

An introduction to International Large-Scale Assessments (ILSAs) for government officials, postgraduate students and applied researchers.

John Jerrim (UCL Social Research Institute)¹

Abstract

International comparisons are of great interest to politicians, policymakers, government officials, journalists and researchers. International large-scale assessments (ILSAs) are one high-profile example. Yet these are complex endeavours, relying on quite sophisticated methods for sampling, test design and consequently analysis. This book is designed to help readers to navigate many of these technicalities. Its central purpose is to communicate the intuition behind many of the key ideas that underpin ILSAs without becoming unduly technical. It is intended to serve as a one-stop resource that covers many of the basics in as accessible a manner as possible, while still giving readers the conceptual tools needed to understand how these studies work, how their results should be interpreted and the caveats that should be applied.

Key words: PISA, TIMSS, PIRLS, PIAAC, ILSAs, International comparisons.

Funding: No funding was received for this work.

AI use: The author has used AI (Chat-GPT) in three primary ways in developing this book. In the writing, the author produced a first draft (or detailed bullet points) for each paragraph in turn. AI was then used to help polish the English and sentence structure. Second, AI was used to construct some images/figures, as documented in the figure notes throughout. Third, AI has been used to verify each chapter section-by-section to help improve accuracy and spot any errors.

¹ J.Jerrim@ucl.ac.uk. 55-59 Gordon Square, Bloomsbury, London. WC1H 0AL.

1. Introduction

1.1. What is this book about?

International comparisons are of great interest to politicians, policymakers, government officials, journalists and researchers. In recent decades, there has been a proliferation of international statistics and datasets designed to compare countries across a wide range of outcomes. When well conducted and used carefully, such comparisons can provide valuable insights into important questions. For instance, how are we doing compared with other countries? Are we improving or declining relative to our international peers over time? These questions arise in many domains, from economic growth (e.g. quarterly GDP figures) to international football rankings (e.g. the FIFA rankings) and – as per the focus of this book - the performance of education systems. For politicians and policymakers, comparative indicators offer a way to monitor national performance and (at times) make claims about the success or failure of the government of the day. For journalists, they often provide interesting stories and compelling narratives. While, for researchers, they provide an opportunity to examine whether social and economic phenomena vary across countries, whether findings generalise across contexts and what individual and country-level factors may help explain these differences.

International large-scale assessments (ILSAs) are one high-profile example of this broader growth in international comparison. Their history dates back to the 1960s, when the first 12-country pilot study was conducted (IEA, n.d.). However, their breadth, scope and prominence have increased substantially since the mid-1990s. The first publication of Programme for International Student Assessment (PISA) 2000 results in 2001 was particularly important in raising the public and political profile of ILSAs, which has continued to this day. These studies are now widely cited by politicians, attract extensive media coverage and receive intense scrutiny whenever new results are released. Some countries have even set explicit targets based on ILSA rankings or performance. For example, Uzbekistan has aimed to be among the top 30 countries in PISA by 2030 (President of the Republic of Uzbekistan, 2019), while Australia previously set a goal of reaching the top five PISA countries by 2025 (Gillard, 2012) and the United Arab Emirates aimed to be in the top 20 by 2021 (UAE Government, 2024). ILSAs have also encouraged “policy tourism” with governments visiting high-performing systems to see what lessons might be imported (e.g. see Chung, 2010, for a discussion with respect to Finland). At the same time, ILSA data are now widely used by researchers, with reviews suggesting that studies drawing on PISA data alone are approaching 4,000 publications (Xu et al., 2026).

ILSAs are, however, complex endeavours. This complexity is inherent in collecting large amounts of data across many countries, languages, education systems and institutional contexts. As a result, ILSAs rely on quite sophisticated methods for sampling, test design and consequently analysis. Researchers also often use these data in innovative and imaginative ways to address original research questions, sometimes well beyond the purposes for which the studies were originally designed. A great deal of documentation is therefore available, including lengthy technical reports (e.g. von Davier et al., 2024), user guides (OECD, 2009) and handbooks on their use (Nilsen et al., 2022). These resources are often written by those closely involved in conducting the studies, explaining the methods in technical detail, but in places give less prominence to limitations than to procedures. The difficulty is that many of these resources are also long, technical and targeted at quantitatively sophisticated users. In practice, they are not always read or followed closely. For example, Troccoli (2020; 2024) found that many published studies in high-profile economics journals failed to account for key technical features of the ILSA data.

This book is therefore primarily aimed at two core user groups. Its central purpose is to communicate the intuition behind many of the key ideas that underpin ILSAs without becoming unduly technical. It is intended to serve as a one-stop resource that covers many of the basics in as accessible a manner as possible, while still giving readers the conceptual tools needed to understand how these studies work, how their results should be interpreted and the caveats that should be applied.

The first audience is those working in or alongside national governments. ILSAs typically have a high level of government engagement, with civil servants often being directly responsible for overseeing their delivery or in close collaboration with national survey agencies. Moreover, a broader group of government officials often have a strong interest in the results, though they lack specialist quantitative training or a background in statistics. It is therefore vital that they are well briefed on the main technical features so they can critically evaluate the methodology and findings.

The second audience is students and early-career researchers, many of whom first encounter ILSA data during a master's course or at the start of a PhD. To pick up and first use ILSA data, these researchers do not necessarily need to understand every technical detail and nuance. They should, however, build a basic understanding of the intuition behind how such studies are designed and what this then implies for their own analyses.

