and (5) the PIRLS (Progress in International Reading Literacy Study). These last two are run by the International Association for the Evaluation of Educational Achievement (IEA). While both of them share the aspiration of influencing policy, this aspiration is not as much to the forefront as it is in the thinking and actions of inter-governmental bodies like the OECD and World Bank. Despite increasing criticism of it in international research literature (e.g. Meyer & Benavot 2013; Sjøberg 2019; Zhao 2020), the PISA programme is now the world's most widely-used evaluation instrument for the work of schools. It has moreover become highly influential in shaping national educational policies, a role which it actively promotes for itself. (OECD 2019a, pp. 4-5). Finally, PISA is also probably the world's single largest exercise in educational research, and in the conduct of this research it reveals a decisive preference for certain forms of evidence rather than others.

The notion of evidence-based policy, whether at international, national, or school level, is to be applauded, indeed promoted, if it is adequate to its task. The conviction that policy needs to be continually informed by the most pertinent and incisive research insights is one that bodes well for educational practice, including educational leadership. Such an apparently promising development can, however, do serious harm if it proceeds from the start with a deficient concep-

[1]   Maynooth University.
       Mail: padraig.hogan@mu.ie
       ORCID: https://orcid.org/0000-0003-2165-1833

tion of evidence. Adequacy in evidence is a basic requirement in all research and this requirement needs continual vigilance and interrogation. Research evidence in educational evaluations can hardly be sufficient if it continually fails to take account of the most significant things that are happening in daily experience in school and college classrooms. Consider, for instance, the following strategy that has become internationally prevalent since the turn of the century. In the ever-expanding business of quality assurance, crucial matters of quality are routinely recast as matters of indexed quantity – viz. of grades, scores, and a variety of other performance indicators. In this recasting the actual quality of educational experience – i.e. its degree of richness or poverty – frequently becomes eclipsed, in favour of certain kinds of learning performances that can be numerically measured, indexed, and ranked.

This eclipse of educational experience diverts attention from the very things that make educational practice itself fruitful, for both students and teachers. As professional efforts get preoccupied with the kinds of goals that are readily measurable, educational practice becomes increasingly misshapen. The evaluation of that practice and its products gets settled, even institutionalised, in a restricted mode. It largely bypasses the incisive kinds of analysis and self-reflection, of discernment and judgement, of perceptiveness and appraisal, which are central to an adequate understanding and pursuit of quality in the daily work of schools and colleges. A similar shift of emphasis has taken place in educational research more widely, especially in large-scale initiatives that seek to influence policy-making. Here, data-gathering casts its nets ever more widely, giving suggestive impressions of incremental advances in thoroughness. In the case of PISA for instance, this now includes tests of "global competence" (2018 assessments) and "creative thinking" (2021 assessments) among 15-year-olds, as well as its established tests of reading, mathematics and science. It also includes assessments of students' "well-being." In the case of SABER-Teachers it includes an ever-extending range of "policy goals" designed to make teaching more strategically effective on a large scale (World Bank 2013, p. 12 ff).

In short, what seem to be promising developments on the face of things – the possibilities offered by a major move to evidence-informed policy-making – turn out on closer inspection to be initiatives laden with a welter of unseen harms and troubles. In the effort to make quality in education a central policy concern, albeit to encourage its incremental improvement, real questions of quality may recede to the side-lines while an elaborate lookalike assumes the centre stage.

This opening analysis may strike many readers as counter-intuitive, or even mistaken, so the kinds of things that warrant it need to be interrogated. And we will turn to this task now. The interrogation, focusing on the major ISLAs, should reveal some crucial differences between contrasting conceptions of evidence; not only between what each can do and cannot, but also between how certain things become highlighted and others become marginalised or excluded. The interrogation will begin by considering the assessment of students (PISA) and will then review the assessment of teachers (TALIS and SABER-Teachers).

Evidence of Quality? Reviewing some prominent examples

(a) *The OECD Programme for International Student Assessment*

In the results of the 2018 round of the PISA programme, China was ranked number one in the world. Its average score of 578, over reading, maths and science, was 90 points above the OECD average of 488. The documents where the results are published refer regularly to "quality" in describing PISA's core work; for instance, "quality of learning outcomes", "quality assurance", "PISA Quality Monitor", "high-quality and equitable education" (OECD 2019a). The clear inference from these frequent references is that the PISA results show that China's 15-year-olds experience the highest quality of education in the world. For a number of reasons however, this would be a quite misleading conclusion to draw, indeed a deceptive inference to suggest. The two main reasons are the selection of evidence and the composition of evidence.[2] We will take the selection issue first. The PISA programme is designed to evaluate certain educational achievements of 15-year-olds in the participating countries every three years. A representative sample of students take the PISA tests in each country. Yet China participates not on the basis of such a representative sample, but on the basis of schools and students selected from its most developed regions. In the PISA 2018 assessments (also 2022) these are Beijing, Shanghai, Jiangsu, Zhejiang.[3] The majority of China's provinces, containing rural and urban populations where economic and social development are much lower, are not included. Results from China appear in the OECD publications under the devised heading "B-S-J-Z (China)", although they often appear in news headlines under "China."[4]

By permitting the inclusion of results from just some of China's most advanced regions and disregarding chronic problems in China's rural schools (Yiu & Yun, 2917), the PISA programme undermines its own scientific preten-

---

[2]   There is a third major reason also, namely the difficulty of ascertaining what actually goes on in teaching and learning in Chinese schools. Contributing factors to a highly complex picture include the Chinese government's insistence on schools' conformity with the official ideology on the one hand and a "vast private tutoring sector worth $100 billion annually" on the other (Jones & Wu, 2021). On the question of recent and current trends in Chinese education see Hessler (2022), Jones & Wu (2021), Park (2022).

