Population definitions for comparative surveys in education

Martin Murphy

Australian Council *for* Educational Research

January 2016

The Australian Council for Educational Research Ltd
19 Prospect Hill Road, Camberwell, Victoria, 3124, Australia.

Contents

Table of Figures

Introduction

This paper provides an overview of population definitions for large-scale comparative educational surveys. It has been prepared to help inform the development of a population definition and sampling framework that will be used in the British Council Global English research project. This paper examines a number of large-scale surveys including the Trends in International Mathematics and Science Study (TIMSS), which is conducted by the International Association for the Evaluation of Educational Achievement (IEA), as well as the Programme for International Student Assessment (PISA), and the Assessment of Higher Education Learning Outcomes (AHELO), both of which the Organisation for Economic Co-operation and Development (OECD) conduct. TIMSS and PISA have each been conducted over multiple administrations over many years, and are regarded very highly for their quality. At the present, AHELO has only been administered once as part of a feasibility study, so it is less fully developed compared to TIMSS and PISA. However, it provides interesting insights into the possibilities of survey work in the higher education area.

All of the surveys discussed in this paper are assessments of students. However, comparisons are not made between individual students' results. Rather, data collected from students sampled to participate in the assessment are used to make inferences to a clearly defined population. By doing this, the results can be used to make comparisons between different populations. These comparisons can help identify factors such as teaching practices that may lead to better outcomes for a particular population compared to others. These comparisons can also help inform governments and policymakers about survey participants as well as more broadly about potential areas for improvement.

There are many potential populations that might be inferred to, for example an entire country, a region or a single institution. Comparisons are of most interest when the populations being compared are as similar to each other as possible. Populations such as countries or institutions are structured very differently and it therefore becomes necessary to have a very clear common starting point for comparison, as well as to thoroughly document and quantify any departures from that common starting point. While differences often exist between populations being compared, reports of survey findings will allow the reader to evaluate the similarities or differences between populations across a number of dimensions to better understand the differences observed in student outcomes.

The paper will examine how populations are defined in these large-scale international comparative educational surveys, examples of how some of these have evolved over time, and the implications of these definitions and evolutions on the interpretation of outcomes. It will also examine the implications of decisions about population definitions on the way in which the survey is conducted as well as the impact on data analysis. Finally the paper will provide some examples of how findings from these surveys are reported.

The target population

Particularly for comparative surveys, it is vital that a clear understanding of the target population is reached well in advance of commencing survey fieldwork. Surveys are complex, challenging and expensive activities. Without a clear target population, resources will likely be wasted. Moreover, a lack of clarity in the population definition may lead to misunderstanding and dissatisfaction among survey participants.

Most of the surveys mentioned above seek to monitor trends over time. Any changes to the population definition over time will impact the capacity to measure those trends. A key principle of large-scale survey work is that 'if you want to measure change, don't change your measure'. This principle certainly extends to the population definition. If we alter the population definition in a later cycle of the survey, we are surveying a differently specified population to earlier cycles. This means it will be much harder to know the extent to which any observed change is just the result of this difference, or is a true trend. Investing the time required to develop a clear and appropriate population definition in the first instance in order to minimise changes later on will pay off handsomely over time.

Although changing the population definition can impact the measurement of trends, there are some instances when change is required. The population definitions used in TIMSS and PISA have both changed over time, and some of these changes have been quite significant. Some of these changes have addressed technical issues; some have been the result of discussions and debates following the publication of outcomes. Changes have also occurred due to the success of the surveys. For example PISA has seen participation broaden far beyond its initial focus on OECD member countries. Currently more than 70 countries participate in PISA. The changes required to broaden the population definition in PISA have not been without controversy, and they have invariably added significant additional operational and analytical complexity to the survey. Some of those changes will be discussed in detail in this paper.

Survey population definitions will evolve somewhat over time. In the context of increasing globalisation and change, this is not surprising. However, developing a very clear understanding of what a survey intends to measure at the very start of planning a survey will greatly improve the chances that the population definition will be sufficiently robust to survive that evolution.

A model for defining the population

The graphic shown in Figure 1, taken from the chapter on sampling design in the 2006 PIRLS Technical Report(Joncas, 2007), provides a model for how populations are typically defined in internationally comparative surveys.

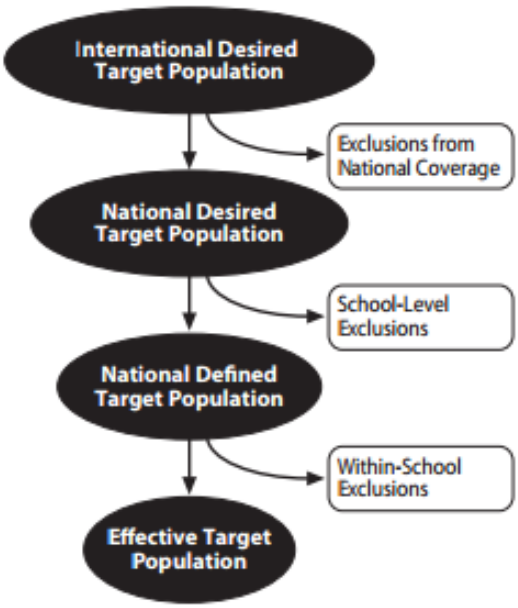

Figure 1: Relationship between desired populations and exclusions

Source. (Joncas, 2007, p. 39).

As shown in Figure 1, the starting point is the International Desired Target Population which is typically quite concise and transferable across participating countries. From this starting point, for various reasons – political, geographical and others – exclusions are identified across a series of levels until the actual population that will be targeted – here described as the Effective Target Population- is arrived at.

It is of course desirable that exclusions be minimised because the distinction between the International Desired Target Population and the Effective Target Population will tend to be overlooked in the reporting of outcomes. As discussed further below, most surveys will set limits on the amount of the population that can be excluded for particular reasons.

It is important that any national departures from the target population such as regional, institution level or student level exclusions are clearly documented and quantified. Technical Reports from large scale surveys will typically include many tables that reflect various aspects of a country's participation, including the way in which they have defined the target population. This information assists the reader to evaluate the quality and comparability of outcomes. Examples of these reports will be discussed later in the paper.

Ultimately the task of accurately defining the population and identifying and quantifying exclusions is the responsibility of the participant. However, at the international level considerable effort needs to be made to ensure that the expectations around population definition and comparability are clearly understood by all participants.

Examples of population definitions

A discussion of the population that a survey seeks to target is included in survey reports. These discussions will typically include:

- an indication of the primary unit of comparison;
- a statement of eligibility with respect to the desired population;
- where appropriate, some form of internationally agreed reference point; and
- some statement about anticipated exclusions and acceptable limits of such exclusions.

The following section discusses each of these aspects of population definitions in detail.

The primary unit of comparison

A key component of the reports of international surveys are comparison tables, where participants are ranked alongside each other across a whole range of survey outcomes. In early administrations of these surveys, participants were generally members of the IEA and OECD respectively, most commonly countries. For example PISA was initially designed as a comparative survey of OECD member states. The population definition for the first cycle of PISA referred simply to 'the country' as the comparative unit. This was presumably intended to mean 'what we typically mean as the country in our work within the OECD'. But even something that is at first glance conceptually simple, like a country can be difficult to clearly define, for example are overseas territories such as Puerto Rico included as part of the United States 'country' sample in this definition?.