The book is also likely to be of interest to data journalists, third-sector researchers, education policy advisers and advanced undergraduate students – many of whom also consume ILSA results and should be aware of their strengths and limitations.

The need for such a resource is particularly important now with the growing use of large-language models (LLM) and artificial intelligence (AI). AI is increasingly moving into the analysis of quantitative data and can draw on software packages that handle many of the technical features of ILSAs (see section 1.7 for an overview of such software). Yet this makes it even more important that government officials and applied researchers understand what is happening “under the hood”. In other words, what increasingly matters is not simply being able to run analyses using data from ILSAs but understanding what the software (and any AI use of it) is doing, why it is doing it, how the resulting estimates should be interpreted and what key caveats that should be applied. This is precisely what this book attempts to do.

1.2. What is an international comparison?

An international comparison is a systematic comparison of outcomes, characteristics or relationships across two or more countries or jurisdictions. In the context of ILSAs, it commonly involves comparing population-level distributions and summary statistics for knowledge, skills and related outcomes (though comparisons may also cover educational experiences, attitudes and features of education systems). Prominent examples include comparing average levels of achievement in key subjects (e.g. reading, science, maths) across countries, examining whether socioeconomic inequalities in achievement are larger in some nations than in others and investigating whether certain features of the education system (or another country-level factor) are linked to higher or lower levels of achievement.

An international comparison requires more than simply placing national statistics side by side. It is vital that the measures being compared are sufficiently comparable, carrying the same meaning across countries. This may in turn require the use of a common measurement framework, comparable assessment and background questions, careful translation and adaptation procedures, appropriate sampling designs, high response rates, scaling methods and harmonised data collection procedures. Comparability is thus the *raison d'être* of ILSAs and other similar data collections; without it, their added value over other datasets is limited.

International comparisons can also take several forms. The simplest are descriptive, such as comparing national mean scores, the percentage of individuals falling above or below a specific achievement threshold and measuring (and comparing) trends over time. Yet more complex comparisons are also possible. For instance, cross-nationally comparable data allow analysts to understand whether a given factor is a stronger independent predictor of achievement or socio-emotional development in some countries than in others. Likewise, they can help establish if a particular social science phenomenon that one observes in a single country (or a small number of countries) generalises across a diverse set of nations (e.g. see Marsh & Hau, 2003, for cross-national evidence on the presence of so-called “Big-Fish Little-Pond” effects). International comparisons may also allow researchers and policymakers to understand whether particular institutional or macroeconomic features of countries are linked to better educational, social and/or labour market outcomes.

Thus, in summary, international comparisons are best understood as a structured attempt to measure similarities and differences in selected inputs and outcomes across national contexts using measures designed to be as comparable as possible. Their value therefore does not simply boil down to identifying which country is doing the “best”. Rather, they can help researchers and policymakers understand commonality and variation across countries, challenge assumptions based on evidence from a single national setting and – in theory - generate insights about how outcomes are shaped by national institutions, policies, cultures and wider social conditions.

1.3. Different forms of international comparative data

Broadly speaking, one can distinguish between two types of international data. I refer to these as “ex-ante” and “ex-post” harmonised datasets.

1.3.1. Ex-post harmonised data

Ex-post harmonised data were originally collected within individual countries. They have typically been designed to produce country-specific research and insights, rather than to draw comparisons to other countries. Nevertheless, researchers or research groups have then attempted to harmonise these country-specific datasets after they have been collected. This involves a process whereby the researchers make variables as comparable as possible across countries (after the data collection is complete). For instance, each national dataset may contain information on parental education recorded using the specific qualification framework of that country. The researchers then attempt to recode these parental education measures within each national dataset into a common international classification (e.g. typically either International

Standard Classification of Education (ISCED) groups or number of years spent in education). Similarly, surveys may ask about family income using different questions, reference periods, currencies and response categories. Researchers may nevertheless convert these responses into a common metric (e.g. purchasing-power-adjusted dollar values) or focus on relative differences in each country (e.g. children in the top and bottom 25% of the family income distribution *within* each individual country). Such endeavours are sometimes undertaken by individual researchers or small teams for specific one-off projects. Examples include:

- Investigations into the role of academic achievement versus individual choice in determining educational attainment across eight countries (Jackson et al. 2013).
- Investigations of inequalities in educational achievement and how these develop during childhood across four English-speaking countries (Washbrook et al., 2012; Bradbury et al., 2015).
- Explorations of how attending a private school is linked to post-secondary educational transitions and labour market outcomes across three countries (Jerrim et al., 2016).

Other ex-post harmonisation efforts have been undertaken by larger research groups. These have typically received greater time and monetary investments and are intended to support a wide range of projects by different research teams. Examples (many outside the field of education) include:

- The Luxembourg Income Study (LIS) and Luxembourg Wealth Study (LWS). These bring together household-level survey data on income, wealth, employment and demographic characteristics from individual country data sources. The LIS includes over 50 countries, with several having data available at multiple time points. These data have been widely used to study topics such as poverty, inequality and the distribution of household assets across national contexts. Further information is available from <https://www.lisdatacenter.org/about-lis/>
- The Cross-National Equivalent File. This brings together household panel surveys from several countries, including Australia, Canada, Germany, the UK, Japan, South Korea, Russia, Switzerland and the United States. It is designed to support cross-national research on topics such as income, employment, family structure, health, education and social mobility, while retaining the longitudinal structure of the original national surveys. The data are made as comparable as possible by creating equivalently defined variables across the participating studies (despite the original household surveys not being designed to facilitate cross-national comparisons). Further details are available from <https://www.cnefdata.org/>