[3]   Just four provinces also represented China in the 2015 round of PISA: Beijing, Shanghai, Jiangsu, Guangdong. In the 2012 round, Shanghai, Hong-Kong and Macau were the only entries under China. If the OECD has a rationale for allowing participation by special regions, as distinct from participation as a whole country, that rationale remains unexplained, even unmentioned, in its publications. By contrast, the technical appendices of PISA reports provide detailed explanations on how schools and students are selected.

[4]   For instance, *The Washington Post* 4 December 2019: *The Irish Times*, 3 December 2019; *Daily Mail*, 3 December 2019; *South China Morning Post*, 3 December 2019.

sions. Abandoning the principle of comparing like-with-like, it skews from the outset the evidence-gathering, and the evidence products, in favour of the higher scores that schools from the entity B-S-J-Z (China) can produce. To place it in an objective perspective, this is comparable to allowing the United States to participate on the basis of four of its most developed states, say Massachusetts, Connecticut, New Hampshire, Vermont, provided that these results would be published under a devised heading like M-C-NH-V (USA). Pursuing to its conclusions the curious logic of this idea, it would allow all federal countries – e.g. Argentina, Brazil, Germany, Mexico, Spain – to enjoy participation on the same basis as China. As well as highlighting the underlying invalidity, this would render the PISA programme largely meaningless as a ranking exercise. In short, its credibility would be gone.

There is a further inequitable aspect to the devised headings rationale. This becomes distinct if we move from the high-performance end to the lower end of the scale. The case of Finland illustrates the point most clearly. The 2018 PISA results from Finland showed a continuing drop from the prime position the country held in 2000, 2003 and 2006. Analyses of the Finnish results since the first round of PISA show that the scores of the top performers have remained consistently high, but that the country's final score has been dragged down by the increased numbers showing up at the lower end of the scale: level 2 or below. For instance, Finland's percentage of those reaching level 2 or below in reading rose from 7% in 2000 to 13.5% in 2018 (Ahonen 2021, pp. 125-6). The biggest performance gaps in Finland moreover occur within individual schools, not between schools in different locations. Unlike China, there is no device available to Finland, or to broadly comparable countries, to exclude beforehand large numbers of low performers from the evidence base of the evaluation exercise.

Turning now to the second issue, the composition of the evidence, the PISA programme in recent years has shown improvements in its design. It has sought to address criticisms of its earlier practices of collecting test results from just three areas: reading, mathematics, science. The design of test items in these three, moreover, has advanced substantially in successive rounds of the programme. The tests increasingly seek evidence of higher-order capabilities such as analysis, judgement and the tracing of inferences. In addition, the questionnaires to students, school principals, teachers and parents that accompany each round of PISA have become more sophisticated. These yield some keen insights on things like co-operation between students, support to students from teachers, students' sense of belonging in schools and teachers' attitudes and practices (OECD 2019a, p. 15). The findings gathered by these means are analysed in detail to identify correlations between positive or negative aspects of school climate and the scores achieved by students in the tests in reading, mathematics and science. These changes to the design and conduct of the programme, and particularly the development of the questionnaire strand, lessen considerably the traditional one-dimensional nature of PISA results. But the changes do little to remedy the deficiencies of PISA as a valid appraisal of the quality of the education, or of the educational environments, experienced by participating students.

It is interesting to note then that the PISA programme is expanding the scope of its research, as it seeks to reach beyond test results into a second strand of data, gathered by questionnaires. Its purpose here is to try to capture the flavour of the learning environments of the schools where its tests are completed. The expansion is connected with PISA's declared aim of focusing on "creativity and critical thinking" and "students' well-being" and on the educational circumstances conducive to these (Schleicher 2019). Volumes III and VI of the 2018 results – respectively titled *What School Life Means for Students' Lives* (OECD 2019b) and *Are Students Ready to Thrive in an Interconnected World?* (OECD 2020) – are based on questionnaires, completed by students, school principals and teachers. The data provided by these, and their analysis by the PISA researchers, give readers some wider context for making sense of the results in the three test areas: reading, mathematics and science. This second strand of data, however, comes essentially in the form of terse responses to pre-designed questions. The respondents, including students, principals, teachers, parents, do not receive an opportunity to speak in their own voice, or articulate their considered views. Accordingly, the responses offered are predominantly perceptions and opinions that are taken at face value rather than interrogated for their soundness and incisiveness. Such probing could be possible if a range of carefully tailored qualitative approaches were also employed. Most importantly however, the dearth of evidence in forms such as case studies, examined narrative accounts, or illustrative action research studies, means that the experienced quality of the learning occurring in the schools remains beyond the reach of PISA research. This experiential dimension – essentially the missing third strand – is the source and core of the matter where the question of quality is concerned. Its absence from the PISA research means that the test results and questionnaire responses present a crowded behavioural surface behind which lies a vast but largely unreached world of educational experience.