As the PISA survey cycles have progressed there has been considerable evolution in the scope of participation. OECD membership has expanded. Within OECD member states, many sub-national entities sought participation as separate entities and as separate entries in comparison tables, for example Scotland, the Flemish community of Belgium, and regions of Spain. Many countries that are not within the OECD have started participating in the survey. In more recent years, sub-national entities such as states (e.g. Tamil Nadu from India) or economies (e.g. Shanghai in China) have been included as survey participants.

In its latest documentation, the comparative unit used in PISA is described more broadly as an 'adjudicated entity':

> *Adjudicated Entity - a country, geographic region, or similarly defined population, for which the International Contractors fully implements quality assurance and quality control mechanisms and endorses, or otherwise, the publication of separate PISA results*(OECD, 2015, p. 23).

Most IEA documentation refers to the primary comparison unit as the country. As with the OECD, membership status to the organisation tends to define the primary comparison units. However, there are some reporting differences between IEA and OECD surveys. For example, in TIMSS 2007 which is conducted by IEA, England and Scotland are listed as participating countries, whereas in PISA, the primary point of

reference is the United Kingdom, with Scotland separately listed as an additional adjudicated region. In IEA studies, non-member states and sub-national entities such as states of the USA or Canadian provinces also participate and are published in IEA reports separately as 'benchmarking participants'.

In the AHELO feasibility study, there was no attempt to infer results to the country level. For this study, outcomes for each participating higher education institution (HEI) were compared with outcomes across all HEIs internationally. Means and distributions of outcomes across all institutions were provided, as was a list of all participating institutions, but it was not possible to directly compare outcomes from one institution to another. Nor was it possible to compare outcomes for an institution with the distribution of outcomes from all participating institutions from that country. These comparisons were beyond the scope of the feasibility study – for example the sampling of HEIs was selected by judgement, and was understood to at best be only broadly representative of the HEIs from the country. For these reasons, it was not seen as appropriate to report the results from the feasibility study at a country level.

For the purposes of this paper, the experiences of AHELO are interesting with respect to two possible designs for surveys of higher education students: surveys of the institutions themselves; or a country or regional level survey of higher education students accessed through their institution of study (in the same manner that PISA and TIMSS access their sample through schools). In the former design, the higher education institution becomes the primary unit of comparison and all of the political and operational considerations associated with ensuring comparability between units at a country or regional level are extended to this level. In the latter design, the institutions serve a similar role to schools in PISA and TIMSS. In this case it becomes important in this case to reflect on how the different nature of institutions can impact on the conduct of the survey and the analysis of outcomes.

Statement of eligibility

The statement of eligibility refers to the description of who is included and excluded from the target population. The PISA population definition is based on the age of participating students:

> *PISA Target Population – students aged between 15 years and 3 (completed) months and 16 years and 2 (completed) months at the beginning of the testing period, attending educational institutions located within the adjudicated entity, and in grade 7 or higher. The age range of the population may vary up to one month, either older or younger, but the age range must remain 12 months in length* (OECD, 2012b, p. 380).

The eligibility for TIMSS is based primarily on number of years of schooling, although with some reference to the average age of students in that year. For example:

> *All students enrolled in the grade that represents eight years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 13.5 years* (Joncas & Foy, 2012, p. 4).

The AHELO statement of eligibility was also based on stage of studies rather than age of students. The target population for the Engineering strand of AHELO was defined as:

> *The target population for the Engineering strand comprises all full-time students at the end of a three- or four-year undergraduate degree in civil engineering of the Engineering department or faculty. It also comprises all full-time students at the end of a three- or four-year undergraduate degree in a multidisciplinary program, who significantly majored in Civil Engineering* (Dumais, Coates, & Richardson, 2011, p. 10).

The choice between an age-based versus a 'stage of studies' based population definition has far-reaching consequences with respect to the administration of the survey, the analyses of survey data and the interpretation of results. These consequences will be discussed later in this paper.

International reference points

International reference points help standardise some of the differences between participants' target populations. Countries define their levels of schooling in different ways, and so the TIMSS eligibility makes reference to the UNESCO International Standard Classification of Education (ISCED). However, these levels are defined within participating countries. Participants must determine – with guidance from the international study centre – which level is the starting point for ISCED level 1 and count up eight years from that point to identify students that meet the eligibility criteria.

Similarly, taking the example of AHELO, civil engineering courses appear in a very wide range of tertiary settings, and their curricula can vary from very academic to very vocational. For these reasons, an international framework was necessary in the AHELO survey to be sure that the same types of civil engineering courses were being included in the Engineering strand across participating countries. Once again, reference was made to the ISCED framework:

> *The programs are referenced by the UNESCO International Standard Classification of Education (ISCED) as level 5A or 5B… Although the ISCED level 5 programs' duration typically ranges from less than 3 years to up to 6 years, the appropriate focus for the feasibility study is on the first range category, i.e. programs with a total cumulative duration of 3 to 4 years* (Dumais et al., 2011, p. 7).

Coverage and Exclusions

While the intention is to maximise coverage of the desired target population for any survey, there will usually be parts of the population that for various reasons cannot be covered. In the TIMSS and PIRLS sample design documentation, the types of potential exclusions are described as follows:

> *… in some rare situations, certain groups of schools and students may have to be excluded from the national target population. For example, it may be that a particular geographical region, educational sub-system, or language group cannot be covered...*
>
> *Even countries with complete population coverage find it necessary to exclude at least some students from the target population because they attend very small schools, have intellectual or functional disabilities, or are non-native language speakers*(Joncas & Foy, 2012, p. 5).

To maximise comparability between participants, limits are set with regard to the amount of the population that can be excluded from the survey. In TIMSS for example, these are stated as:

- The overall number of excluded students must not account for more than 5% of the national target population of students in a country. The overall number includes both school-level and within-school exclusions.
- The number of students excluded because they attend very small schools must not account for more than 2% of the national target population of students (Joncas & Foy, 2012, p. 6).

In the higher education context, exclusions can be due to a variety of reasons. These could be due to unit-level (as distinct from whole-course) enrolments, part-time study, gaps in students' study and that some units may be offered at different levels or stages of students' course. The issues surrounding which students are counted within the target population can become extremely complex:

> *… selected students … absent for the duration of the assessment period but who would likely be present if the collection period were shifted a little (e.g. absent due to short-term illness) are not exclusions but are rather non-respondents. Selected students and faculty who would refuse or would otherwise be unable to take part in the assessment are also to be considered as non-respondents. However, in-scope individuals who are unavailable for the duration of the assessment regardless of the actual collection window (e.g. parental or sabbatical leave, internship) are excluded. Exclusions are to be determined before the sample of students and faculty is selected and the assessment takes place. Exclusions cannot be a by-product of the assessment* (Dumais et al., 2011, p. 10).

A key activity prior to fieldwork is negotiating with participants to reach a common understanding of the institutions and individuals to be counted as 'in-scope' and to identify and quantify exclusions from the target population. This is typically managed through the completion of a series of standardised forms where all proposed exclusions and reductions in coverage are documented, quantified and negotiated between the participant and the coordinating centre.

During data collection, students may be sampled who are subsequently identified as valid exclusions. These exclusions need to be identified and coded accurately,

and distinguished from other exclusions – for example eligible students who are absent or who refuse to participate – so that accurate estimates of the extent of exclusions overall for each participant can be reported.

Implications of decisions surrounding population definition

The units of comparison

As noted above, both TIMSS and PISA have had increased levels of participation in their surveys, such that participation now extends beyond the member states of their respective organisations. In addition to 'countries', there are now 'adjudicated regions' and 'benchmarking participants' who take part in the surveys. For each survey, the distinctions between member countries and non-member participants is maintained in reporting. For example, these are shown by divisions within comparison tables or the use of colours or other graphic elements in publication. However, the reporting does make some comparisons across these units where appropriate and possible.