- Education at a Glance. An annual summary of education systems across countries published by the OECD. While some topics vary between years, issues that have been covered include educational attainment, participation, finance, teachers and institutional structures. The data are drawn from a range of both national and international data sources, with further details available from OECD (2025). Comparability is supported through common classifications and definitions of variables, though there are important cross-national differences in data sources, reference periods, coverage, reliability thresholds and classification practices. Further details are available from https://www.oecd.org/en/publications/2025/09/education-at-a-glance-2025_c58fc9ae.html
- Multinational Time Use Study (MTUS). This resource brings together diary-based time-use surveys from multiple countries. It allows researchers to compare how people allocate time to paid work, unpaid work, education, leisure, sleep and other daily activities. It converts national time-use surveys into a common internationally standardised format, including harmonised background variables and activity categories. Nevertheless, the individual national surveys vary in terms of sampling, response rates, survey mode, diary design, reference days and activity coding (amongst other issues). Further details are available from <https://www.timeuse.org/mtus>

A major advantage of some of these ex-post harmonised datasets is that they provide longitudinal information on participants. In other words, some follow individuals over time, making it possible to examine developmental trajectories and longer-term outcomes in more than one country. This is a significant advantage over and above most ex-ante harmonised datasets (including most ILSAs) which are usually cross-sectional at the individual level. They can therefore provide an international perspective on questions that would otherwise be difficult or impossible to study comparatively. However, ex-post harmonised data also have important limitations. Most obviously, the original datasets were not designed or collected with international comparison in mind. There thus remains questions as to whether the harmonised variables – and other key aspects of the data (e.g. national representativeness) – are sufficiently comparable across countries to facilitate robust international comparisons. Moreover, many ex-post datasets only include a small number of countries – particularly those based on longitudinal national datasets. This then limits the type of research questions that the international comparative aspect of these data affords (see section 1.5 for further discussion of this issue).

My overarching view is hence that comparisons of precise estimates across countries should be treated with caution when using ex-post harmonised international comparative data. There is

often considerable uncertainty surrounding its true comparability, with this being an issue that is hard to quantify and clearly convey. Such data are however valuable in terms of establishing whether broad patterns and relationships hold in more than a single national context (e.g. does the socio-economic gap in achievement persist or grow from pre-school and into compulsory schooling across English-speaking countries – as per Bradbury et al., 2015).

1.3.2. Ex-ante harmonised data

Ex-ante harmonised data - including ILSAs - are different. These studies have been designed from the outset with cross-national comparisons in mind. Indeed, facilitating such comparisons is usually their central purpose. This means that issues of international comparability are built into almost every stage of their development, including the definition of target populations, sampling procedures, translation, survey administration, assessment design, data collection windows and technical quality standards. It means that common frameworks are developed before data collection begins, rather than attempts to shoehorn variables into international schemas after the data have already been gathered. Countries are also expected to meet minimum standards (e.g. in terms of response rates and sample exclusions) with any deviations clearly justified and documented in international technical reports.

As facilitating cross-national comparisons is the main purpose of ex-ante harmonised data collections, the number of countries included is often larger than in many ex-post harmonised datasets. This is especially true in comparison to ex-post harmonised longitudinal data or those harmonised by a small team of researchers. For example, recent cycles of PISA have included more than 80 countries, while more than 50 education systems have participated in TIMSS, PIRLS and TALIS. This is considerably more than in many of the ex-post harmonised datasets discussed above (some of which included less than 10 countries).

Together, these features mean that one should generally have greater confidence in the cross-national comparability of ex-ante harmonised data (at least when such studies are well executed). The larger number of participating countries also makes such data especially valuable for research questions where the relevant variation occurs at the country level (discussed further in section 1.5). On the other hand, as most ex-ante international data collections are cross-sectional, they do not usually follow the same individuals over time (though there are some recent limited exceptions – see section 1.4.3 below). They also have limited space in their background questionnaires, meaning they cannot collect the same depth of information on certain topics as some specialised ex-post harmonised datasets (e.g. detailed information on time use - where the ex-post harmonised Multinational Time Use Study is likely to be a more appropriate resource).

The remainder of this book is focused on ex-ante harmonised data in the specific form of ILSAs. Nevertheless, many of the topics and issues that shall be discussed also apply to other international data collections as well. This includes both the ex-post datasets overviewed above, along with other ex-ante international data collections that are available for research (e.g. the Health Behaviour in School-aged Children - <https://www.hbsc.org/>).

1.4. An overview of major ILSAs.

While some clearly receive more attention than others, a wide variety of ILSAs now exist. These cover different populations and can be used to investigate a multitude of different issues. Short summaries of some of the most prominent are provided below.

1.4.1. The Programme for International Student Assessment (PISA)

PISA is an international assessment of 15-year-olds led by the OECD. It was first conducted in 2000 and has been repeated approximately every three years since. However, following the 2025 results, PISA will move to a four-year cycle. PISA assesses pupils aged 15 years 3 months to 16 years 2 months who are enrolled in school and have completed at least six years of formal education (OECD, 2023a). Its core domains are reading, mathematics and science, with one “major domain” (focal subject) each cycle. However, it also includes other innovative domains (e.g. creative thinking and financial literacy). In most countries, PISA is now a computer-based assessment taking around two hours to complete. Participants and headteachers also complete background questionnaires. Moreover, there are additional optional questionnaires collected in some countries covering issues such as wellbeing or attempting to collect additional data from parents. The most recent cycles of PISA collect data from over 700,000 pupils from more than 80 countries, including all members of the OECD. In addition to the achievement tests, PISA has been used to study a diverse range of issues including educational inequalities, learning environments, teaching approaches and various socio-emotional outcomes of young people. It is perhaps the most widely known ILSA and has been used extensively in secondary research. For instance, a systematic review published in 2018 identified 654 peer-reviewed studies using these data (Hopfenbeck et al., 2018) though more recent studies with broader inclusion criteria put the figure closer to 4,000 documents in total (Xu et al., 2026). Further details about PISA are available from <https://www.oecd.org/en/about/programmes/pisa.html> with the data downloadable from <https://www.oecd.org/en/data/datasets.html?orderBy=mostRelevant&page=0&facetTags=oecd-languages%3Aen&q=PISA>