This absence is significant not only from the standpoint of the thoroughness of an evaluation and research exercise where the stakes have become very high for the participating countries. The omission also ill-serves the OECD's own declared purpose of seeking to illuminate the kinds of educational environments that are conducive to the growth of creativity and critical thinking. An adequate comprehension of such environments, where the quality of the students' participation and benefit is the pivotal issue, would extend well beyond the limits of the PISA design to date. There is little sign so far in PISA of anything like the third strand of evidence just mentioned. We will come back to this question in the last section, in connection with exploring the features of a research strategy that might give a fuller account of educational experience and its fruits.

(b) *The TALIS and SABER-Teachers Programmes*

The OECD conducts its TALIS programme at five-year intervals. It began in 2008 with 24 countries, increasing to 48 in 2018. Like PISA, it concentrates mainly on lower secondary education, but both primary and upper secondary

schools participated in the 2018 survey. Its central research strategy is the use of questionnaires issued to teachers and school principals. In relation to the aims and purposes of TALIS, these are described by the OECD in the following words:

> Its main goal is to generate internationally comparable information relevant to developing and implementing policies focused on school leaders, teachers and teaching, with an emphasis on those aspects that affect student learning. It gives a voice to teachers and school leaders, allowing them to provide input into educational policy analysis and development in key areas. It is also a collaboration between participating countries and economies, the OECD, an international research consortium, teachers' unions and the European Commission (OECD 2019c, p. 19).

A concern with evidence-informed policies, based on the like-with-like principle, is clear from the opening lines of this statement. Also, there appears to be an essentially consultative tenor running through the TALIS rationale, which would be in keeping with its aim of giving teachers and school principals a voice. From the official descriptions given, the programme's authorities emerge as being in a listening mode where evidence from teachers and principals are concerned. Likewise, the descriptions suggest a collaborative mode in the programme's dealings with representatives of participating countries, teacher unions and the European Commission.

A searching critique of TALIS by Susan Robertson and Tore Sørensen, published in the *European Educational Research Journal* in 2018, questions this benign picture. The critique argues, firstly, that the OECD has taken over much of the power and influence in educational policy matters previously enjoyed by bodies like UNESCO and the International Labour Organisation. Secondly, it contends that the OECD has effected a similar acquisition of many of the educational policy powers historically held by national governments. Thirdly, it highlights the point that the consultations engaged in by TALIS include influential voices that are not explicitly found in the OECD's own descriptions. Fourthly, it charges that the OECD "has brought in a new set of more globally-oriented actors to flank its governing activity in the field of symbolic control – from the European Commission to transnational education companies, global academics and global consultancies" (Robertson & Sørensen, 2018, p. 478). Robertson & Sørensen conclude that the most instrumental forces in designing and refining the TALIS research model are those concentrated in the TALIS Governing Board, not those arising from responsive listening to teachers and principals. These are serious criticisms, but it is hard to uncover the true picture here. For instance, an examination of TALIS literature, including Acknowledgements and Annexes that record the various TALIS contributors, found no mention of "transnational education companies." But the "cooperation of the teachers and principals" is mentioned as "crucial" in an Annex to Volume I of the TALIS report for 2018 (OECD 2019c, p. 211). Clearly however, the aim of "giving a voice to teachers" is not matched by a research strategy that enables teachers and principals to articulate – i.e., to put into their own words – their key professional experiences.

Robertson and Sørensen point out that while Education International (EI – the world federation of teachers' unions) has gained "permanent observer status" on the TALIS Governing Board, this falls short of full participation in the Board's deliberations and decision-making. EI is consulted on drafts of material in preparation and can offer comments and ideas. "However, EI does not have the opportunity to decide on the encodings of what a good teacher is, only to react and comment on them" (Robertson & Sørensen, 2018, p. 479). The contrast between what TALIS says of itself and what the critique of Robertson and Sørenson says of it makes it difficult to ascertain clearly who is really involved in designing the research procedures, or the strength of that involvement. It would be reasonable however to draw the two following conclusions. Firstly, despite the commitment to giving teachers and principals a voice, the TALIS programme excludes the very instruments that are best suited to pursuing this goal. Secondly, it seems that TALIS consults seriously with participating teachers, but when it comes to influencing its research design and procedures, teachers' representatives are not among the decision makers.

Having a consultative role rather than full membership may well place practitioners in an inferior position where designing the features of high-level research on their own work is concerned. But to exclude them from any meaningful role turns an inferior position into an invidious one. This is what criticisms of the SABER-Teachers project have alleged and we shall review these criticisms briefly now.

