Chapter 15

Understanding the Policy Influence of International Large-Scale Assessments in Education

David Rutkowski, Greg Thompson, and Leslie Rutkowski

Abstract International large-scale assessments (ILSAs) require national governments to invest significant resources in both time and money. With such investment national governments expect the results of ILSAs to provide policy and research communities with evaluative information on their educational system. Armed with this evaluative information, policymakers in many participating countries have used the results to stimulate reform. However, systematically tracking this influence and judging the validity of the claims has proven difficult for both the policy and research communities. There can be an erroneous expectation in the wider education community, and sometimes amongst policymakers themselves, that ILSA data automatically suggests policy solutions. Because of this error, a means for better systematizing the policymaking process responding to ILSA results is required. The model developed here can assist the policy and research community to better understand whether ILSAs are providing valid evidence to support their use in policy formation enactment and can be used to analyze ongoing consequences of that influence. Two worked examples demonstrate the utility of the model.

Keywords Evaluation · Large-scale assessment · Policy impact · Policy influence

D. Rutkowski (✉) · L. Rutkowski
Indiana University, Bloomington, IN, USA
e-mail: drutkows@iu.edu

L. Rutkowski
e-mail: lrutkows@iu.edu

G. Thompson
Queensland University of Technology, Brisbane, Australia
e-mail: g6.thompson@qut.edu.au

15.1 Introduction

Countries and educational systems participate in international large-scale assessments (ILSAs) for a variety of reasons, including educational system monitoring and comparison. As taking part in an ILSA requires the investment of significant money and time (Engel and Rutkowski 2018), it is important that countries derive value and use from their participation. One way to justify participation is to demonstrate the ways in which ILSA results are used (or are claimed to be used) as a lever for policy formulation and as a means to change policy trajectories in order to improve educational outcomes. That is, successfully attributing policy changes to ILSA results can be seen as a rationale for continued participation, building a case for the further outlay of time and money. It makes sense that testing organizations want to argue to their stakeholders that the resources spent on their assessment tools are worthwhile. However, demonstrating how ILSAs impact systems and nations remains difficult. First, this is because policymaking itself is an emotive and politicized domain informed more by what can be sold to the public for electoral success than what research might suggest could be the best direction (see for example Barber et al.'s (2010) concept of "Deliverology"). Second, even if an association is identified, it remains challenging to establish the direction of the relationship or the amount of influence ILSAs had in any policy change that resulted. In other words, it is difficult to prove the counterfactual that the policy change would not have occurred in the absence of the ILSA. Third, evidence of the influence of ILSAs on policy can be inflated or misleading. For example, there are a number of cases demonstrating that governments made use of ILSA results simply to justify policy reforms that were already set to be implemented (Gür et al. 2012; Rautalin and Alasuutari 2009; Takayama 2008).

In this chapter, we explore how participation in ILSAs, and the subsequent results, could reasonably be said to "influence" policy. In other words, how can ILSA results, or any proposed policy attributed to those results, be shown to be the reason that a policy or policies change? In complex systems the attribution of singular causes that can explain an altered state of affairs is always difficult because of the multiple forces at work in that system. Further, few would argue that the ILSA-policy nexus is easy to understand given: (1) the complex social, cultural, historical, economic and political realities within each system; (2) the complexities between systems; and (3) the limitations of what the tests themselves can measure on any given topic and the difficulty in measuring policy change. This, then, leads to a key problem that confronts policymaking communities; how can system leaders properly understand and manage ILSAs' influence on their system?

A second question that drives this chapter asks what are the overall consequences of participating in an ILSA? This question works from the premise that there are always intended and unintended consequences when an ILSA has influence at the national level. When ILSA results are used to set policy goals or are the impetus for educational change this creates the conditions for a range of consequences. For example, implementing a particular kind of science curriculum as the result

of middling science performance on an ILSA will have consequences that might include money spent training teachers and abandoning other teaching approaches, and so on. Correspondingly, where a system's leaders become convinced that doing well on rankings will lead to better educational outcomes, a variety of perverse incentives can emerge, resulting in attempts by a range of stakeholders to "game" the test. This is evidenced through the multitude of high stakes testing cheating that took place in the USA (Amrein-Beardsley et al. 2010; Nichols and Berliner 2007) and the fact that some countries participating in ILSAs are removed from the results for "data irregularities," including being too lenient when marking open-ended questions, which resulted in higher than expected scores (OECD [Organisation for Economic Co-operation and Development] 2017). Stakes at the student and school level may remain low; however, at the national level there is growing evidence that the stakes of participation are high.

One challenge in the ILSA-policy nexus that encompasses both understanding ILSA influence and the consequences of that influence is that there is rarely, if ever, systematic analysis undertaken of the data in the context of the whole system. When ILSA data is released media, policymakers, and other stakeholders tend to sensationalize and react, often quickly, without a full accounting of the evidence (Sellar and Lingard 2013). To underline this problem, we present two cases that highlight the problem of simply claiming ILSA "influence." Subsequently, we describe a model as a means for better systematizing how influence ought to be attributed to policy processes as a result of participation in ILSAs and the publication of subsequent results. This model that can assist the policy and research communities to better understand whether ILSAs are providing valid evidence to warrant their influence on educational policy formation and debates and to analyze the consequences of that influence.