Participants, be they countries, regions or institutions are understandably very sensitive to being compared against others. This is particularly the case when comparisons may suggest that their students are not performing as well as others. They will have many questions that relate to the way in which comparisons are being made. Has the local context been adequately accounted for? Are the comparisons 'fair' across participants? Are the necessary structures in place with respect to the governance of the survey to ensure that any comparisons are valid? These can be very difficult challenges and controversies can, and do arise with respect to these issues.

The inclusion of Shanghai – as an adjudicated entry – in PISA is one example of a recent controversy. In PISA 2012, Shanghai's outcomes were at the very top of the international comparison tables across Mathematics, Science and Reading. Questions have been raised about how the Shanghai economy is defined for the purposes of PISA, who is included in the population of 'Shanghai 15 year olds', who makes these population decisions, and on what basis those decisions are made, see Loveless (2014). On the one hand, as noted in OECD (2013a), there is considerable interest in knowing more about the education systems of important emerging economies of the world. On the other hand, as the education community is encouraged to look towards Shanghai as a leading 'strong performer and successful reformer in education(OECD, 2013a) there are legitimate concerns about whether the comparisons being made are truly 'like for like'. Loveless (2014) argues that this brings into question the whole governance and decision-making structure of the PISA survey.

It is important to keep in mind that different groups, both internationally and within participating countries, will have different perspectives on survey outcomes. While all of the surveys being discussed in this paper are fundamentally about improving educational policy and provision by identifying the factors that appear most related to improvements in student outcomes, there is no doubt that broader political considerations can be at play. Instances of 'TIMSS shock' and 'PISA shock' caused

by countries being placed in an unexpectedly low position on comparison tables of results from these surveys have been well documented (Döbert, Klieme, & Sroka, 2004). Those coordinating the survey internationally must be able to demonstrate that every opportunity was afforded to the various stakeholders to ensure that local contexts and circumstances were adequately taken into account and that comparisons are 'fair'. One must also expect that some participants, under pressure to be 'favourably located' with respect to particular comparisons may seek to subvert the conduct of the survey, and provision must be made for sufficient monitoring and quality control to avoid the publication of false or misleading outcomes. These issues also apply to any survey where results from participants will be made publicly available, including surveys of institutions rather than countries.

From an international perspective, the focus of reporting is on the outcomes for the country as a whole, but within many countries the main responsibility for the provision of education does not lie at the national level but rather at lower levels of government. In the case of Australia for example, the national estimates from surveys like PISA provide a snapshot of the 'educational health' of the country as a whole, and how this compares to others. However education in Australia is primarily the responsibility of States and Territories and it would most likely be state educational departments who would enact changes to education policy or practice. For example, they would likely be actioning any recommended changes to practices or reallocating resources with respect to the effective teaching of foreign languages in schools. So while the international report limits comparisons to the country level, there will often be supplementary national reporting that provides more detailed analyses by state and territory to assist relevant levels of government and policymakers. Often sample sizes will be boosted for these subpopulations so that reliable estimates can be obtained at these levels.

The evolution of large scale surveys such as TIMSS and PISA over the last two decades shows that the unit of comparison of most value towards meeting the needs of individual participants and also of the survey overall is not necessarily a country. Sub-populations, including adjudicated regions, benchmarking participants, economies, and oversampled regions, are now all very much a part of the picture of these surveys.

There appears to be increasing recognition that useful information from an international perspective can be obtained about educational systems that do not necessarily extend to a whole country. It is also clear that within many countries themselves, participation and comparisons of subgroups of the population can be at least as relevant as comparisons for the country as a whole.

As argued by Wu (2009), there can be a risk that participants in these complex and expensive assessments, by trying to 'solve it all', fail to achieve outcomes at levels of most relevance to policy reform.

> *Focused studies may well be more suited to establish the effectiveness of a particular intervention, or a particular policy change. Smaller, purposeful and targeted assessment programs may achieve a narrow*

> *but well-defined set of objectives rather than a large-scale assessment system that does not providing any useful data*(Wu, 2009).

While it is essential to address any political or operational concerns of participants with respect to comparisons that will be made, designs which allow for reliable reporting at the levels at which policies and reforms are most likely to be enacted are most likely to meet participants' needs.

Age-based versus grade-based eligibility

A key point of difference between the population definition for TIMSS and PIRLS, and the PISA population definition is that the TIMSS and PIRLS populations are both defined according to a particular grade of schooling, whereas the PISA population is defined according to the age of the student. The focus of the IEA studies is on measuring student outcomes and linking these to the curricula taught in schools. PISA instead seeks to investigate the preparedness of students, typically at the end of the compulsory levels of schooling, to meet the challenges of their post-school lives. These different population definitions have important implications for sample design, survey administration, and the analysis of outcomes.

A challenge with the grade-based approach is that of aligning the grade level structures of participants to an international framework. As observed in the TIMSS population definition, this is done via alignment with the ISCED framework, but also with reference to the mean age of students at the particular grade:

> *… because educational systems vary in structure and in policies and practices with regard to age of starting school and promotion and retention, there are differences across countries in how the target grades are labelled and in the average age of students…*(Joncas & Foy, 2012, p. 4)

TIMSS surveys students from two grades – fourth grade and eighth grade. In relation to TIMSS conducted with fourth grade students, this meant that for the countries England, Malta and New Zealand, whose students begin schooling at an earlier age, the most comparative year level was actually the fifth year of schooling. Even then, the students from this target grade were relatively young compared to other countries (Joncas, 2012c)

While there are some complexities associated with a grade-based population definition, they pale in significance when compared to an age-based population definition, such as the one used in PISA. With a grade-based definition, it is typically quite straightforward to determine who is eligible and who is not. The whole notion of an 'eligible educational institution' is also clearer – the institution either does or does not offer that grade. In contrast, defining the population based on an age-range will mean the potential inclusion of many more institutions. There are relatively few 15 year olds in 'junior schools', or in senior secondary schools but there will be some. There will be other institutions offering educational programs – for example work-based vocationally oriented programs – that include some 15 year old students. Should students attending these institutions be considered 'PISA-eligible'?

A comparison between an age-based and a grade-based population brings into focus two contrasting explanatory variables for measuring outcomes – age and years of schooling. With different student entry points into formal education, younger students who have had more time in school might, at least at some points in time, perform better than newer entrants to the school system who are older. Clearly these issues will affect the validity of comparisons. Inevitably whether age- or grade-based, some control with respect to the other variable becomes important. As noted above, while the TIMSS population definition is predominantly grade-based, there is reference made to the average age of students in the target grade and adjustments made when the difference of a population with respect to age is considered too great. In the PISA context, the issue is more focused on the timing of the assessment. It would, for example, affect comparisons if most 15 year olds in one country are at the start of their tenth year of schooling, whereas in another country are at the end of their tenth year of schooling. For this reason, there are some controls on testing windows, with most Northern hemisphere countries conducting assessments between March to May, and Southern hemisphere countries testing later in the calendar year, reflecting the differences in academic year timings across different parts of the world. In addition, it is a PISA standard that the testing period is not held within the first six weeks of the national school academic year. This is to avoid so-called 'holiday effects' experienced by students at that time of year.