When the World Bank began the SABER-Teachers programme it took the view, as its "Framework Paper" of 2013 explains, that the evidence was "scattered and incomplete" on "the impact of teacher policies on the quality of the teaching and learning process" (World Bank 2013, p. 7). Based on that view, it adopted a twofold approach to developing its research and evaluation framework: firstly a "policy mapping" step and then a "policy guidance" step (p. 7). The mapping exercise produced ten headings, or "dimensions", on which data on teachers would be gathered. These were: entry and retainment requirements; initial preparation; recruitment and employment; workloads and autonomy; professional development; salary and benefits; retirement and benefits; monitoring and evaluating the quality of teaching; teacher representation and voice; school leadership (p. 12). Despite important omissions (e.g., the quality of teachers' relations to students, to their teaching subjects, to colleagues, school principals, parents), each of these headings recognises a pertinent and potentially valuable field for evidence gathering. Questions like the following then become central: What kind of evidence is to be gathered under each heading? Who does the gathering? Who decides the scope of the research and designs the procedures?

The answer to the third question – who decides and who designs? – will have a critical bearing on how the other two are answered. A search of SABER publications failed, however, to find an explicit answer for this third ques-

tion.[5] In any case, the policy guidance part, or second step, of the SABER framework produced eight policy goals to guide the actions of governments in order to lead to "improved student outcomes." These eight are:

> 1. Setting clear expectations for teachers; 2. Attracting the best into teaching; 3. Preparing teachers with useful training and experience; 4. Matching teachers' skills with students' needs; 5. Leading teachers with strong principals; 6. Monitoring teaching and learning; 7. Supporting teachers to improve instruction; and 8. Motivating teachers to perform (World Bank 2013, p. 24).

One of the headings in the mapping exercise, "teacher representation and voice", was dropped from the policy guidance stage. Critics of SABER-Teachers agree that this omission of teachers is associated more with the World Bank's own neoliberal orientation than with the technical reason offered by the World Bank itself in its 2013 "Framework Paper" (Robertson 2016, Baker Robbins 2020, Klees et al. 2020). That technical reason was that no clear evidence could be established on a relationship between "teacher voice" in the shaping of teacher policies and "student outcomes" (World Bank 2013, p. 24). Baker Robbins illustrates that "there are other policies that are in line with World Bank ideology that are included in spite of unclear or contested research." She mentions in this connection policy goal no. 8, Motivating Teachers to Perform, and that dropping this goal would not sit well with the World Bank's predilections for performance-related pay (Baker Robbins 2020, pp. 2-3). Excluding teachers means, in Robertson's words, that "a select group of economists of education have drawn up the benchmarking and data collection protocols aimed at generating an evidence base on teacher policies and student performance" (2012 p. 584). The exclusion of teachers as a source of evidence is clearly more pronounced in the case of SABER-Teachers than in the case of TALIS. In a detailed article in the *Comparative Education Review* in 2020, Klees et al examine a range of shortcomings in SABER-Teachers and conclude that:

> Instead of being presented as those who should have autonomy, decision-making authority, and a level of control over their professional responsibilities, in the SABER-Teachers' framework, teachers are presented as those who should be monitored, led, and motivated by others (p. 54).

On this understanding of matters, research is something that is performed on teachers; something that is done to them, with little regard to the experienced realities of their practice. Such a standpoint ignores the possibility that research could be something much more fertile: something to which teachers might become committed as participants with valuable funds of evidence to offer on the fulness of what happens in classrooms.

Through the eyes of others

What can be done about the routine failure of international large-scale assessments (ILSAs) to include the more illuminating sources of evidence where the actual quality of education is concerned? In addressing this question I would like to call on the work of two East Asia-based researchers, Jeremy Rappleye (Japan) and Hikaru Komatsu (Taiwan), who have studied ILSAs in-depth from a comparative and historical perspective. In recent years they have published a succession of research papers containing incisive but constructive critiques of ISLAs. Their research highlights a parochialism in what presents itself as the global, or universal thinking of bodies like OECD and the World Bank. Such parochialism is difficult to detect, unless, as Hannah Arendt memorably put it, one can succeed in "training the imagination go visiting" (Arendt, 1982, p. 43), This involves a self-critical as well as a critical perceptiveness: the disciplined effort to see with one's own eyes from someone else's standpoint.

The research of Rappleye and Komatsu supplies frequent and productive examples of this kind of effort. Their analyses discovered anomalies associated with Japan's consistently high performance in PISA assessments. Firstly, PISA data itself showed that, despite Japan's high scores, study time among Japanese students was below the OECD average. Rappleye and Komatsu were aware that prevalent Western perceptions of Japanese schooling were badly out of date. It was the anomaly itself however – higher performance with less study time – that really demanded their further scrutiny. Secondly, there was a divergence between the kind of active learning that PISA tests are designed to reward and the pedagogies prevalent in Japan. Rappleye and Komatsu point out that these pedagogies feature "strong authority by a teacher matched with students' trust, obedience and persistence" (Rappleye & Komatsu 2021 pp. 251). This divergence should have resulted in lower, not higher, scores for Japanese students. The anomaly remains puzzling – unless there were other decisive factors at play that the supposedly neutral PISA research instruments failed to capture. This is in fact what they discovered. The teaching approaches and study habits that would better explain the Japanese results were rooted in intricate pedagogical inheritances on which there was a considerable research literature in Japanese, but little in English (p. 251). As their investigations progressed, what became increasingly clear to the researchers was the presence of unconscious cultural biases in the research frameworks not just of PISA, but of the ISLAs more widely. On this issue they conclude:

---

[5] In this connection the 2013 "Framework Paper" says only that "The framework has been developed by SABER-Teachers – a work program within the Human Development Network's Education Sector of the World Bank" (World Bank 2013, p. 6). Presumably, this was "the core team of SABER" whose eight names and affiliations are mentioned in the Acknowledgements at the start of the document. They include economists, academics, senior World Bank staff and consultants. Evidently there was no participation by teachers, school leaders or teachers' representatives.