15.2 Impact, Influence, and Education Policy

To understand influence, we first differentiate between what we view as ILSA's impact on policy (which is hard to demonstrate) and ILSA's influence on policy (a concept that is still difficult but easier to demonstrate than impact). For the purposes of this chapter, we define impact as a *difference in kind* while influence is defined as a *difference in degree*. To show policy impact, we would have to isolate an ILSA result and prove that this caused a significant shift in a policy platform. We should expect to see clear evidence that there was a rupture, such as a new national policy direction being caused by ILSA data. However, making causal claims such as ILSA *X* caused Policy *Y* requires a methodological framework that may simply not be possible because of the complexity of most national systems. Moreover, many of the claims made in reports that attribute impact to an ILSA result exemplify what Loughland and Thompson (2016) saw as a post hoc fallacy at work rather than identifying a causal mechanism. They explained that, "when an assumption is made based on the sequence of events—so the assumption is made that because one thing occurs after

another, it must be caused by it” (Loughland and Thompson 2016, pp. 125–126). This is particularly true of ILSA results where the data often appear to be used to maintain current policy directions in the interests of political expediency, even where the data suggests this may be having unhelpful consequences. Finally, impact is notoriously difficult to demonstrate because education policy agendas are often politically rather than rationally decided (Rizvi and Lingard 2009). As such, even if ILSAs provided perfect information they will only be one factor among many that influence policymaking. For these reasons, when ILSAs are mentioned together with policy impact, we suspect that it would be better to frame this in terms of evaluating the influence that ILSAs have on policy agendas and trajectories within given contexts.

Policy influence can be viewed as the use of ILSA results to buttress or tweak policy settings that already exist. However, establishing exactly what constitutes influence remains difficult for a number of reasons. For example, similar to impact, much of the policy literature fails to define influence (Betsill and Corell 2008). The lack of a clear definition in the literature leads to (at least) three problems. First, without an established definition for influence, it is difficult to determine the type of evidence needed to demonstrate influence. This is a particular problem, as ILSA literature tends to report evidence of influence on the basis of the particular case at hand without consideration of wider application and with a pro-influence bias, rarely examining evidence to the contrary (e.g., Breakspear 2012; Schwippert and Lenkeit 2012). Second, to make a strong case that ILSAs influence policy, some consensus as to what data should be collected to mount a sufficient argument is needed. Finally, this lack of definition makes cross-case comparisons potentially unstable because different stakeholders risk measuring different things and claiming them as demonstrating influence. In other words, if each claimant develops their own ideas around influence and collects data accordingly, they may simply end up comparing different things.

In this chapter we borrow from Cox and Jacobson (1973), who defined influence as the “modification of one actor’s behavior by that of another” (p. 3). With this definition we take a broad view and define ILSAs as policy actors that are as involved in creating meaning in a variety of contexts as much as they are created artefacts of organizations or groups of countries. As an actor, an ILSA represents multiple interests and ambitions, and intervenes in social spaces in a variety of ways. For example, in the case of OECD’s Programme for International Student Assessment (PISA) study, the results are intended to serve at the behest of the OECD’s and member countries’ policy agendas (OECD 2018). However, the declared explicit use of ILSAs for policy modifications are less clear for the International Association for the Evaluation of Educational Achievement’s (IEA) Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), perhaps because of the IEA’s history as a research rather than policy focused organization (Purves 1987). That said, ILSAs are evaluative tools partly sold to “clients” such as nation states with an assumption that the assessment will help judge the merit and worth of a system by measuring performance. Demonstrating positive impact and/or influence to those jurisdictions who have paid to have the

tests administered would make commercial sense regardless of the methodological concerns outlined above. Rightly so, testing organizations and countries alike want to know whether the tests they design and administer are having a positive influence on systems, at least partly because organizations then have a compelling narrative to sell to other potential participants and countries have a legitimate reason for participating.

We do not, however, want to ground our discussion of ILSA influence on policy in a naive caricature of ILSAs as authoritarian tools thrust onto nations by some evil council of neoliberals such that nations have little choice but to participate (the critical ILSA literature is full of this). In most cases, nations willingly sign up to testing regimes because they have come to believe that ILSAs offer their systems something that they either do not have or should have more of. If coercion is “the ability to get others to do what they otherwise would not” (Keohane and Nye 1998, p. 83), then “influence is seen as an emergent property that derives from the relationship between actors” (Betsill and Corell 2008, p. 22). Influence differs from coercion. There are obvious power imbalances in regards to intergovernmental organizations like the OECD, where more powerful countries have larger voices; however, that does not always result in coercive leverage towards less powerful actors. In other words, ILSAs may have the potential to be leveraged over educational systems to compel actor behavior, but that is not always the case and most national systems choose to participate, as evidenced by the growing number of participants that are self-electing to join the studies.