1.4.2. Trends in International Mathematics and Science Study

TIMSS is led by the International Association for the Evaluation of Educational Achievement (IEA). It assesses mathematics and science achievement amongst fourth-grade (≈age 9/10) and eighth-grade (≈age 13/14) pupils. First conducted in 1995, TIMSS has been conducted every four years since. In the most recent cycles, TIMSS has collected data from over 600,000 pupils from more than 60 countries. As TIMSS samples entire classrooms within schools, it is possible to link teachers to their pupils. In addition to the mathematics and science test, TIMSS also collects a range of background questionnaires completed by pupils, their teachers, headteachers and parents (fourth grade only). These cover a range of topics including curriculum implementation, family background, classroom resources, teaching practices, school climate and attitudes towards mathematics/science. Further details about TIMSS are available from <https://www.iea.nl/studies/iea/timss> with data available from <https://www.iea.nl/data-tools/repository/timss>

1.4.3. The TIMSS-Longitudinal study (TIMSS-L)

TIMSS-L is a longitudinal extension of the TIMSS study. It was first implemented after TIMSS 2023, making it the first pupil-level longitudinal extension of TIMSS. Specifically, pupils that participated in the TIMSS-L 2023 study were reassessed one year later to measure mathematics and science achievement growth. It also resurveyed participating pupils, providing longitudinal information on their attitudes to learning and socio-emotional skills. A total of nine countries participated in the fourth-to-fifth grade longitudinal extension, with three of these participating in the eighth-to-ninth grade extension as well. These data can thus be used to explore changes in academic achievement, school climate, home resources, wellbeing, absenteeism, bullying and attitudes to mathematics and science over the course of an academic year. Further details are available from <https://www.iea.nl/studies/iea/timss/2023-Longitudinal> with data available from <https://timss2023.org/longitudinal/data/>

1.4.4. Progress in International Reading Literacy Study (PIRLS)

PIRLS is also led by the IEA and is focused on the assessment of reading skills of grade 4 (9/10-year-old) pupils. It was first conducted in 2001 and has been conducted every five years since. The most recent cycles have included data from over 50 countries including around 400,000 pupils. Similar to TIMSS, it is possible to link teachers to pupils, with background questionnaires completed by pupils, their teachers, headteachers and parents. These questionnaires cover a range of issues including the home literacy environment, classroom instruction, school context and attitudes to reading. A 2023 systematic review identified 221 academic publications using or discussing PIRLS data (Stiff et al., 2023). Further details about PIRLS are available from

<https://www.iea.nl/studies/iea/pirls> with data available from <https://www.iea.nl/data-tools/repository/pirls>

1.4.5. The Teaching and Learning International Survey (TALIS)

TALIS is an OECD-led international survey of teachers and school leaders. It was first conducted in 2008, with subsequent waves in 2013, 2018 and 2024 (with the fifth cycle due in 2030). Its core target population is lower-secondary school teachers and headteachers, although recent cycles also offer additional primary and upper-secondary options (OECD, 2025a). TALIS does not test teachers' knowledge or skills. Rather, teachers and school leaders complete self-report questionnaires, typically taking around 45–60 minutes. The 2024 cycle collected data from about 280,000 educators from across 55 education systems. Topics covered include information on teachers' backgrounds, professional learning, classroom practices, school leadership, collaboration, wellbeing, job satisfaction and technology use. It has also been widely used in secondary education research; a systematic review conducted in 2024 synthesised a total of 238 peer-reviewed journal articles using TALIS data (Veletić et al., 2024). Further details about TALIS are available from <https://www.oecd.org/en/about/programmes/talis.html> with data available from <https://www.oecd.org/en/about/programmes/talis.html#data>

1.4.6. TALIS video study

The TALIS Video Study (also known as Global Teaching InSights) was an OECD study designed to provide direct observational evidence on classroom teaching (OECD, 2020). Implemented during 2017/18, it was a one-off international study with no subsequent data collections announced. It focused exclusively on lower-secondary mathematics teachers teaching a unit on quadratic equations to 13-15-year-olds across eight participating education systems. Unlike the main TALIS survey, the video study collected data from multiple participants in multiple ways, including the ability to link teachers to the pupils they teach. Specifically, the data collected included video recordings of lessons, teacher questionnaires, pupil questionnaires, lesson artefacts and pre/post-tests of pupil's mathematical knowledge and skills. The study sought to identify features of effective teaching by examining instructional quality, classroom interactions, cognitive demand, teacher practices, classroom climate and the relationships between teaching and pupils' academic and socio-emotional outcomes. As a lesser-known ILSA, secondary research using these data is relatively limited (though see, for example, papers by Jerrim et al. 2025, Herbert et al., 2022 and Charalambous et al., 2025). Further details are available from <https://www.oecd.org/en/about/projects/global-teaching-insights.html> with data available from <https://www.oecd.org/en/about/projects/global-teaching-insights.html#technical>