ILSAs are almost always understood using the cultural 'sense making' frameworks of the Western world (particularly Anglo-American). We have rarely, if ever, come across non-Western theoretical frameworks, either in the creation of the PISA questions/questionnaires or in the subsequent scholarly analyses (p. 251).

We have already noted the exclusion of teachers' voices in the design of the PISA research procedures. The last point in the passage just quoted calls attention to a further, albeit unwitting, exclusion: the voices of educational expertise and experience from non-Western educational traditions. This kind of exclusion is compounded when the expansion of the OECD's work into the field of students' well-being is considered. On examining the first of the OECD's reports on student well-being (OECD, 2018), Rappleye and Komatsu were confronted with a disconcerting finding.

While there might be many reasons to applaud this shift away from cognitive indicators, we were instantly troubled by the fact that at the very bottom of the OECD's 50 nation well-being league table were the East Asian countries, including Japan and Taiwan. Both of us had lived in Japan and Taiwan, and our experience told us that people were not necessarily unhappy (p. 252).

Rappleye and Komatsu consulted prominent cultural psychologists from Japan and the US to find out to what extent cultural factors played a part in conceptions of well-being in different countries. These investigations discovered that the instrument used by the OECD for gathering evidence on students' well-being was based purely on the "Cantril Ladder of Life Satisfaction." This instrument, they point out, focused mainly on individual well-being and was unable to capture the features of well-being that are most significant in East Asian cultures. These latter include "a much stronger emphasis on relationships, attunement to feelings of others, management of personal thought, and staying calm" (p. 252). Here the authors also call attention to instruments such as the East Asian "Interdependent Happiness Scale", built and tested by cultural psychologists to counter the bias in supposedly universal instruments that conceive of well-being mainly as something pertaining to individuals (p. 252).

Any probing reflection on the example of the East Asian countries just considered will bring to the foreground not just the obscuring of the East Asian experience but of educational experience itself. To understand what makes educational experience properly fruitful in a Japanese, or Taiwanese context, we need to gather evidence on *what is actually happening* while that experience is taking place. This involves focusing on developments not just in cognitive progress, but also in the collateral learning that unavoidably happens at the same time, and that John Dewey highlighted as being of major importance in the long run (1938/1997 p. 48). Collateral learning includes the *attitudes* the students learn to take to their work, to their teacher(s), to their fellow students and their life context more widely. It also includes *practices* of study in which the students become habituated – competitive, collaborative, investigative, conformist, or whatever. What holds true, in these respects, of educational experience in Japan or Taiwan holds equally true of Germany, Argentina, Zambia, the United States, or any other country. Unlike many of the conceptual categories of ISLA research instruments, collateral learning can properly be said to be universal. It is something that inescapably happens when learning of any kind – fruitful, harmful or other – takes place in human experience.

Evidence on what actually happens

The question raised at the start of the previous section can now be asked with a sharper focus: What can be done about the routine failure of international large-scale assessments (ILSAs) to include the more illuminating sources of evidence where the actual quality of education is concerned? Any adequate answer must focus initially on highlighting to the designers of ISLA research the necessity of trying to capture educational experience in its manifold fullness. A meaningful commitment to research integrity requires this as an essential first step. Rappleye and Komatsu urge moves of this kind, although they add, "we do not expect that OECD analysts will actually listen" (2021 p. 253). Hopefully, their insights will be acknowledged in PISA's expanding plans. To their credit moreover, Rappleye and Komatsu distance themselves from research efforts that are "for" or "against" PISA. They raise instead the nonpartisan possibility of "what might come into view if we were to transcend those positions" (p. 246).

Secondly, tackling the "what can be done?" question in a practical manner calls for identifying, crediting, and urging further any initiatives by ISLA authorities that expand the reach of evidence-gathering beyond the limits imposed by a behavioural sciences approach. The extension of PISA tests to include "creativity and critical thinking", "students' well-being", and "global competence" point ultimately in this direction. But the research approaches adopted so far in gathering evidence in these additional areas remain within a behavioural sciences template. To what extent they can continue to do so if the OECD's own educational research priorities toward 2030 are to be successfully advanced now becomes a key issue.

Here it is instructive to consider the remarks of the OECD's Director for Education and Skills, Andreas Schleicher, in his Forewords to PISA reports, or in announcing new developments in OECD educational research. In a 2019 address to a conference in London, Schleicher stressed a range of OECD's educational priorities, offering concise remarks on each. These priorities include: observing creativity in the classroom; promoting changes in learning environments to nourish creative and critical thinking; strengthening teachers as creative designers; cultivating the capacity to manage complex information; working capably with people from diverse backgrounds and cultures; recognising the importance of "social and emotional skills" in a classroom climate. (Schleicher, 2019). Schleicher's

list of priorities was longer, but this sample of prominent items highlights the increased scope and diversity of the task the OECD seeks to pursue.