Here we present two cases that illustrate the policy influence of ILSAs. We chose these cases because they demonstrate possible overclaiming that ILSA results influenced change and evidence of ILSA influence that resulted in an unusual policy.

15.3 Policy Influence?

15.3.1 Case 1: PISA Shocks and Influence

Given the explicit goal of the OECD to inform national policy of member nations there is a considerable amount of research concerning the policy influence and impact of its flagship educational assessment PISA (Baird et al. 2011; Best et al. 2013; Breakspear 2012; Grek 2009; Hopflins et al. 2008; Kamens et al. 2013). PISA-inspired debates have resulted in a range of reforms including re-envisioning educational structures, promoting support for disadvantaged students (Ertl 2006) and developing new national standards aligned to PISA results (Engel and Rutkowski 2014), to name a few. The term “PISA shock” is now commonly used to highlight participating countries that were surprised by their sub-par PISA results and subsequently implemented educational policy reforms. Perhaps the most notable of these shocks occurred in Germany after its initial participation in PISA 2000. In response to lower than expected PISA scores, both federal and state systems in Germany implemented significant educational reforms (Ertl 2006; Gruber 2006; Wadow 2009). However,

Germany was not alone, and other countries, including Japan (Ninomiya and Urabe 2011) and Norway (Baird et al. 2011), experienced PISA shocks of their own.

In general, “shocks” attributed to ILSAs tend to be focused on PISA results rather than other studies. Notably, Germany, Norway, and Japan participated in the TIMSS assessment five years prior with similar results (in terms of relative rankings) to PISA (Beaton et al. 1996), yet this resulted in significantly less public discourse and little policy action. Of course, the perceived lack of a TIMSS shock could be for a variety of reasons. First, it is possible that the IEA simply does not have the appetite and/or political muscle to influence policy debates, leaving any discussions of results to academic circles. Second, the idea of a PISA shock may be misleading, representing an engineered discourse rather than a true social phenomenon. For example, Pons (2017) contended that much of the academic literature claiming that there is a PISA shock is biased because it contributes to a particular representation of what effect PISA “is expected to produce in conformity with the strategy of soft power implemented by the OECD” (p. 133). Further, similar to our discussion of the term influence, PISA shock is never fully conceptualized in the literature, making it difficult to compare and assess within and across systems. Pons (2017) further contended that assessing the effects of PISA on education governance and policy is difficult because the scientific literature on PISA effects are heterogeneous and fueled by various disciplines and traditions that ultimately lead to findings corresponding to those traditions.

15.3.2 Case 2: An Australian Example, Top Five by 2025

In 2012, the Australian Federal Government announced hearings into the Education Act 2012, which was subsequently passed and enacted on the January 1, 2014 (The Parliament of the Commonwealth of Australia 2012). This referred specifically to five agendas that were linked to school reform. These five reform directions were “quality teaching; quality learning; empowered school leadership; transparency and accountability; and meeting student need”. These five reform directions were to improve school quality and underline the commitment of the Federal Government to have a system that was both high quality and high equity. The Act went on to outline that the third goal was:

“...for Australia to be ranked, by 2025, as one of the top 5 highest performing countries based on the performance of Australian school students in reading, mathematics and science, and based on the quality and equity of Australian schooling” (The Parliament of the Commonwealth of Australia 2012, p. 3)

The Explanatory Memorandum that accompanies this Act includes in brackets, “(These rankings are based on Australia’s performance in the Programme for International Student Assessment, or PISA.)” There are a number of curious things about this legislation that binds the Australian education system to be “top 5 by 2025.” The first of these is that it shows, in the Australian context, that PISA and no doubt other

ILSAs have had an influence on policymakers. But the nature of the influence remains problematic, focused more on national rankings outcomes rather than considering what PISA tells Australia about its system and the policy decisions that have been made. While generic references to quality teaching and so on might work as political slogans, the reality is that they contain no specific direction or material that could ever be considered as a policy intervention or apparatus.

Second, it is curious that Australia legislated for a rank rather than a score or another indicator of the type that PISA provides such as some goal regarding resilient students. This would seem to suggest that this is how PISA is understood by policymakers in Australia, as a competitive national ranking system of achievement in mathematics, science, and reading. It would be easy to lay this solely at the feet of the policymaker, but it is probably not helped by the way that the OECD itself presents PISA as rankings to its member nations. Third, it seems fairly obvious that this use of ILSAs opens a system up to perverse incentives.

In the case of being “Top 5 by 2025,” the influence of ILSAs falls short because it lacks intentionality. It also shows that those who are charged with making policy do not understand the data that they see paradoxically as: (1) determining a lack of quality and equity, and (2) clearly indicating what needs to be done as a result. In other words, without identifying what policy agendas in specific contexts could best respond to highlighted problems, ILSA data is often left to speak for itself as regards to what must be done within systems. The Education Act 2012 points to the impact and influence of PISA on Australian policymakers, yet paradoxically that impact and influence comes at the cost of policymaking itself.