1.4.7. The Programme for the International Assessment of Adult Competencies (PIAAC)

PIAAC (also known as the Survey of Adult Skills) is a household survey and assessment led by the OECD. It was first implemented in 2011/12, with further first-cycle rounds (adding data from additional countries) in 2014/15 and 2017/18. The second PIAAC cycle was conducted in 31 countries (including 27 members of the OECD) in 2022/23. The focus is adults aged 16 to 65 (though some collect additional data for those up to age 74) assessing their numeracy, literacy and adaptive problem-solving skills. In total, around 160,000 adults participated. The assessment is administered through a personal interview and assessment, with background questions on education, work, skills use, learning, social-emotional skills and non-economic outcomes. A recent review found 880 publications drawing upon PIAAC data, with around half of these being academic journal articles (Maehler et al., 2025). Further details are available from <https://www.oecd.org/en/about/programmes/piaac.html> with data available from <https://www.oecd.org/en/about/programmes/piaac/piaac-data.html>

1.4.8. The International Civic and Citizenship Education Study (ICCS)

The ICCS is led by the IEA and is an assessment of civic and citizenship education amongst eighth-grade (≈age 13/14) pupils. It was first implemented in 2009, and has since been conducted in 2016, 2022 and 2027. The focus of ICCS is civic knowledge, civic understanding, attitudes, engagement and intended participation, with background questionnaire data collected from pupils, teachers and schools. The 2022 round of ICCS covered 24 education systems, with around 80,000 participating pupils. Key topics include democracy, citizenship, institutions, civic participation, sustainability, diversity, school climate and opportunities for civic learning. A systematic review (Myoung & Liou, 2022) identified 135 peer-reviewed articles using ICCS and its predecessor (the Civic Education Study). Further details are available from <https://www.iea.nl/studies/iea/iccs> with data available from <https://www.iea.nl/data-tools/repository/iccs>

1.4.9. The International Computer and Information Literacy Study (ICILS)

ICILS is an IEA study that measures eighth-grade pupils' (≈13-14-year-olds) computer and information literacy skills. It began in 2013 and has been conducted every five years since. The computer-based assessment is complemented by background questionnaires that capture pupils' access to technology, digital learning experiences and attitudes. Data are also gathered from schools on related issues such as resources available, teacher preparedness and leadership. At the time of writing, around 35 education systems participated in ICILS, with data

gathered from over 100,000 pupils. Although the secondary literature remains considerably smaller than for studies such as PISA or TIMSS, a recent scoping review synthesised findings from 114 peer-reviewed journal articles using ILSA data focused on information and communication technology (Mejía-Rodríguez et al., 2026). Further details are available from <https://www.iea.nl/studies/iea/icils> with data available from <https://www.iea.nl/data-tools/repository/icils>

1.4.10. Regional assessments

Alongside global international large-scale assessments, a number of regional large-scale assessments have been developed to address the educational priorities and policy contexts of specific world regions. These typically measure pupils' foundational learning outcomes while collecting comparable information on schools, teachers and learning environments, thereby facilitating comparisons against regional rather than global peers. For instance, the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) is a regional assessment programme for Southern and Eastern Africa, which has assessed Grade 6 pupils in periodic cycles since 1995. It focuses mainly on reading/literacy and mathematics/numeracy, alongside contextual information on pupils, teachers and schools. Similarly, the Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN (PASEC) has monitored primary education quality in mainly francophone African education systems since the 1990s. It assesses pupils near the beginning and end of primary schooling (e.g. Grades 2 and 6) in language/reading and mathematics, alongside teacher and school questionnaires. Likewise, the Pacific Islands Literacy and Numeracy Assessment (PILNA) assesses Year 4 and Year 6 pupils' reading, writing and numeracy skills across 15 Pacific Island nations. Compared with studies such as PISA, PIRLS and TIMSS, microdata from these regional assessments are generally less openly available as direct public downloads. However, they may be accessed via online request for research purposes.

1.5. **Why do cross-national comparisons?**

ILSAs have been used to address a wide range of research questions and policy-relevant issues. However, they are better suited to some purposes than others. This subsection introduces a series of broad types of questions that researchers and policymakers have commonly examined using ILSA data.

1.5.1 Benchmarking

One of the major uses of ILSAs is to allow countries to benchmark their performance against that of other education systems. This often involves comparing average achievement in subjects such as reading, mathematics and science. Yet benchmarking can also extend to many other outcomes. For example, teacher workload is an important policy concern in numerous countries. TALIS allows education systems to compare teachers' (self-reported) total working hours, as well as how their time is divided between different tasks (e.g. teaching, marking, lesson preparation, administration). ILSAs also allow countries to benchmark themselves against their own previous performance. Whenever new results are released, there is great interest in whether performance has improved, declined or remained stable since the previous cycle. Repeated participation therefore enables countries to track trends over time across a range of outcomes.

Studying trends through an international comparative perspective provides at least two advantages over examining changes within a single country.

First, the measures used in ILSAs are likely to be less susceptible to domestic political and policy pressures than national measures. Democratically elected governments often have a strong incentive to demonstrate that educational standards have improved during their time in office. This may create conditions in which grade inflation is allowed to occur in national examinations, for example through more lenient marking or changes to grade boundaries over time. ILSAs are in contrast developed and administered through international organisations, making the measures less directly influenced by the political priorities of a single government. They may therefore provide a more independent basis for assessing changes in educational performance over time.