Regarding the first item, observation studies of creativity in the classroom can be pursued by researchers-as-observers, by teaching colleagues, by teachers themselves, or by some combination of these. Such studies are essentially qualitative forms of research, as distinct from approaches grounded in behavioural sciences. The latter approaches may *infer* somethings about an experience of learning. But they cannot capture its quality: its emotional tenor, its vitality or dullness, its intellectual vigour or weakness, its effects for better or worse on later experiences, and so on. Observation studies are likely to include some video recordings of teachers and students engaged in innovative practices of teaching and learning. The recordings provide invaluable forms of evidence on what actually happens: on challenges encountered, breakthroughs achieved, wrong turnings taken, reactions from students, revisions to original plans and so on.

These points can be expanded, and find fuller application, where topics like the others on the OECD priorities list are concerned. For instance, action research approaches are particularly suited to goals like building learning environments that nourish critical thinking or strengthening teachers' capacities for pedagogical leadership. All forms of action research seek to bring about some enduring changes in practice. They involve introducing new initiatives, monitoring these closely over a definite period of time, reviewing developments regularly and making revisions to practice in the light of what criticisms and feedback have brought to light. The common ground with lesson study, a discipline of Japanese origin, should not be missed here. The involvement of teachers becomes central – as researchers, as originators of fresh ideas, as colleagues critiquing their own efforts, and not least as leaders of educational innovations; or to use Schleicher's words, as creative designers. While observation studies can be once-off recordings of specific educational practices, action research initiatives are of longer duration and reach deeper into practitioners' work. They track closely the fortunes of one or more innovations over a definite period, ranging typically from a school term to two school years. They also necessarily produce an inclusive account, supported by various forms of record, of successes, failures and insights gained.

A particular feature of action research is its systematic interrogation of the changes, both expected and unexpected, that come about following the introduction of one or more new pedagogical initiatives. Such changes can be observed and recorded in both the overt and collateral dimensions of learning, thus combining to provide a threefold yield of evidence: i.e., evidence of changes in the students' *achievements in* learning, in their *attitudes to* learning and in their *practices of* learning. Discerning and appraising the significance of such changes calls for regular feedback from the students. In addition to a range of everyday response strategies this includes, crucially, structured interviews at regular intervals with representative groups of students to probe their responses to the different paths the changes have led their learning to follow. Unlike questionnaires, where answers are radically truncated, and channelled into pre-designed categories, well-designed regular interviews allow students to speak in their own voice, to develop their thinking, and to defend and amend their standpoints. Students thus become contributors to the items that feature on the interview agenda, and to how that agenda is pursued. While enabling students to elaborate their views more fully and coherently, the interviews provide teachers with valuable opportunities to delve into the real currents of students' experiences and to become adept in learning from these. Where such encounters become habitual through persistent pedagogical effort, they build the kinds of relationships that make learning environments more fertile and that enable a diversity of outlooks to work together in some productive concord (Malone & Hogan 2020; Hogan & Malone 2023).

Advancing Beyond Misperceptions

Clearly, the approaches reviewed in the account just given involve teachers and also students in significant ways in research activity. Teachers here are afforded opportunities for influential and fruitful participation research work. Also, opportunities arise to achieve a vibrant ownership of their own practice. In the case of students, they are afforded a stronger stake in their own schooling by coming to appreciate the venturesome, yet the safe nature of genuine learning.[6] This kind of experience can also do much to overcome some historically commonplace misconceptions of teaching among the cultures of students: e.g., as a body of rules and procedures to be imposed and followed, or perhaps resisted. In short, participating in these forms of research strengthens teachers' sense of authorship of their own work and their capacity to partake in informed policy debates.

We begin to realise now just how much is omitted in the kind of research and evaluations characteristically conducted by ISLAs, and to appreciate the necessity for including approaches like those reviewed above in the design of ISLAs. Notwithstanding this necessity however, questions may still be raised about the scientific standing of such research approaches. Such questions usually focus on things like validity, reliability and universalisability (or transferability) of findings. For instance: (a) How does the criterion of validity apply in research if precise numerical indicators are replaced by stories and narrative accounts? (b) Qualitative research studies and case studies are tied to specific times, geographic locations and socio-cultural circumstances. To what extent can they meet the

---

[6]   In fact, Schleicher's Foreword to the report on TALIS 2018 strongly recommended such ownership for both teachers and students: "But simply perpetuating a prescriptive model of teaching will not produce creative teachers: those trained only to reheat pre-cooked hamburgers are unlikely to become master chefs. By contrast, when teachers feel a sense of ownership over their classrooms, and when students feel a sense of ownership over their learning, that is when productive teaching takes place" (OECD, 2019c, p. 4).

scientific demands of replicability, and thus reliability? (c) Where such studies produce findings of what works, can this success be transferred across geographical and cultural boundaries? These three questions provide a rough summary of the reservations that are typically expressed about qualitative studies. Dealing with them should clear up misperceptions that might underlie the questions themselves and should also remove any reasons for lingering doubts about the soundness of this kind of research, when properly conceived and pursued.