Demonstrating the impact and influence of ILSAs on national policymaking is not the same as demonstrating that ILSAs are having a positive, or beneficial, impact or influence on policymaking. Legislating to be “Top 5 by 2025” is a prime demonstration of impact on policy that consequently opens the test up to perverse incentives. It is an absurd example, but should New Zealand outperform Australia on PISA, then invading them and taking over their country necessarily brings Australia closer to its goal. This neatly illustrates the problem of influence: how can society think about making the influence that ILSAs are having more useful than an obsession with rankings? This begins by considering how ILSAs might be used to hold policymaking to account, particularly at a time where “top down” accountability in most contexts seems to be about protecting policymakers from repercussions based on their policy decisions (see Lingard et al. 2015).

These two cases are illustrative in two ways. The first case of “PISA shock” shows that while influence is easy to claim, it is invariably linked to pre-existing interpretations and expectations. In other words, it appears that ILSAs are often used to buttress the preconceived policy frames rather than interrogating them. The second case shows that even where influence can be demonstrated, this does not necessarily improve policymaking nor does it improve the understanding that policymakers have regarding their system. Both cases exemplify the problem of influence. ILSAs may or may not influence change in systems and, where they do, the resultant change may be artificial, superficial, or downright silly. What is needed, then, are better tools to inform decision making through understanding, predicting, and evaluating influence.

This may go some way to help policymakers become more purposive in their use of ILSAs as evaluative tools.

15.4 A Model for Evaluating Influence

Oliveri et al. (2018) developed a model to assist countries in purposeful, intentional ILSA participation. Although the model was originally designed as a means for countries to evaluate whether their educational aims can be met by what an ILSA can deliver, it is generalizable for other uses. The model encourages intentionality by carefully considering whether claims that are made about what an ILSA can reasonably be expected to do are valid in a given country. Further, it helps establish a set of more valid interpretations of ILSA data for policymakers to use in their decision making. In our retooling of Oliveri et al.’s model, we use the same general structure (Fig. 15.1).

Using the model begins with a matching exercise between the influence attributed to ILSA results and what evidence the ILSA in question can provide to motivate changes that are said to be directly caused by ILSA results. In this step, the national system or other stakeholder must clearly articulate all the ways ILSAs have influenced or are anticipated to influence their educational system and the policy process. As an example, assume that country X scores lower than desirable results on the TIMSS grade 8 science assessment and that, as a consequence, policymakers in country X propose a policy that requires all science teachers to obtain an advanced degree

Fig. 15.1 Model for systematically understanding international large-scale assessment (ILSA) influence and the associated consequences for policymaking

(e.g., master's degree) by some set date. The matching analysis involves querying whether this policy can be attributed to TIMSS results. One line of argument might go like this: TIMSS provides results in science and teachers are asked about their level of education. Thus, initially, TIMSS appears to show evidence that science teachers with master's degrees on average teach classes with higher performance. This causal claim, then, is said to be initially consistent with the evidence that TIMSS can provide, setting aside formal causality arguments.

The next step in the process is a formal logic argument. The logic model (Fig. 15.2) is a derivative of Toulmin's (2003) presumptive method of reasoning. The process involves a claim supported by a warrant and additional evidence, which is often provided through a backing statement. In contrast, rebuttals provide counterevidence against the claim. The process allows for an informed decision to be made on whether and to what degree an ILSA can be reasonably said to have influenced a proposed or enacted policy. In the case of requiring master's degrees, the logic argument could proceed as follows. TIMSS was said to influence policymakers' decision that all science teachers should have a master's degree. The warrant for this decision is that better educated teachers produce higher average student achievement. The backing could be that in country X (and maybe other countries), teachers with master's degrees teach in classrooms with higher average TIMSS science achievement. Then, a possible rebuttal could be multifold. First, TIMSS does not use an experimental design. In the current example, teachers are not randomly assigned to treatment (master's degree) and control groups (education less than a master's degree). Thus, assuming that the teacher sample is strong enough to support the claim, a plausible explanation for this difference in country X is that teachers with master's degrees command a higher salary and only well-resourced schools can afford to pay the master's premium. A second plausible explanation is that only the very highly motivated seek a master's degree, which, rather than serving as an objective qualification is instead a signal of a highly motivated and driven teacher. A final decision might be that, given the observed associations (e.g., more educated teachers are associated with higher performance), country X decides to pursue the policy, thereby ignoring the evidence in the rebuttal.

Fig. 15.2 Logic model for evaluating plausibility of influence

Returning to our model (Fig. 15.1), an analyst could conclude that, in spite of alternative explanations for the observed achievement differences between classes taught by teachers with different education levels, TIMSS results influenced policymakers' decision to enact a policy requiring all science teachers to have a master's degree. As noted, this conclusion could be by degrees and the analyst should include the warrant and rebuttal as the basis for this conclusion. The final, and perhaps most important step in the process is to consider the consequences of attributing influence and enacting a policy change based on ILSA results. This is in the form of a conditional consequential statement (CCS). Again, considering the TIMSS example, the CCS might be as follows: if country X requires science teachers to earn a master's degree, then the educational systems in country X can expect mixed achievement results, given that attributing influence to TIMSS results is not fully supported. Of course, there could be other consequences (i.e., medium-term teacher shortages, overwhelming demand for teacher training programs, and so on). However, it is important to delineate these consequences from those that are directly attributable to the influence of the ILSA in the given setting. To highlight this point, imagine that TIMSS did use an experimental design that randomly assigned teachers to different training levels. Further imagine that the TIMSS results showed that teachers with master's degrees taught classes that consistently outperformed classes taught by teachers with bachelor's degrees. Then, without going through the full exercise, assume that the decision from the logic argument is that the ILSA influence is fully supported. Then, a different CCS could be that if country X requires science teachers to have master's degrees, country X can expect higher average achievement in science on future TIMSS cycles.