Second, ILSAs allow countries to assess relative (as well as absolute) changes in performance over time. Suppose that children's mathematics knowledge and skills in Country A have improved over the past 10 to 15 years. Viewed in isolation, this would appear to provide positive evidence that the education system is improving well. However, ILSAs make it possible to compare this trend with developments in several other countries. It may be that achievement has risen elsewhere at an even faster rate. In that case, Country A would have improved in absolute terms but could also be said to be in *relative* decline compared with its international peers.

International benchmarks can also make otherwise abstract statistics easier to interpret. For instance, a common measure of inequality is the Gini coefficient. This ranges from 0 (complete equality) to 1 (complete inequality). Yet a Gini coefficient such as 0.34 is difficult to interpret in

isolation – is inequality high or low? The figure becomes more meaningful when it is compared with the Gini coefficients of other countries or with earlier values from the same country. Such benchmarking provides the necessary context to the result that aids interpretation.

1.5.2. Generalisability and exceptionality of educational and social phenomena

Many studies in education and the social sciences are conducted within a single national context. Indeed, systematic reviews and meta-analyses are often dominated by evidence from the United States and other English-speaking countries. This creates an important limitation; relationships identified in one country may not necessarily hold elsewhere (Oo et al., 2026). Another important use of ILSAs is therefore to examine the generalisability of research findings across national settings. By applying similar measures and analytical approaches across countries, researchers can investigate whether a particular association appears consistently, varies in strength or is present only under certain conditions.

A useful example stems from the gender gap in reading achievement, which has been documented in national assessments and academic studies within individual countries; girls tend (on average) to outperform boys in this subject (Fernández Darraz & Mora Guerrero, 2017; Reilly et al., 2019; Hofmeyr, 2022; Manu et al., 2023). An important question, however, is whether this pattern holds true across every country across the world (or, at least, every country where data from ILSAs are available)? Data from PIRLS 2021, for example, allow us to see that the estimated average reading score for girls was higher than that for boys in every participating education system (Mullis et al., 2023)². A very similar pattern emerged in PISA 2022. Girls achieved a higher average reading score than boys in every participating country, again with only a few differences falling just short of conventional statistical significance (OECD, 2025, p.125). Data from these ILSAs thus help us to demonstrate how the gender gap in reading skills is close to universal.

ILSAs can also help identify countries that stand out from broader international patterns (i.e. instances where countries are somewhat exceptional). For example, research in the United States has long suggested that teenagers often hold highly ambitious educational expectations - many more young people expect to graduate from university than eventually do. Viewed only through national evidence, it is difficult to know whether this is a widespread feature of adolescence or something distinctive to the United States.

² Though a handful of these differences did not quite reach statistical significance at the 5% level.

Jerrim (2014) investigated this question using data from PISA and TIMSS and found evidence of a degree of American exceptionalism. Although the United States is not the only country in which more teenagers expect to obtain a university degree than eventually do so, this pattern is not universal. In some countries, the opposite relationship appears, with fewer young people expecting to graduate from university than the proportion who ultimately do. Jerrim (2014) also showed that the gap between teenagers' expectations and actual university completion rates is especially large in the United States compared with many other countries. This was particularly evident among lower-achieving American pupils, a substantial proportion of whom still anticipated completing a university degree.

Thus, as well as helping researchers assess generalisability, ILSAs can also reveal when a country appears to be “exceptional” (i.e. when a feature or phenomenon within the country stands out from other countries across the globe).

1.5.3. Impact of institutional / national features

A further use of ILSAs is to examine the effects of institutional structures (or national features) such as how specific aspects of education systems are organised. Certain characteristics (e.g. accountability arrangements; academic selection in school admission) often vary little within a country but differ substantially across countries. International data therefore provide an opportunity to investigate whether these system-level differences are associated with pupils' outcomes.

A concrete example that has attracted substantial attention is between-school tracking (e.g. Jakubowski, 2010; Ruhose & Schwerdt, 2016; Strello et al., 2021). In some education systems, children are separated at a relatively early age into different types of schools with more academic versus more vocational pathways. This is common in countries such as Germany, where selection typically occurs at around age 10, and the Netherlands, where it occurs around age 12. As these arrangements usually operate across entire national education systems there is often little variation within countries³. The main variation is therefore between countries. Researchers have thus attempted to exploit this cross-national variation to investigate whether early tracking

³ Some within-country variation does exist in the use of between-school tracking in some countries, such as England. See Jerrim and Sims (2019) for empirical evidence on this matter from this particular country.

is associated with outcomes such as average levels of achievement and educational inequality. See chapter 7 (section 7.6) for a selection of studies using ILSA data to study this topic.

Using ILSAs to investigate the effects of institutional structures is, however, not without its challenges. For instance, there are often only a reasonably small number of countries that can be included due to data availability. Yet, since the key source of variation occurs at the country level, this can lead to limited statistical power (as well as making estimates sensitive to the inclusion or exclusion of individual countries). Moreover, studies using such approaches sometimes attempt to combine data across several ILSAs (e.g. measuring change from the 4th grade PIRLS and TIMSS samples through to PISA results). Yet these studies differ in their target populations, assessment frameworks, survey instruments and implementation, limiting their comparability (see section 7.6 for further details). A more fundamental difficulty concerns causal inference. Countries that adopt different institutional arrangements are also likely to differ in many other respects, making it difficult to isolate the effect of any single feature. This issue is discussed further in chapter 7.