These issues would also be relevant in the context of a survey at the senior secondary or higher educational levels. There would clearly be concerns about the comparability of a population definition that meant that in some countries most students in their second year of higher education were being assessed, while in other countries most students were in their third or later years. This is one reason why AHELO pursued a 'stage of studies' model for their population definition, rather than using an age-based definition. On the other hand, in the higher education context, it can be very difficult for institutions to clearly identify individual students within a particular year of study because of the way in which students enrol. Students often enrol in modules rather than whole degrees, they sometimes study on a part-time basis, and they may have gaps in study. In addition, different HEIs have different semester structures, with teaching periods starting at different times, and some offering courses across two semesters, while others have a trimester system. These differences can make it difficult to make comparisons between institutions.

Table 1 shows the distribution of eligible students by grade level and by school type (lower secondary (ISCED 2) versus upper secondary (ISCED 3)) for the PISA 2006 survey.

Table 1: Percentage of students per grade and ISCED level, by country (PISA 2006)

	Grade								
	7	8	9	10	11	12	13	ISCED 2	ISCED 3
AUS		0.1	9.2	70.8	19.8	0.1		80.1	19.9
AUT	0.3	6.4	44.6	48.7	0			6.7	93.3
BEL	0.4	4.4	31.1	63.2	1			6.8	93.2
CAN	0	1.7	13.3	83.8	1.2	0		15	85
CHE	0.8	16.1	62.6	20.3	0.3	0		82.8	17.2
CZE	0.6	3.5	44.3	51.5				50.4	49.6
DEU	1.6	12.3	56.5	29.3	0.3			97.3	2.7
DNK	0.2	12	85.3	1.4	1.1			98.9	1.1
ESP	0.1	7	33	59.8	0			100	0
FIN	0.2	11.7	88.1	0				100	0
FRA	0	5.2	34.8	57.5	2.4	0		40	60
GBR			0	0.9	98.4	0.7		0.5	99.5
GRC	0.5	2.1	5.3	78.8	13.3			8	92
HUN	2.2	5.5	65.7	26.6	0			7.7	92.3
IRL	0	2.7	58.5	21.2	17.5			61.3	38.7
ISL			0.2	99.2	0.6			99.4	0.6
ITA	0.3	1.5	15	80.4	2.8			1.7	98.3
JPN				100					100
KOR			2	97.3	0.7			2	98
LUX	0.2	11.8	53.4	34.4	0.1			64.2	35.8
MEX	2.3	8.1	33.5	48.9	5.1	2		44.5	55.5
NLD	0.1	3.7	44.9	50.7	0.4	0		73.5	26.5
NOR			0.5	99	0.5			99.5	0.5
NZL			0	6.2	89.4	4.4	0	6.2	93.8
POL	0.6	3.8	95	0.6				99.4	0.6
PRT	6.6	13.1	29.5	50.7	0.2			50.2	49.8
SVK	0.7	2.2	38.5	58.7				38.6	61.4
SWE		1.9	95.9	2.2				97.8	2.2
TUR	0.8	4.5	38.4	53.7	2.6			5.3	94.7
USA	0.8	1	10.7	70.9	16.5	0.1		12.4	87.6

Source. (OECD, 2009, p. 144).

For some countries, e.g. the Scandinavian countries, Japan and New Zealand there is a strong relationship between age and grade, with a high proportion of 15 year olds in the same grade. In other countries, the distribution across grade is more mixed. In some cases, for example France, the Czech Republic and Mexico, there are substantial proportions of 15 year olds attending different types of schools. These factors lead to significant additional analytical complexities. For example, where the population substantially divides into different school types, one would expect, and often finds, that the 15 year olds in lower level school types would generally have different performance outcomes than the 15 year olds in upper level school types. The impact of grade retention – a common practice in some countries and vary rare in others – is a major additional complicating factor (OECD, 2013b, p. 73). Analyses that involve variance components such as multi-level modelling are made considerably more complex in these situations.

In relation to the possibility of a target population at the upper secondary level for the Global English study, similar problems will arise, particularly in countries which have a significant separate 'senior secondary school' component within their educational structures.

Following data collection, weights will be calculated for each participating student reflecting sample selection probabilities and adjusting for non-response. The weighting calculation is made much more complex in the PISA environment due to the age-based population definition. The formation of weighting classes in PISA for example takes grade level and sex into account with higher priority than the school itself, so that in some cases eligible students from the upper grades of a school might reside in a different weighting class than students from the lower grades of the same school.

The field operations are also much more complex with an age-based sample design. For a grade-based study, the sample design commonly involves the equal probability selection of an intact class from the list of classes at the target grade from the sampled school. For an age-based design with eligible students across multiple year levels (as shown in Table 1) the sampling operation required involves the preparation of a list of all eligible students at the school from which an equal probability sample is selected, a considerably more complex operation. With students selected across multiple grades and multiple year levels, the disruption to the school is also much greater, as students are likely to come from several different classes, and this may impact upon survey response.

Comparing outcomes from the TIMSS and PISA surveys

One observation that has been quite clearly identified and has been the subject of a number of papers, for example Hutchison and Schagen (2007) and Wu (2010) is that countries' relative outcomes on the TIMSS and PISA surveys and their locations on comparison tables can differ. These differences in relative performance are at least partly explained by the differences arising from how the respective populations are defined. Wu for example concludes that:

> *…. a country with a high score in PISA shows that the students are good at "everyday mathematics", while a high score in TIMSS shows that the students are good at "school mathematics". … The fact that there are differences in country rankings between PISA and TIMSS results suggests that, at least in some countries, school mathematics has not prepared students as well in the application of mathematics as in academic mathematics. Conversely, there are countries that have not prepared students as well in specialist areas of mathematics, such as algebra and geometry, as they have prepared students in solving mathematics problems in everyday life. The question of which approach is better or which curriculum balance is the best will be for the education policy makers in each country to consider in their own context, and, certainly, neither PISA nor TIMSS alone should set the directions for future mathematics curriculum reform*(Wu, 2010, p. 96).

Once again, the message here is that prior to launching a large-scale comparative survey, a very clear understanding of the aims and objectives needs to be achieved. These will feed into the development of a population definition. Another important message is that how the outcomes from a survey are used to inform policy change may vary from country to country, and will depend on a much deeper understanding of local contexts and priorities than is obtained by only reviewing the outcomes from a particular survey.

The institution

While three of the surveys discussed above – PISA, TIMSS and AHELO – are expressly surveys of students, for each survey, the 'educational institution' is a component of the population definition for each.

- The sample designs for both PISA and TIMSS involve accessing the target population from within educational institutions. The samples are designed to be optimised towards the selection of students. Both studies sample schools with the probability of being selected proportional to enrolment size. This optimises the sample with respect to students as students from a stratum are sampled with equal probability, the most efficient design. However, as the term 'probability proportional to size' indicates larger schools are included with higher probability than smaller schools. Particularly in the case of an age-based design with significant portions of the population in different school types, this adds further analytical burdens to the survey. These are discussed further below.
- In the case of the AHELO feasibility study, the institution was used as the primary unit of comparison. There was no attempt to infer outcomes to countries or economies in this survey.

As noted above, one issue is determining the institutions in-scope for the survey. With respect to the upper secondary levels of schooling, there is a diversity of provision for students at this level, with schools and programmes aimed towards higher education, vocational programs, tertiary education and 'post-secondary non-tertiary education' (Wu, 2010). Indeed the distinction between 'education' and 'work' is sometimes not clear cut, for example the case of Austria where students enrolled in vocational educational programs spend periods of the year in school and other times in work environments. This issue led to bias in the outcomes for Austria in the 2000 PISA cycle and comparability problems for subsequent PISA cycles.