15.4.1 Case 1: Worked Example

Norway had lower than expected PISA 2000 results and one resultant policy change was to implement a national quality assessment system (Baird et al. 2011). In fact, the OECD's report *Reviews of Evaluation and Assessment in Education: Norway* explained that poor PISA results spurred Norwegian policymakers to "focus attention on the monitoring of quality in education" (Nusche et al. 2011, p. 18). Using our model, a supporting analysis might proceed as follows. The first step is a matching analysis. That is, can PISA reasonably provide the necessary evidence to drive an expansion of a national evaluation system? Norwegian policymakers used PISA data, which showed that Norwegian students were lower performing than other peer industrialized nations, even though spending per child was one of the highest in the world. Initially, this claim might be regarded as consistent. This, then, triggers an analysis through the logic model. PISA is regarded by the OECD as a yield study (OECD 2019), measuring literacy and skills accumulated over the lifespan. It takes place at one point in time, when sampled students are 15 years old. Assuming this OECD claim is reasonable, PISA outcomes are attributable to a lifetime of learning. Then, the warrant for implementing a national assessment system could be that PISA performance was low relative to industrialized peers, and a national assessment

system will help Norwegian policymakers understand why. The backing is that PISA measures the accumulated learning through to the age of 15 and underachievement can be linked to learning deficiencies at some point between birth and age 15. A rebuttal to this argument, however, is that PISA does not explicitly measure schooling or curriculum, but rather, the totality of learning, both inside and outside of school. A national assessment that is (and should be) linked to the national curriculum will not fully align to PISA and risks missing the source of the learning deficiencies that lead to underperformance. The original argument also relies on the assumption that lower than desirable performance in the 2000 PISA cohort will be stable in future cohorts.

A CCS in this case might be: given that PISA showed Norway's achievement was lower than its industrialized peers and that PISA is a yield study, implementing a comprehensive national assessment system could reasonably be attributed to PISA results. But the fact that PISA does not measure curriculum imposes challenges in assessing the educational system and improving PISA outcomes. Thus, Norway can expect mixed results in future PISA cycles from enacting a policy that dictates a national assessment system. Here, a fairly simple but systematic analysis indicates that PISA is a mediocre evidentiary basis from which to enact such a policy and, although speculative, Norway might have used PISA as justification for a policy that they already wanted to initiate. This claim is substantiated to some degree by the fact that Norway's performance in TIMSS in 1995 was also relatively low; however, no similar policy reforms were enacted.

15.4.2 Case 2: Worked Example

The "Top 5 by 2025" case can be used to illustrate another worked example. PISA results clearly influenced Australian policymakers' desire to climb the ranks in the PISA league tables. Here, then, is a clear consistency; Australia's ranking in PISA drove a desire to improve on that position. The logic model becomes a somewhat trivial exercise where the warrant is that PISA rankings show the relative ordering of Australia's 15-year-olds in mathematics, science, and reading. The backing, again somewhat trivial, is the evidence that higher scores map onto better achievement in these domains. A plausible rebuttal is that simple rank ordering changes are somewhat meaningless without considering measures of uncertainty. For example, if Australia moves up two or three places in the league table, this change might not be statistically significant. Nevertheless, the decision is relatively straightforward: PISA results can reasonably be attributed to influencing the decision to seek a top five position in the PISA league tables. However, a desire for improvement, as understood by a position on a ranking (top five) within a timeframe (by 2025), does little to demonstrate improved decision making or better understanding of policy settings and their impact. A desire for improved rankings is the opposite of influencing policy; in fact it may act as a barrier for making policy changes necessary for that improved ranking.

Then, a conditional consequential statement (CCS) might be: if Australia uses PISA results to influence a decision to move up the league table, then Australian schools can expect initiatives intended to drive improvement in mathematics, science, and reading. Certainly, as with any CCS, there is no guarantee that these consequences will happen. Perhaps policymakers will take no concrete action to realize the gains necessary to move into the top five. Further, downstream consequences also become important in this example. If initiatives to improve mathematics, science, and reading come at the cost of other content areas (e.g., art, civics, or history), second order consequences might involve narrowing of the curriculum or teaching to the test. Depending on the incentives that policymakers use to achieve the top five goal, there might be undue pressure to succeed, raising the risk of cheating or otherwise gaming the system (e.g., manipulating exclusion rates or urging low performers to stay home on test day). Clearly, this is not a full analysis of the potential consequences of such a policy; however, this and the previous examples demonstrate one means of using the model to systematically evaluate whether ILSA results can reasonably influence a policy decision and what sort of consequences can be expected.