In response to the first question, it needs to be stressed that validity in all forms of qualitative research is centrally concerned with the appropriateness, the adequacy and the credibility of the evidence that is sought and gathered. Here, accordingly, the germane question is: Has the research sought to include all pertinent sources of evidence to the question at issue, and is that evidence convincing beyond a reasonable doubt? Because experiences, desires, motives, aspirations and the like are central to the subject matter of the research, it would be mistaken from the start to seek here the theoretical regularities, or predictive sameness, sought by empirical sciences, viz. natural sciences and behavioural sciences. In response to the second question, clearly no form of research in the human sciences (behavioural or otherwise) can reproduce the exact conditions of the initial situation in second and later cycles of research. In fact, any demand of this kind is self-defeating as it amounts to denying the historical character of human experience. In this instance the germane question is: What steps have been taken to ensure that the evidence presented has been scrutinised for coherence and interrogated for robustness, to ensure its reliability? In response to the third question – the transferability of what works – the notion of transferability itself embodies an error here. To hold that what has worked impressively in a school in a particular socio-cultural setting can be transferred to other settings is to acquiesce in a manifold and mistaken assumption. It is to assume, for instance, that countless contingencies of unfolding experience can be identified, and stabilised as controllable variables, to make any notion of transfer in a scientific sense serviceable in this context. The more pertinent – and promising – notion to invoke here is that of *resonance*, not transferability. To put the proper question in practical terms: to what extent do the impressive fruits of a research initiative in school X resonate with educational practitioners in other settings? "Practitioners" here can be taken to mean teachers, school leaders, educational researchers, schools inspectorates, support services, and so on; in short, all those who might be included under the term "educational professionals."

An action research project that has succeeded impressively in a particular learning environment, or a case study that yields an illuminating account of similar advances, can quicken the interest of practitioners nearby and much farther afield, often strikingly. This quickening of interest is not born of any desire to reproduce the same model elsewhere. Rather, it springs from a keen perceptiveness and attunement on practitioners' part that policy-borrowers repeatedly lack. It is recognisable in remarks like: "variants of key initiatives they have progressed here could be very promising in our school"; "they have ingeniously got around the timetabling and teamwork issues that have been stumbling blocks for us"; "I always thought lesson study meant extra work, but they've made it an inspiration here"; and so on. Insofar as educational researchers – and the professional cultures of educational research – are strangers to such perceptiveness and attunement, they are also strangers to the notion of resonance and its energising effects. There is a likelihood that such deafness will be an institutionalised feature of international large-scale assessments unless the researchers "go visiting", as suggested earlier, and dwell attentively among teachers. The World Bank's SABER-Teachers programme seems to have closed the door on this possibility, at least for now but hopefully not for good. The OECD, by contrast, seems to be moving toward a more serious involvement of teachers in expanding its priorities towards 2030 (OECD 2018, pp. 4, 6; OECD 2019a, p. 3).

This expansion of the OECD's research scope to include realities residing in the experience of students and teachers requires a simultaneous advance in its evidence-gathering rationale and procedures. Such restructuring is essential to enable the manifold character of quality in education to show itself more fully and more naturally. By way of recap then, in addition to what the OECD research instruments have traditionally gathered, such disclosure would include evidence from qualitative studies on things like the following: the quality of educational relationships between teachers and students; influential factors in achieving participatory learning environments; promising and less promising pedagogical ideas; creative and restrictive practices of teaching and assessment; and so on. Such themes can only be touched on lightly, and distantly, by questionnaires, but they are central to the quality of educational practice: i.e., to the quality of work pursued daily by teaching practitioners and school leaders. OECD reports occasionally cite qualitative studies in passing (e.g., OECD 2019b, p. 88), but to date the OECD has excluded from its evidence-gathering strategies any of the primary means of capturing the core and quality of educational experience. Nevertheless its international prominence now puts it in a key position to undertake a range of potentially highly illuminating qualitative studies in various parts of the world, and to publish the findings globally. It might thus identify some fresh and promising pathways for its own expanding work. In the process, moreover, it may give pause to some of its many critics and set an example for ISLAs more widely.

References

Ahonen, A. K. (2021) Finland: Success Through Equity – The Trajectories in PISA Performance, in N. Crato, ed. *Improving a Country's Education: Pisa 2018 Results in 10 Countries*, pp. 121-136. Cham: Springer https://doi.org/10.1007/978-3-030-59031-4_6

Arendt, H. (1982) *Lectures on Kant's Political Philosophy*. University of Chicago Press

Baker Robbins, M. (2020) *Teachers in the World Bank's SABER*, #SABERexposed https://www.ei-ie.org/en/item/23129:saber-exposed-teachers-in-the-world-banks-saber-by-melanie-baker-robbins

Bassey, M. (1992) Creating Education through Research, *British Educational Research Journal*, 1992, Vol. 18, Issue 1, pp. 3-16. https://doi.org/10.1080/0141192920180101