Thus, although ILSAs have been used to study the “impact” of certain education-system level factors such as between-school tracking, the resulting evidence should be interpreted cautiously. In general, ILSAs are better suited to answering broad, descriptive questions about patterns and differences across education systems. Estimating the causal effects of country-level institutional/structural features is considerably more difficult to do well.

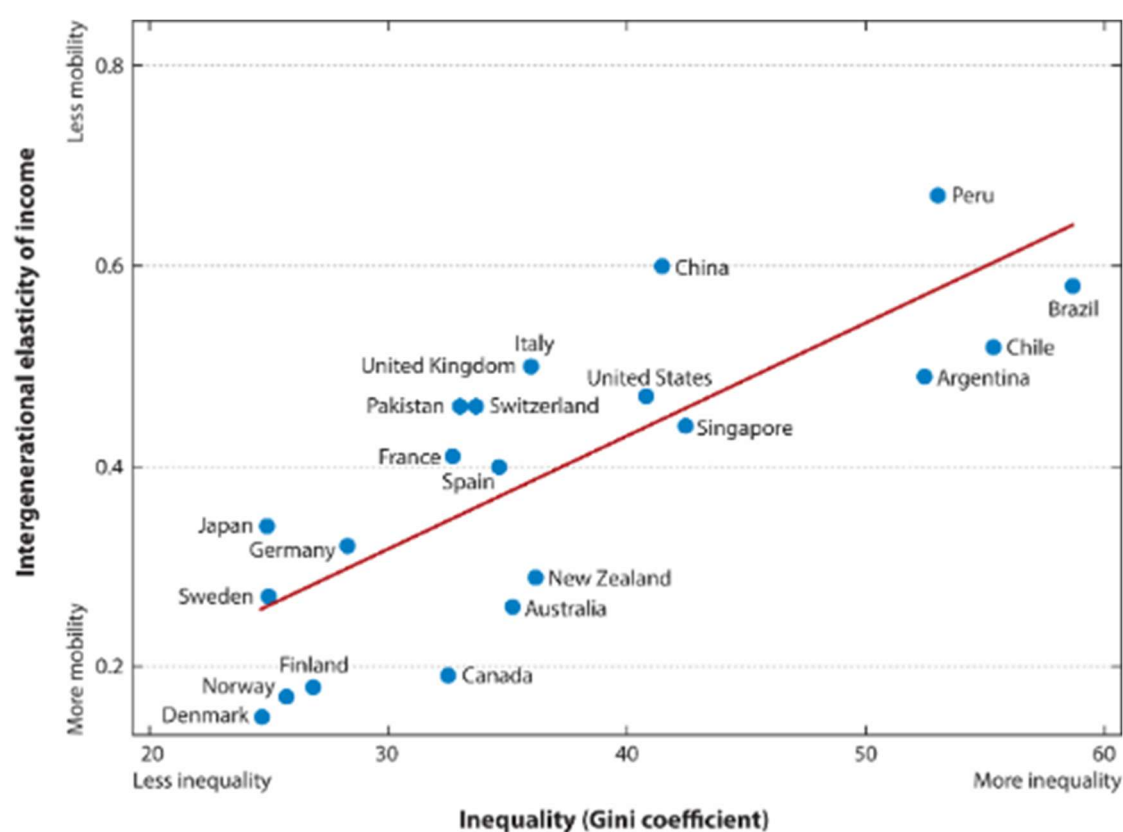
1.5.4. “Impact” of macro-level forces

Building on the previous subsection, a related use of ILSAs is to investigate how broader “macro-level” forces are associated with individual outcomes. A prominent example is income inequality, which is usually measured and reported at the country level. Researchers can combine such national indicators with individual-level ILSA data to examine whether – for instance – young people living in more unequal societies tend to have different educational outcomes from those living in more equal ones. A well-known example of this broader approach is *The Spirit Level* (Wilkinson & Pickett, 2009) which presents a series of cross-country correlations between income inequality and outcomes relating to the environment, social cohesion, education and health (Pickett, 2024). Although factors such as income inequality can sometimes be measured at lower geographic levels, estimating their “effects” becomes more complicated because of

issues such as internal migration and the non-random selection of where individuals choose to live.

An example of such a cross-country correlation is illustrated in Figure 1.5.1, in what has become widely known as the Great Gatsby Curve (Durlauf et al., 2022). The horizontal axis plots the level of income inequality within each country, while the vertical axis captures intergenerational persistence (i.e. the extent to which an individual's earnings are related to their family background). This now famous graph suggests that countries with higher levels of income inequality also tend to have lower levels of social mobility. In other words, in countries where the income distribution is more unequal, the socio-economic position into which individuals are born appears to be more strongly related to how much they earn as adults.

Figure 1.5.1. The Great Gatsby Curve



 Durlauf SN, et al. 2022
Annu. Rev. Econ. 14:571–605

Source: Durlauf et al. (2022).

This type of analysis shares many of the limitations discussed in the previous subsection on institutional structures. The number of countries available for comparison is often relatively small, with the analyses having limited statistical power and maybe sensitive to the inclusion or

omission of individual countries. In addition, the relationships identified are usually correlational rather than causal. Indeed, we return to the use of cross-country scatterplots later in this book, including cautions regarding their interpretation in section 7.5 (as well as guidance on their presentation in section 8.5). Despite these limitations, such analyses can produce interesting descriptive plots, highlight important country-level associations and identify countries which may be outliers internationally. This is often where the real strength of ILSAs lies; identifying important associations and generating questions for further investigation. Using the data to try and credibly address causal questions is a much more challenging task.