15.5 Discussion and Conclusions

Ensuring that ILSAs do not have undue influence on national systems requires active engagement from the policy community to include an examination of the intended and unintended consequences of participation. Thus, while ILSAs can be an important piece of evidence for evaluating an educational system, resultant claims should be limited to and commensurate with what the assessment and resulting data can support. It is imperative to recognize that ILSAs are tasked to evaluate specific agreed upon aspects of educational systems by testing a representative sample of students. For example, PISA generally assesses what the OECD and its member countries agree that 15-year-olds enrolled in school should know and do in order to operate in a free market economy. To do this, they measure the target population in mathematics, science, and reading. Importantly, PISA does not measure curriculum, nor does it measure constructs such as history, civics, philosophy, or art. Other assessments such as TIMSS have a closer connection to national curricula. Nevertheless, even TIMSS is at best only a snapshot of an educational system taken every four years. As such, inferences can be made but only provide a cross-sectional perspective of a narrowly defined population regarding its performance on a narrowly defined set of topics. Although the majority of ILSA data is collected based on rigorous technical standards and is generally of good quality, it is not perfect and includes error, some of which is reported and some of which is not.

Given the high stakes of ILSA results in many participating countries, it is not surprising that there are both promoters and detractors of the assessments. For example, in the academic literature there exists a strong critical arm arguing that some of the most prominent ILSAs do more harm than good to educational systems (Berliner 2011; Pons 2017; Sjøberg 2015). Questions around the value of ILSAs have

also been posed by major teacher unions (Alberta Teachers' Association 2016) and in the popular press (Guardian 2014) and, with a specific focus on PISA, by the director of the US Institute of Education Sciences (Schneider 2019). Importantly, the USA is one of the largest state funders of the most popular ILSAs (Engel and Rutkowski 2018). In the face of these and other criticisms, promoters of ILSAs contend that the tests have important information to offer and have had a positive influence on educational systems over time (Mullis et al. 2016; Schleicher 2013). Yet, as we have argued in this chapter, demonstrating the specific influence that ILSAs have had on educational systems is not always straightforward given the differing definitions of influence in the literature, along with the inherent complexity of isolating influence in large, complex national educational systems.

Once a definition of influence is established, as we have done in this chapter, it is possible to demonstrate instances when ILSAs clearly have an influence. Our two examples are both problematic for a number of reasons. First, both examples misuse ILSA results in order to encourage and implement policy change. In the case of Norway, poor results on PISA changed how their entire educational system is evaluated. In the case of Australia, policymakers set unrealistic goals and failed to explain how a norm referenced moving goal is justifiable as the ultimate benchmark of educational success, superseding the more common goals a citizenry and its leaders have for its education system. We contend that both cases demonstrate how, without a clear purpose and active management, ILSAs can influence policy in ways that were never intended by the designers.

Although admittedly not foolproof, we argue that one way to properly manage ILSA influence on educational systems is for participating systems to purposefully nominate their reasons for participation and forecast possible intended and unintended consequences of their own participation. Our proposed model depends on an empirical exercise with an assumption that it is possible to establish direct relationships in a complex, multifaceted policy world. We accept this criticism, but note that this is, in many ways, what ILSAs are attempting to do by collecting empirical data on large educational systems. Just like ILSAs, we do not contend that results from our empirical exercise will fully represent the ILSA/policy interaction. However, results from the model should provide more information than is currently available to countries and provide them with: (1) a better understanding of how participation may influence or be influencing their educational systems; and (2) what valid interpretations and uses of ILSA data can and should be made.

We realize that this is a serious endeavor fraught with difficulty but, without a clear purpose and plan for participation in ILSAs, those who understand what claims can and cannot be supported by the data are often sidelined once the mass hysteria of ILSA results enter the public sphere. As such, we developed our model as tool for those who fund participation in ILSAs to be more purposeful concerning the process. We realize the suggested model will require most systems to engage in additional work, but we maintain that systematically evaluating the degree to which ILSA results can serve as the basis for implementing policy changes will help prevent misuse of results. Further, documenting the process provides transparency surrounding what national governments expect from the data and enables testing

organizations to better explain to their clients what valid information the ILSAs can provide. Outside of education, similar forecasting models are well established in the policy literature and common practice in many governmental projects around the world (Dunn 2011, p. 118). Given the high stakes of ILSA results, anticipating or forecasting consequences from the perspective of what ILSAs can and cannot do is a step toward better informing policymakers in their decision making. Our model can also be used to link the influence of ILSAs to any proposed or enacted policy. In other words, the model can work as a tool to understand whether the results from ILSAs are an adequate evidentiary basis to support or inform the policy. We recognize our model will not prevent all misuse or unintended influence of ILSAs, but it invites a more purposeful process.

Educational systems and policies designed to guide and improve them are extremely complex. We are not so naive as to believe that it will ever be possible to document or even understand exactly how ILSAs influence or impact participating educational systems. However, that does not mean that the endeavor is fruitless. We contend that defining terms and participating in an intentional process are two important ways toward understanding how ILSAs influence policy and holding policymakers and testing organizations accountable in how they promote and use the assessments.