Dan, W. (2023) *Families Are Abandoning China's Rural Schools. Here's Why That Matters. The Sixth Tone: Fresh voices from Today's China*. Available at https://www.sixthtone.com/news/1012872

Dewey, J. (1997) *Experience and Education*. Touchstone. [Original published in 1938]

Gillard, D. (2018) *Education in England: a history*. Retrieved of http://www.educationengland.org.uk/history/

Hessler, P. (2022) A Teacher in China Learns the Limits of Free Expression, *New Yorker*, 9th May 2022. https://www.newyorker.com/magazine/2022/05/16/a-teacher-in-china-learns-the-limits-of-free-expression

Hogan, P. & Malone, A. (2023) Tackling a Lingering Infirmity: On the nature and warrant of action research in education, *British Educational Research Journal*, 49(3), pp. 439-454. https://doi.org/10.1002/berj.3849

Jones, G. & Wu, Y. (2021) *The Business of K-12 Education in China, Harvard Business School*, Working Paper 22-022 https://www.hbs.edu/faculty/Pages/item.aspx?num=61340

Klees, S. J. et al. (2020) The World Bank's SABER: A Critical Analysis. *Comparative Education Review*, 64(1), 46-65. https://doi.org/0010-4086/2020/6401-0003

Komatsu, H. & Rappleye, J. (2021) Rearticulating PISA, *Globalisation, Societies and Education*, 19(2), 245-258. https://doi.org/10.1080/14767724.2021.1878014

Malone, A. & Hogan, P. (2020). Evidence and its Consequence in Educational Research. *British Educational Research Journal*, 46(2), pp. 265-280. https://doi.org/10.1002/berj.3580

Meyer, H-D. & Benavot, A. (2013) *PISA, Power and Policy: the emergence of global educational governance*. Oxford: Symposium Books.

OECD (2018) *The Future of Education and Skills: Education 2030*. OECD Publishing https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf

OECD (2019a) *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing https://www.oecd.org/publications/pisa-2018-results-volume-i-5f07c754-en.htm

OECD (2019b) *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. OECD Publishing https://www.oecd.org/publications/pisa-2018-results-volume-iii-acd78851-en.htm

OECD (2019c) *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. OECD Publishing https://www.oecd.org/education/talis-2018-results-volume-i-1d0bc92a-en.htm

OECD (2020) *PISA 2018 Results (Volume VI): Are Students Ready to Thrive in an Interconnected World?* OECD Publishing https://www.oecd.org/publications/pisa-2018-results-volume-vi-d5f68679-en.htm

Park, J. (2022) Beijing's Mission to Make China's Youth More Chinese, *Geopolitical Monitor,* 22 January 2022. https://www.geopoliticalmonitor.com/beijings-mission-to-make-chinas-youth-more-chinese/

Rappleye, J. (2020) Comparative education as cultural critique. *Comparative Education*, 56(1), pp. 39-56. https://doi.org/10.1080/03050068.2019.1701247

Rappleye, J., Komatsu, H., Ukiko, U., Kuba, K., Markus, H. (2020) Better policies for better lives?: constructive critique of the OECD's (mis)measure of student well-being*, Journal of Education Policy*, 35(2), pp. 258-282. https://doi.org/10.1080/02680939.2019.1576923

Rappleye, J. & Komatsu, H. (2021) Is knowledge capital theory degenerate? PIAAC, PISA, and economic growth, *Compare: A Journal of Comparative and International Education*, 51(2), pp. 240-258. https://doi.org/10.1080/03057925.2019.1612233

Robertson, S. L. (2016) The Global Governance of Teachers' Work. In Mundy, K., Green, A., Lingard, B. and Verger, A (Eds) *The Handbook of Global Education Policy* Wiley, pp. 275-290.

Robertson, S. L. & Sørensen T. (2018) Global transformations of the state, governance and teachers' labour: Putting Bernstein's conceptual grammar to work. *European Educational Research Journal*, 17(4), pp. 470-488. https://doi.org/10.1177/1474904117724573

Schleicher, A. (2019) Andreas Schleicher on Creativity and Critical Thinking. Video of a presentation in London, 24 September 2029 https://www.youtube.com/watch?v=M3zEig3FpSw

Sjøberg, S. (2019) The PISA-syndrome – How the OECD has hijacked the way we perceive pupils, schools and education. *Confero,* 7(1), pp. 34-88. https://doi.org/10.3384/confero.2001-4562.190125

World Bank (2013) *What Matters Most for Teacher Policies: A Framework Paper*, SABER Working Paper Series, Number 4, April 2013. The World Bank http://wbgfiles.worldbank.org/documents/hdn/ed/saber/supporting_doc/Background/TCH/Framework_SABER-Teachers.pdf

Yiu, L. & Yun, L. (2917) China's Rural Education: Chinese Migrant Children and Left-Behind Children, *Chinese Education and Society*, 50(4) pp. 307-314 https://doi.org/10.1080/10611932.2017.1382128

Zhao, Y. (2020) Two Decades of Havoc: A synthesis of criticism against PISA, *Journal of Educational Change*, 21(22), pp. 245-266. https://doi.org/10.1007/s10833-019-09367-x