Perhaps more importantly, the notion of what constitutes 'a school' itself, i.e. the first-stage sampling unit, can be quite complex. The preferred unit is a 'whole school' but for various reasons there are cases where programs within schools are identified as separate schools for sampling, and alternatively in other cases the school is a larger administrative unit with multiple campuses. Another complicating factor is whether different shifts using the same buildings should be considered as separate schools. In some cases, for example completely separate staffing and management in each shift, this may be more appropriate. When staffing and management is shared across the different shifts, the decision becomes more complex.

Decisions about what constitutes the first-stage sampling unit might vary for different parts of a country, or for different school types. For example, in PISA 2012 the description of the sampling units used for Belgium was: 'A combination of whole schools (French- and German-speaking communities) and implantations (Tracks/programmes taught on a single address/location [administrative address]) (Flemish Community)(OECD, 2014, p. 86). For Slovenia the units were described: 'Study program in ISCED3 schools and whole ISCED2 schools'(OECD, 2014, p. 86).

Once again, these different arrangements lead to complexities in the analysis of data:

> *The structure of education systems also affects the school variance and any multilevel regression analyses. Indeed, the distinction between upper and lower secondary education is part of the within-school variance in some countries where both lower and upper secondary education are provided in one educational institution. On the contrary, in other countries where lower and upper secondary education are provided in separate educational institutions (e.g. in France), this distinction will contribute to the between-school variance* (OECD, 2009, p. 32).

Coverage and exclusions

It is essential for comparability purposes that exclusions are applied on the same basis across participating countries. The process for achieving this outcome begins with an internationally-agreed classification of exclusion categories. For example, the international categories of exclusions identified under PISA were:

- "Intellectually disabled students are students who have a mental or emotional disability and who, in the professional opinion of qualified staff, are cognitively delayed such that they cannot be validly assessed in the PISA testing setting. This category includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students were not to be excluded solely because of poor academic performance or normal discipline problems.
- Functionally disabled students are students who are permanently physically disabled in such a way that they cannot be validly assessed in the PISA testing setting. Functionally disabled students who could provide responses were to be included in the testing.
- Students with insufficient assessment language experience are students who need to meet all of the following criteria: *i)* are not native speakers of the assessment language(s); *ii)* have limited proficiency in the assessment language(s); and *iii)* have received less than one year of instruction in the assessment language(s). Students with insufficient assessment language experience could be excluded.
- Students not assessable for other reasons as agreed upon. A nationally-defined within-school exclusion category was permitted if agreed upon by the PISA Consortium. A specific sub-group of students (for example students

with dyslexia, dysgraphia, or dyscalculia) could be identified for whom exclusion was necessary but for whom the previous three within school exclusion categories did not explicitly apply, so that a more specific within-school exclusion definition was needed" (OECD, 2012b, p. 67).

- Students whose language of instruction for mathematics (the major domain for 2012), was one for which no PISA assessment materials were available. Standard 2.1 of the PISA 2012 Technical Standards "…notes that the PISA test is administered to a student in a language of instruction provided by the sampled school to that sampled student in the major domain of the test. Thus, if no test materials were available in the language in which the sampled student is taught, the student was excluded" (OECD, 2012b, p. 67).

These categories must then be adapted to suit local contexts. It will generally be a person at the sampled school, in conjunction with the centre coordinating the survey within the country, who will be making these exclusion decisions. It can be quite challenging in some contexts to address the need to limit exclusions overall to meet internationally imposed limits on exclusions, while also addressing national and local needs and expectations. For example, when PISA was conducted in the United States, it was not permissible to administer assessments to students who were under individualised educational plans (IEPs) without special accommodations being offered. In addition to these national requirements, schools and teachers may also have different views about the merits of participation for individuals or groups of students.

While the issue of negotiating adaptations to international exclusions categories to suit local contexts is primarily a role for those in charge of field operations, it is important to be able to distinguish between students who were sampled but were subsequently identified as ineligible for the survey from sampled students who were absent or were otherwise non-respondents. Part of the field operation therefore involves the collection and the careful classification of these data about sampled students via tracking forms completed by school level and/or test administration staff.

Inevitably there will be some differences across participants in how exclusions are applied and these of course affect comparability. Table 2 shows an extract from the exclusion rate tables published in the PISA 2012 Technical Report (OECD, 2014) that provides an example of the variations in rates of exclusion across selected countries. This highlights that exclusion rates are an important consideration when comparing, for example, Korea's outcomes with those of Canada or Norway.

Table 2: Variations in rates of exclusion at a school level, within-school level and overall for PISA 2012

Country	School level exclusion rate (%)	Within school exclusion rate (%)	Overall exclusion rate (%)
Australia	1.98	2.06	4
Canada	0.73	5.69	6.38
France	3.63	0.82	4.42
Germany	1.37	0.17	1.54
Japan	2.15	0	2.15
Korea	0.45	0.37	0.82
Norway	1.16	5.01	6.11
United Kingdom	2.66	2.85	5.43
United States	1.01	4.39	5.35

Source. (OECD, 2014).

Survey response

A key point of comparison between participants will be the rates of response of sampled institutions and students to the survey. With non-response comes the possibility of bias in the estimates derived from the responses. In other words, the possibility that respondents and non-respondents differ with respect to survey outcomes. The lower the response rate, the greater the chance of non-response bias. While measures are taken – particularly through the weighting of survey data – to address non-response, these can only attempt to ameliorate the potential effects and are no guarantee against the possibility of non-response bias.

The surveys discussed in this paper set response rate standards prior to survey fieldwork commencement that participants strive to achieve. When response rates are not achieved by a participant, options to address this include providing further evidence to show that the responding sample is unbiased, attaching 'data flags' to outcomes in comparison tables, 'above- and below the line' reporting, or the removal of a participant's outcomes from comparison tables.

In PISA, for example, response rate standards are clearly presented to participants at the start of the survey. There are response rates which are identified as clearly meeting the standard, and other rates that clearly do not meet the standard. Then there is an area in between, where participants have the opportunity to present further evidence that their responding sample is not biased. Those cases are assessed by the sampling contractor, the separately appointed international sampling referee and the Technical Advisory Group as part of the PISA Data Adjudication process. The outcomes of data adjudication decisions are reported in the PISA Technical Report (see for example, OECD, 2014).

In PISA, there have been a number of cases where a participant's data has been deemed of insufficient quality for inclusion in the international comparison tables. For example the Netherlands was removed from international comparison tables in 2000 and the United Kingdom in 2003.

Survey reporting

As illustrated in this paper, while the basic population definition that forms the basis of the survey will be concise and transferable across multiple contexts, the key components that underpin that definition lead to international, national and local variations in participation. An important part of the reporting of these surveys is to quantify these variations in as much detail as possible.

Below are some examples of reporting with respect to the TIMSS 2011 survey of students at Grade 8. With such a long experience in comparative survey work in education, reports from the IEA studies give an excellent insight into how local variations to international population definitions can be reported. (The same also applies with respect to the reports from the PISA survey).

Following the examples from TIMSS reporting are some extracts from an institutional report developed for the AHELO study. The AHELO reporting compares an individual institution with all institutions that participated internationally, as well as making comparisons against various profile markers such as the size of the institution, the source of funding or the highest degree offered.