References

- Alberta Teachers' Association. (2016). Association to push for PISA withdrawal [webpage]. Edmonton/Calgary, Canada: The Alberta Teachers' Association. <https://www.teachers.ab.ca/News%20Room/ata%20news/Volume%2050%202015-16/Number-18/Pages/PISA-withdrawal.aspx>.
- Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, 18, 1–36.
- Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford, UK: Oxford University Centre for Educational Research. <http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf>.
- Barber, M., Moffit, A., & Kihn, P. (2010). *Deliverology 101: A field guide for educational leaders*. Thousand Oaks, CA: Corwin.
- Beaton, A. E., Mullis, I., Martin, M., Gonzalez, E., Kelly, D., & Smith, T. (1996). *Mathematics achievement in the middle school years. IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College. <https://www.iea.nl/publications/publications/matematics-achievement-middle-school-years>.
- Berliner, D. C. (2011). The context for interpreting PISA results in the USA: Negativism, chauvinism, misunderstanding, and the potential to distort the educational systems of nations. In M. A. Pereyra, H. G. Kotthoff, & R. Cowen (Eds.), *Pisa under examination* (pp. 75–96). Comparative Education Society in Europe Association, Vol 11. Rotterdam, the Netherlands: Sense Publishers.
- Best, M., Knight, P., Lietz, P., Lockwood, C., Nugroho, D., & Tobin, M. (2013). *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries*. Final report. London, UK: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. https://research.acer.edu.au/ar_misc/16.

- Betsill, M. M., & Corell, E. (2008). *NGO diplomacy: The influence of nongovernmental organizations in international environmental negotiations*. Cambridge, MA: MIT Press.
- Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance*. OECD Education Working Papers No. 71. Paris, France: OECD Publishing. <https://doi.org/10.1787/5k9fdqffr28-en>.
- Cox, R. W., & Jacobson, H. K. (1973). *The anatomy of influence: Decision making in international organization*. New Haven, CT: Yale University Press.
- Dunn, W. N. (2011). *Public policy analysis* (5th ed.). Boston, MA: Pearson.
- Engel, L. C., & Rutkowski, D. (2014). Global influences on national definitions of quality education: Examples from Spain and Italy. *Policy Futures in Education*, 12(6), 769–783. <https://doi.org/10.2304/pfie.2014.12.6.769>.
- Engel, L. C., & Rutkowski, D. (2018). Pay to play: What does PISA participation cost in the US? *Discourse: Studies in the Cultural Politics of Education*, 1–13. <https://doi.org/10.1080/01596306.2018.1503591>.
- Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, 32(5), 619–634. <https://doi.org/10.1080/03054980600976320>.
- Grek, S. (2009). Governing by numbers: The PISA “effect” in Europe. *Journal of Education Policy*, 24(1), 23–37.
- Gruber, K. H. (2006). The German “PISA-Shock”: Some aspects of the extraordinary impact of the OECD’s PISA study on the German education system. In H. Ertl (Ed.), *Cross-national attraction in education: Accounts from England and Germany* (pp. 195–208). Oxford, UK: Symposium Books Ltd.
- Guardian. (2014, May 6). OECD and Pisa tests are damaging education worldwide. *The Guardian*. <https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics>.
- Gür, B. S., Celik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, 27(1), 1–21.
- Hopkins, D., Pennock, D., Ritzen, J., Ahtaridou, E., & Zimmer, K. (2008). *External evaluation of the policy impact of PISA*. Report no. EDU/PISA/GB(2008)35/REV1. Paris, France: OECD.
- Kamens, D. H., Meyer, H.-D., & Benavot, A. (2013). *PISA, power, and policy: The emergence of global educational governance*. Oxford, UK: Symposium Books Ltd.
- Keohane, R. O., & Nye, J. (1998). Power and interdependence in the information age. *Foreign Affairs*, 77(5), 81–94. <https://doi.org/10.2307/20049052>.
- Lingard, B., Martino, W., Rezai-Rashti, G., & Sellar, S. (2015). *Globalizing educational accountability*. Abingdon, UK: Routledge.
- Loughland, T., & Thompson, G. (2016). The problem of simplification: Think-tanks, recipes, equity and “Turning around low-performing schools”. *The Australian Educational Researcher*, 43(1), 111–129. <https://doi.org/10.1007/s13384-015-0190-3>.
- Mullis, I. V., Martin, M. O., & Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/20-years-timss>.
- Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America’s schools*. Cambridge, MA: Harvard Education Press.
- Ninomiya, A., & Urabe, M. (2011). Impact of PISA on education policy: The case of Japan. *Pacific-Asian Education*, 23(1), 23–30.
- Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Norway*. Paris, France: OECD Publishing.
- OECD. (2017). *PISA 2015 technical report*. Paris, France: OECD. <http://www.oecd.org/pisa/data/2015-technical-report/>.
- OECD. (2018). FAQ: PISA [webpage]. Paris, France: OECD. <http://www.oecd.org/pisa/pisafaq/>.

- OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. Paris, France: OECD Publishing. <https://doi.org/10.1787/b25efab8-en>.
- Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). Bridging validity and evaluation to match international large-scale assessment claims and country aims. *ETS Research Report Series*, 2018(1), 1–9. <https://doi.org/10.1002/ets2.12214>.
- Pons, X. (2017). Fifteen years of research on PISA effects on education governance: A critical review. *European Journal of Education*, 52(2), 131–144. <https://doi.org/10.1111/ejed.12213>.
- Purves, A. C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review*, 31(1), 10–28. <https://doi.org/10.1086/446653>.
- Rautalin, M., & Alasuutari, P. (2009). The uses of the national PISA results by Finnish officials in central government. *Journal of Education Policy*, 24(5), 539–556. <https://doi.org/10.1080/02680930903131267>.
- Rizvi, F., & Lingard, B. (2009). *Globalizing education policy*. Abingdon, UK: Routledge.
- Schleicher, A. (2013). Lessons from PISA outcomes. *OECD Observer* No. 297 Q4 2013. Paris, France: OECD. http://oecdobserver.org/news/fullstory.php/aid/4239/Lessons_from_PISA_outcomes.html.
- Schneider, M. (2019). Mark Schneider: My response to essay rebuttal—I’m concerned about the PISA exam’s future and the implications of its sponsor’s global ambitions. *The 74 million news site* [webpage]. <https://www.the74million.org/article/mark-schneider-my-response-to-essay-rebuttal-im-concerned-about-the-pisa-exams-future-and-the-implications-of-its-sponsors-global-ambitions/>.
- Schwippert, K., & Lenkeit, J. (Eds.). (2012). *Progress in reading literacy in national and international context: The impact of PIRLS 2006 in 12 countries*. Münster, Germany: Waxmann Verlag GmbH.
- Sellar, S., & Lingard, B. (2013). The OECD and global governance in education. *Journal of Education Policy*, 28(5), 710–725. <https://doi.org/10.1080/02680939.2013.779791>.
- Sjøberg, S. (2015). OECD, PISA, and globalization: The influence of the international assessment regime. In C. H. Tienken & C. A. Mullen (Eds.), *Education Policy Perils* (pp. 114–145). Abingdon, UK: Routledge.
- Takayama, K. (2008). The politics of international league tables: PISA in Japan’s achievement crisis debate. *Comparative Education*, 44(4), 387–407. <https://doi.org/10.1080/03050060802481413>.
- The Parliament of the Commonwealth of Australia. (2012). *Australian Education Act 2012*. Pub. L. No. 223, C2012B00223 (2012). Canberra, Australia: Australian Government. <https://www.legislation.gov.au/Details/C2012B00223>.
- Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge, UK: Cambridge University Press.
- Waldow, F. (2009). What PISA did and did not do: Germany after the “PISA-shock”. *European Educational Research Journal*, 8(3), 476–483.

David Rutkowski is an Associate Professor with a joint appointment in Educational Policy and Educational Inquiry at Indiana University (IU). Prior to IU, David was a Professor of Education at the Center for Educational Measurement (CEMO) at the University of Oslo, Norway. He also worked as a researcher for the International Association for the Evaluation of Educational Achievement (IEA) in Hamburg Germany. David’s research is focused in the area of educational policy, evaluation, and educational measurement with specific emphasis on international large-scale assessment. David has collaborated with or consulted for national and international organizations including the US State Department, USAID, UNESCO, IEA and the OECD. David has worked on and lead evaluations and projects in over 20 countries to include Afghanistan, South Sudan, Trinidad and Tobago and the US. He currently is the editor of the IEA policy brief series, co-editor of the journal *Discourse*, serves on the IEA publication editorial committee (PEC), and is a board member of several academic journals.

Greg Thompson is Associate Professor of Education Research at Queensland University of Technology (QUT). Prior to entering academia, he spent 13 years as a high school teacher in Western Australia. Thompson's research focuses on educational theory, education policy, and the philosophy/sociology of education assessment, accountability and measurement with a particular emphasis on large-scale testing. Recent books include *The Global Education Race: Taking the Measure of PISA and International Testing* (Brush Education), *National Testing in Schools: An Australian Assessment* (Routledge) and *The Education Assemblage* (Routledge).

Leslie Rutkowski is Associate Professor of Inquiry Methodology at Indiana University. She earned her Ph.D. in Educational Psychology, specializing in Statistics and Measurement, from the University of Illinois at Urbana-Champaign. Leslie's research is in the area of international large-scale assessment. Her interests include latent variable modeling and examining methods for comparing heterogeneous populations in international surveys. In addition to a recently funded *Norwegian Research Council* grant on developing international measurement methods, Leslie published the edited volume *Handbook of International Large-Scale Assessment* (Rutkowski, von Davier, and Rutkowski, 2014) with Chapman & Hall. She is also currently writing a text book on large-scale assessment under the Guilford stamp. Leslie currently serves as associate editor for the Springer journal, *Large-scale Assessments in Education*, and as executive editor for the Springer book series, *IEA Research for Education*.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.