Reporting Example 1: TIMSS 2011 Grade 8

Table 3: Coverage of TIMSS 2011 target population – grade 8 (extract)

Country	International Target Population		Exclusions from National Target Population		
	Coverage	Notes on Coverage	School-level Exclusions	Within-sample Exclusions	Overall Exclusions
Armenia	100%		1.5%	0.0%	1.5%
Australia	100%		1.3%	1.9%	3.2%
Bahrain	100%		0.5%	1.1%	1.6%
Chile	100%		1.1%	1.7%	2.8%
Chinese Taipei	100%		0.1%	1.2%	1.3%
England	100%		2.2%	0.1%	2.2%
Finland	100%		2.6%	0.9%	3.4%
[1 a] Georgia	93%	Students taught in Georgian	0.9%	3.7%	4.5%
Ghana	100%		0.6%	0.0%	0.6%
Hong Kong SAR	100%		3.9%	1.3%	5.3%
Hungary	100%		2.3%	2.1%	4.4%
Indonesia	100%		3.2%	0.0%	3.2%
Iran, Islamic Rep. of	100%		2.2%	0.0%	2.2%
[3] Israel	100%		16.4%	6.1%	22.6%

1 National Target Population does not include all of the International Target Population.
2 National Defined Population covers 90% to 95% of National Target Population.
3 National Defined population covers less than 90% of National Target population (but at least 77%).
a Exclusion rates for Georgia are slightly underestimated as some conflict zones were not covered and no official statistics were available.

Source. (Joncas, 2012b).

The information included in Table 3 summarises the overall rate of coverage of each countries' target population and the rates of exclusions at a school-level, within-sample level and overall for each country. As shown here, the coverage of students in Georgia has been reduced to students taught in the national language. This reduction in coverage has been quantified to help aid in the interpretation of comparisons. Note that a much higher rate of exclusions occurred in Israel at both

the school-level and within–school level. This clearly will affect comparability between Israel and other participants.

Table 4: Weighted school, class and student participation rates – TIMSS – grade 8 (extract)

Country	School Participation		Class Participation	Student Participation	Overall Participation	
	Before Replacement	After Replacement			Before Replacement	After Replacement
Armenia	100%	100%	100%	97%	97%	97%
Australia	96%	98%	100%	90%	87%	88%
Bahrain	99%	99%	100%	98%	97%	97%
Chile	88%	99%	100%	95%	84%	95%
Chinese Taipei	100%	100%	100%	99%	99%	99%
‡ England	75%	79%	100%	89%	67%	70%
Finland	97%	98%	100%	95%	91%	93%
Georgia	97%	98%	100%	98%	96%	97%

TIMSS guidelines for sampling participation: The minimum acceptable participation rates were 85% of both schools and students, or a combined rate (the product of school and student participation) of 75%. Participants not meeting these guidelines were annotated as follows:
† Met guidelines for sample participation rates only after replacement schools were included.
‡ Nearly satisfied guidelines for sample participation rates after replacement schools were included.
‡ Did not satisfy guidelines for sample participation rates.

Source. (Joncas, 2012a).

As shown in Table 4, England experienced a relatively low rate of school participation. The data flag for England notes that they required replacement schools to 'nearly satisfy' the guidelines for participation rates. Table 5: Information about the students assessed in TIMSS 2011 (extract)

Country	Grade 4		Grade 8		Information About Age of Entry, Promotion, and Retention
	Country's Name for Fourth Year of Formal Schooling*	Average Age at Time of Testing	Country's Name for Eighth Year of Formal Schooling*	Average Age at Time of Testing	
Ghana			Junior High School Form Two	15.8	Children begin school the calendar year of their 6th birthday. Promotion is automatic in Grades 1–6 and dependent on academic progress for Grades 7–9. Promotion is mostly automatic in public schools.
Hong Kong SAR	Primary 4	10.1	Secondary 2	14.2	Children begin school the September after they turn 5 years, 8 months old. Representatives of the Education Bureau may prescribe a maximum rate of repetition.
Hungary	Grade 4	10.7	Grade 8	14.7	Children begin school during the calendar year they turn 6 if their birthday is before May 31st; however, children may begin during the calendar year of their 6th, 7th, or 8th birthday at parental request. Promotion is automatic in Grades 1–3, and dependent on academic progress for Grades 4–8.
Indonesia			Grade 8	14.3	Children must be 7 years old by the end of June to begin on July 12th, although parents have some choice in starting children at age 6. Promotion is dependent on academic progress for Grades 1–8.
Iran, Islamic Rep. of	Grade 4	10.2	Grade 8	14.3	Children must be 6 years old by September 22nd to begin school September 23rd, although there are few private schools that allow registration at 6.5 years. Students with failing grades in June must take a cumulative exam in September to determine promotion or retention.
Ireland	Fourth Class	10.3			The Education (Welfare) Act of 2000 requires children to attend primary schools from the time that they are 6 years old but not before they are 4. In practice, nearly half of 4-year-olds and almost all 5-year-olds are enrolled in infant classes in primary schools. Children only are allowed to repeat a year for educational reasons and in exceptional circumstances.
Israel			Grade 8	14.0	The official policy is that children begin school the calendar year of their 6th birthday, but parents have the final say if they feel their children are not ready to begin. There is retention only in exceptional cases.
Italy	Grade 4	9.7	Grade 8	13.8	Children begin school the calendar year of their 6th birthday, but parents can enroll children who will turn 6 years old by April 30th of the following calendar year in the calendar year of their 5th birthday. The age of entry policy has been revised within the past ten years. Promotion is dependent on academic progress for Grades 1–8.
Japan	Grade 4	10.5	Grade 8	14.5	Compulsory schooling begins at age 6, and children must be 6 years old by April 1st to start school. There is no policy for promotion and retention.

Source. (Mullis, Martin, Foy, & Arora, 2012).

TIMSS provides quite detailed information, as shown in Table 5, about the national name for the grade, the average age at the time of testing, and details regarding entry age and promotion and retention.

Table 6: School sample sizes from TIMSS 2011 (extract)

Country	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample that Participated	Number of Replacement Schools that Participated	Total Number of Schools that Participated
Armenia	153	153	153	0	153
Australia	290	287	276	1	277
Bahrain	97	96	95	0	95
Chile	197	196	166	27	193
Chinese Taipei	150	150	150	0	150
England	150	150	113	5	118
Finland	150	148	143	2	145
Georgia	180	175	171	1	172
Ghana	163	161	161	0	161
Hong Kong SAR	150	150	116	1	117
Hungary	150	147	144	2	146
Indonesia	154	153	153	0	153
Iran, Islamic Rep. of	250	238	237	1	238
Israel	152	151	143	8	151
Italy	204	204	166	31	197
Japan	150	150	128	10	138
Jordan	232	230	230	0	230

Source. (Mullis et al., 2012).

Table 6 includes a summary of the school sample, participation and number of replacement schools. This shows that Chile, Italy and Japan each used a relatively high number of replacement schools in their participation relative to other countries.

Table 7: Student sample sizes – TIMSS 2011 (extract)

Country	Within-school Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Eligible Students	Number of Students Absent	Number of Students Assessed
Armenia	97%	6,057	0	0	6,057	211	5,846
Australia	90%	9,007	192	141	8,674	1,118	7,556
Bahrain	98%	4,960	185	27	4,748	108	4,640
Chile	95%	6,290	95	82	6,113	278	5,835
Chinese Taipei	99%	5,166	34	22	5,110	68	5,042
England	89%	4,382	88	3	4,291	449	3,842
Finland	95%	4,549	16	26	4,507	241	4,266
Georgia	98%	4,779	66	51	4,662	99	4,563
Ghana	97%	8,073	486	0	7,587	264	7,323
Hong Kong SAR	96%	4,261	42	55	4,164	149	4,015
Hungary	96%	5,489	28	55	5,406	228	5,178
Indonesia	96%	6,201	190	0	6,011	216	5,795
Iran, Islamic Rep. of	99%	6,264	141	0	6,123	94	6,029
Israel	92%	5,174	19	64	5,091	392	4,699
Italy	96%	4,379	23	210	4,146	167	3,979
Japan	94%	4,747	14	46	4,687	273	4,414
Jordan	96%	8,439	344	28	8,067	373	7,694
Kazakhstan	98%	4,551	70	25	4,456	66	4,390
Korea, Rep. of	99%	5,315	43	42	5,230	64	5,166
Lebanon	96%	4,231	103	0	4,128	154	3,974
Lithuania	93%	5,285	50	100	5,135	388	4,747
Macedonia, Rep. of	95%	4,360	67	23	4,270	208	4,062
Malaysia	98%	6,209	334	0	5,875	142	5,733
Morocco	94%	9,869	333	0	9,536	550	8,986
New Zealand	90%	6,079	128	41	5,910	574	5,336

Source. (Mullis et al., 2012).

Table 7 shows the student sample sizes by country, and includes information on the participation rates of students, the number of students sampled, number of students who withdrew from the class or school, the number of exclusions, the number of absent students and the total number of students who participated in the assessment. This information shows that the numbers of students withdrawn, excluded or absent vary quite considerably between participating countries.

Croatia

Fourth Grade

Coverage and Exclusions

- Coverage is 100 percent.
- School-level exclusions consisted of very small schools (MOS < 6), hospital schools, schools for minority groups (language and writing, and models A and B), schools in which the majority of the classes are composed of solely Roma children, and private elementary schools.
- Within-school exclusions consisted of students with special needs and special program teaching.

Sample Design

- Explicit stratification by school type.
- Implicit stratification by region (Središnja, Istocna, Sjeverna, Zapadna, Južna, and Zagreb) or area (21).
- Sampled two classrooms in large schools in the "One Building School" stratum (MOS > 90) and sampled two classrooms in each sampled school in the "Multiple Building School" and "Minority School" strata.
- Satellite schools of mother schools were treated as classrooms of the mother school for purposes of sampling.
- School sample overlap between TIMSS (Grade 4) and PIRLS: 1) Samples were drawn all at once; 2) All sampled schools for TIMSS were asked to participate in PIRLS; and 3) All sampled students for TIMSS also were asked to take PIRLS.

Exhibit 16: Allocation of School Sample in Croatia, Fourth Grade

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Refusal Schools	Excluded Schools
			Original Schools	1st Replacements	2nd Replacements		
One Building School	59	0	57	2	0	0	0
Multiple Building School	91	0	91	0	0	0	0
Minority School	2	0	2	0	0	0	0
Total	**152**	**0**	**150**	**2**	**0**	**0**	**0**

Figure 2: Croatia's TIMSS 2011 sampling summary

Source. (Joncas, 2012d)

TIMSS provides a report for each country with quite detailed information about coverage and exclusions, stratification, institution type and participation over the sampled strata. An example of the type of information given to countries is shown in Figure 2.

Reporting Example 2: AHELO Engineering Strand Institutional Report

This section includes extracts from the AHELO Engineering strand institutional reports. These extracts provide an example of the type of information that institutions receive about sampling, participation and results. The extracts are taken from one of the participating institution's report.

Table 8: AHELO Engineering Strand participation statistics

		All institutions		This institution	
		#	%	#	%
Institutions	Participation	92	100	1	100
Students	Population	10875	100	174	100
	Sample	8223	76	174	100
	Response	6078	74	174	100
Faculty	Population	2312	100	22	100
	Sample	2015	87	22	100
	Response	1480	73	22	100

Source. (OECD, 2012a).

As shown in Table 8, internationally, 92 institutions participated in the Engineering strand of the AHELO feasibility study. The names of the participating institutions and their countries are provided at the end of the institution report[1]. This table shows that the populations of Engineering students and staff from these 92 institutions were 10,875 and 2,312 respectively. This institution had 174 students and 22 faculties in the target population, with all participating in the survey.

Table 9: AHELO Engineering Strand institution characteristics and scores

		#	%	X	SD
THIS INSTITUTION'S RESULTS		**1**	**100**	**506**	**97**
Institution size	20000 or more full time students	29	33	527	92
	5000 to 19999 full time students	41	46	499	100
	Fewer than 5000 full time students	19	21	459	96
Institution source of funding	≥ 75% Public funding	26	31	488	93
	50-75% Public funding	20	24	510	95
	25-50% Public funding	17	20	523	99
	< 25% Public funding	22	26	473	100
Highest qualification offered	Bachelor	5	6	397	84
	Masters	10	11	436	85
	Doctorate	74	83	514	96
Bachelor curriculum	Specialised	13	15	491	118
	General/broad	2	2	426	78
	Blend	70	82	501	96
Institution emphasis	Research	7	8	570	95
	Teaching	18	20	465	90
	Balance research/teaching	63	72	503	98
Institution location	Major city	54	61	501	103
	Regional city	32	36	494	98
	Small town/rural area	2	2	548	76
Student pathway	≥ 25% non-traditional pathway	10	16	514	98
	≥ 75% traditional pathway	53	84	500	101

Source. (OECD, 2012a).

The institution's performance on the Engineering assessment (the mean (X) and standard deviation (SD) in the top row) is shown in Table 9. This also shows these

[1] In the case of Japan, the institution names are provided as 'Institution 1', 'Institution 2', etc.

results compared against aggregated results for all participating institutions by different institutional profiles. The institution can identify its profile among the different profiles and use this to compare its performance with others of a similar profile, as a way of evaluating its performance relative to others similar institutions. There are several such profile comparisons provided in the full report.

Table 10: AHELO Engineering Strand demographic characteristics and scores

		All institutions				This institution			
		#	%	X	SD	#	%	X	SD
Total		6078	100	500	100	174	100	506	97
Gender	Male	4440	74	512	99	122	71	523	93
	Female	1552	26	471	93	51	29	468	96
Age	≤ 20 years	406	7	488	86	10	6	-	-
	21 years	1799	30	520	99	52	30	527	86
	22 years	1861	31	517	97	67	39	520	98
	23 years	932	16	494	93	16	9	487	101
	≥ 24 years	956	16	453	94	27	16	465	92
Home language	Language of instruction	5600	93	502	99	166	96	506	97
	Other language	392	7	498	95	7	4	-	-
Enrolment	Part time	3349	56	489	92	104	60	504	93
	Full time	2603	44	517	106	68	40	510	104

Source. (OECD, 2012a).

Table 10 provides an indication of the demographic profile of students who participated in the AHELO feasibility study at this particular institution as well as all other institutions that participated in the survey.

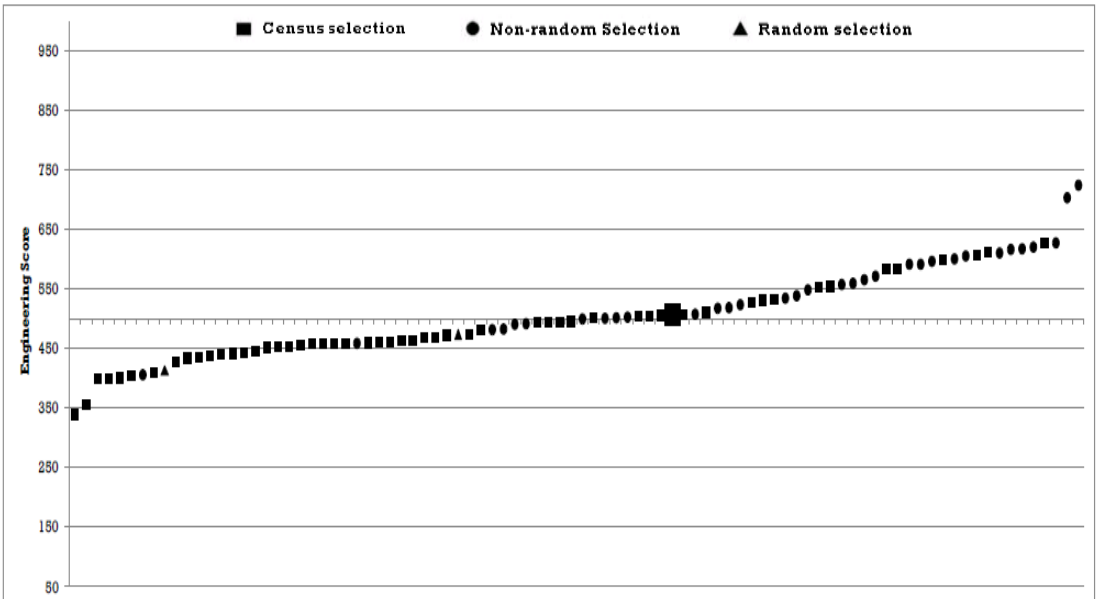

Figure 3: AHELO Engineering Strand mean scores for all participating institutions and this institution

Source. (OECD, 2012a).

Figure 3 displays the mean score of the institution (shown as a larger point on the graph). This is shown compared with the means for all other participating institutions. This graph also indicates whether the student sample for the institution was random or non-random.

Table 11: AHELO Engineering Strand education characteristics and scores

		All institutions				This institution			
		#	%	X	SD	#	%	X	SD
Total		6078	100	500	100	174	100	506	97
Teacher availability	Low	810	14	499	102	31	18	520	92
	Medium	1761	29	499	100	50	29	507	99
	High	3428	57	503	97	92	53	503	99
Frequency of teacher comments	Low	1339	22	507	99	48	28	525	101
	Medium	1890	32	504	100	55	32	503	96
	High	2770	46	497	98	70	40	498	95
Satisfaction with entire experience	Poor	245	4	474	88	7	4	-	-
	Fair	1556	26	491	96	50	29	488	85
	Good	3349	56	501	99	92	53	512	105
	Excellent	846	14	531	100	24	14	533	88
Professional skills and knowledge development	Not at all	49	1	469	86	3	2	-	-
	Very little	388	6	480	91	10	6	-	-
	Some	1612	27	500	101	51	29	491	104
	Quite a bit	2777	46	504	99	84	49	516	96
	Very much	1168	19	508	99	25	14	527	87
Graduate expectations	Related job	4161	69	500	95	116	67	503	91
	Unrelated job	283	5	466	97	11	6	-	-
	Further study	1312	22	514	109	39	23	520	121
	Other	242	4	502	95	7	4	-	-

Source. (OECD, 2012a).

Table 11 provides a summary of students' overall score for this particular institution reported by student responses relating to their course experience, the extent to which they report they have developed professional skills and knowledge, and their future plans. These comparisons also help institutions to contextualise the results from the survey, and to understand differences among their student cohort.

Conclusion

Population definitions have evolved in the major international comparative surveys in education – TIMSS, PISA, PIRLS, AHELO and others – over time. As these surveys have matured, and as participation has broadened beyond the member states of the IEA and the OECD, the surveys have better been able to accommodate sub-national as well as national comparisons. At least in some cases, those comparisons involving sub-national entities – for example states or provinces of a country – are likely to provide insights at least as useful to policy makers within those countries because these levels are where much of the educational policy development and practice is driven. Along with the benefits and experiences that come from participating in these high quality surveys at the international level – such as building capacity in the conduct of large scale surveys for national survey work; building networks with like-minded colleagues; obtaining useful insights into outcomes nationally and how they relate to educational structures and other background

factors – countries can consider more specific national needs and consider participation of sub-national entities.

How a survey population is defined can lead to profound consequences in the operations of a survey and with respect to data analysis and reporting. The difference between the grade-based, curriculum-based IEA studies and the age-based PISA survey for example has led to different interpretations of outcomes: one more focussed on schools and curriculum, the other with a broader 'literacy' perspective. Both have resonated strongly with the educational community internationally as evidenced by the continued strong participation in these surveys.

A key component of the success of these surveys has been the extensive work in documenting national variations to the international population definition framework. These allow the reader to evaluate comparability across a wide range of factors. Given the sheer scale of the activity of these international surveys, outcomes are necessarily somewhat 'broad brush'. But with the very detailed reporting of national variations within the international framework, researchers and policy makers from a particular country or sub-national entity have a much better chance of identifying similarities and differences with other countries and contexts and be motivated to investigate these contexts more deeply as part of their efforts to improve the provision of education within their own system.

References

Döbert, H., Klieme, E., & Sroka, W. (Eds.). (2004). *Conditions of school performance in seven countries: A quest for understanding the international variation of PISA results*. Münster, Germany: Waxmann Verlag.

Dumais, J., Coates, H., & Richardson, S. (2011). *AHELO sampling manual*. *Assessment of Higher Education Learning Outcomes (AHELO)*. Paris, France. Retrieved from http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=EDU/IMHE/AHELO/GNE(2011)21/ANN3/FINAL&doclanguage=en

Hutchison, D., & Schagen, I. (2007). Comparisons between PISA and TIMSS: Are we the man with two watches? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 227–262). Washington, DC: Brookings Institution Press. http://doi.org/10.7864/j.ctt12800b.13

Joncas, M. (2007). PIRLS 2006 sample design. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/PDF/P06_TR_Chapter4.pdf

Joncas, M. (2012a). Meeting TIMSS 2011 standards for sampling participation. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/pdf/T11_Standards_Sampling.pdf

Joncas, M. (2012b). TIMSS 2011 population coverage and exclusions. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/pdf/T11_Pop_Coverage.pdf

Joncas, M. (2012c). TIMSS 2011 target population sizes. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/methods/pdf/T11_Pop_Sizes.pdf

Joncas, M. (2012d). TIMSS 2011: Characteristics of national samples. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/pdf/T11_Characteristics.pdf

Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

Loveless, T. (2014). *The 2014 Brown Center report on American education: How well are American students learning? 2014 Brown Center Report on American Education*. Washington, DC. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/03/18-brown-center-report/2014-brown-center-report_final.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center,

Boston College. Retrieved from http://timssandpirls.bc.edu/timss2011/downloads/T11_IR_M_FrontMatter.pdf

OECD. (2009). *PISA data analysis manual: SPSS* (2nd ed.). Paris: PISA, OECD Publishing. http://doi.org/10.1787/9789264056275-en

OECD. (2012a). *AHELO feasibility study institution report: AHELO Engineering University: Civil engineering learning outcomes. Unpublished report.*

OECD. (2012b). *PISA 2009 technical report*. PISA, OECD Publishing. http://doi.org/10.1787/9789264167872-en

OECD. (2013a). *Lessons from PISA 2012 for the United States, strong performers and successful reformers in education*. OECD Publishing. http://doi.org/10.1787/9789264207585-en

OECD. (2013b). *PISA 2012 results: What makes schools successful? Resources, policies and practices* (Volume IV). PISA, OECD Publishing. http://doi.org/10.1787/9789264201156-en

OECD. (2014). *PISA 2012 technical report*. PISA, OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2015). *PISA 2015 technical standards*. Paris, France. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2015-Technical-Standards.pdf

Wu, M. (2009). Issues in large-scale assessments. In *Pacific Rim Objective Measurement Symposium 2009*. Hong Kong. Retrieved from http://www.edmeasurement.com.au/_publications/margaret/Issues_in_large_scale_assessments.pdf

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS* (OECD Education Working Papers No. 32). Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/5km4psnm13nx-en