1.5.5. Provide novel data

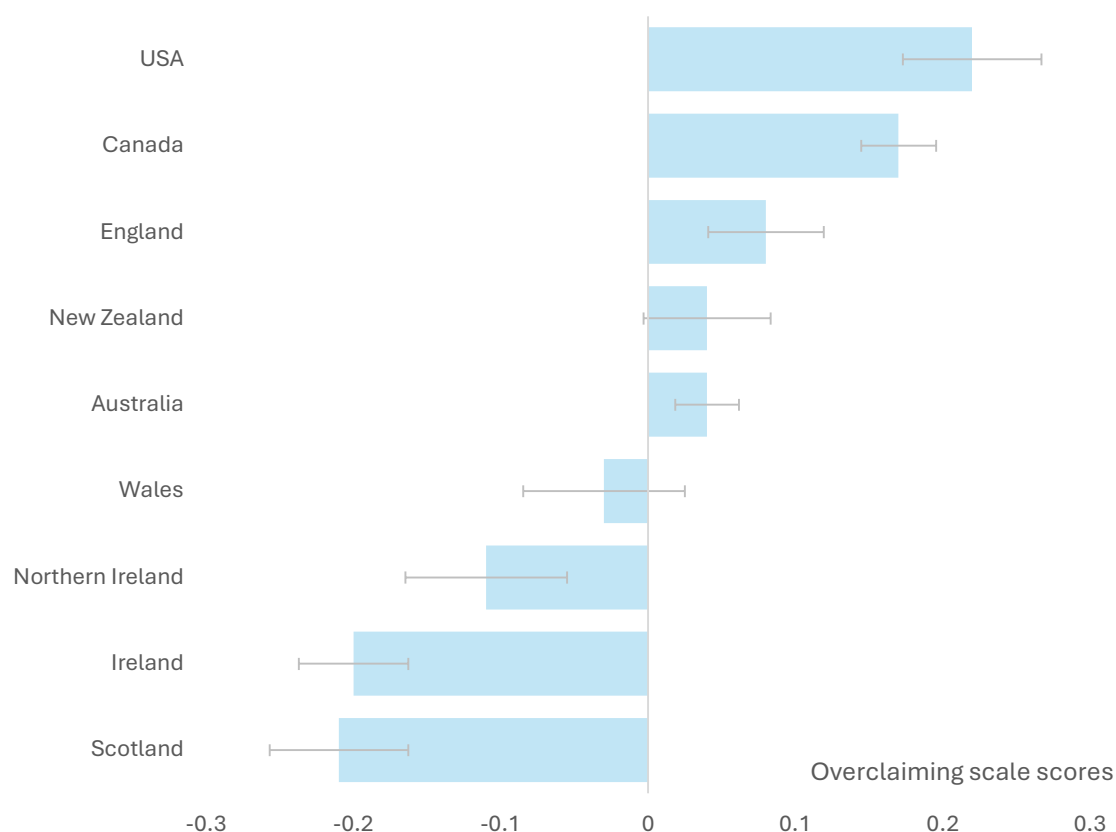
ILSAs are best known for measuring achievement in mathematics, reading and science. However, they also provide a wide range of other valuable (and sometimes novel) data. This includes assessments of young people's knowledge and skills in areas that are not routinely or formally tested in many education systems. Examples include financial literacy and creative thinking in PISA, which offer additional insight into broader aspects of young people's knowledge and skills.

ILSAs also collect rich information through their contextual background questionnaires, much of which is not available in other large-scale datasets. Longitudinal studies of children and young people conducted in several countries (e.g. the Millennium Cohort Study in the United Kingdom, the Singapore Longitudinal Early Development Study and the Young Lives survey conducted in Ethiopia, India, Peru and Vietnam) often cover a broad range of topics including education, health, family relationships and wider social circumstances. This means they do not have the space to include detailed subject-specific information that ILSAs provide (e.g. attitudes towards maths and science). The value of ILSAs therefore extends beyond international comparisons. They can also support novel research within individual countries, while allowing findings to be placed within a wider cross-national context.

One particularly fun example is provided by Jerrim et al. (2023), who used PISA data to investigate how much young people “overclaim” across countries (i.e. the extent that teenagers exaggerate their knowledge and skills). Specifically, in some PISA background questionnaires, pupils were asked to rate how familiar they are with a range of mathematical concepts, such as probability and linear equations. The catch is that not all the concepts listed were real. Some were entirely fictional (“declarative fraction”). Jerrim et al. (2023) used responses to these fake items to construct a measure of overclaiming and compare its prevalence internationally. Their results,

summarised in Figure 1.5.2, provide an entertaining illustration of the kinds of unusual questions that can be explored using ILSAs. As it turns out, teenagers in North America are biggest “overclaimers” across these English-speaking countries. In other words, they are particularly likely to say they understand mathematical concepts that do not even exist!

Figure 1.5.2. Average country scores on the PISA overclaiming scale



Source: Jerrim et al. (2023).

TALIS also deserves particular mention. Unlike most ILSAs, it focuses specifically on teachers and school leaders rather than on pupils. Many countries find it difficult to collect high-quality survey data from teachers, with response rates in national studies often falling well below 30% (Jerrim, 2023). TALIS is a notable exception, typically achieving substantially higher participation rates through its carefully organised survey procedures. In some countries, it therefore probably provides the most reliable nationally representative survey evidence available on teachers. TALIS illustrates how data from ILSAs are also amongst the highest-quality survey data available on certain topics and groups within individual nations.

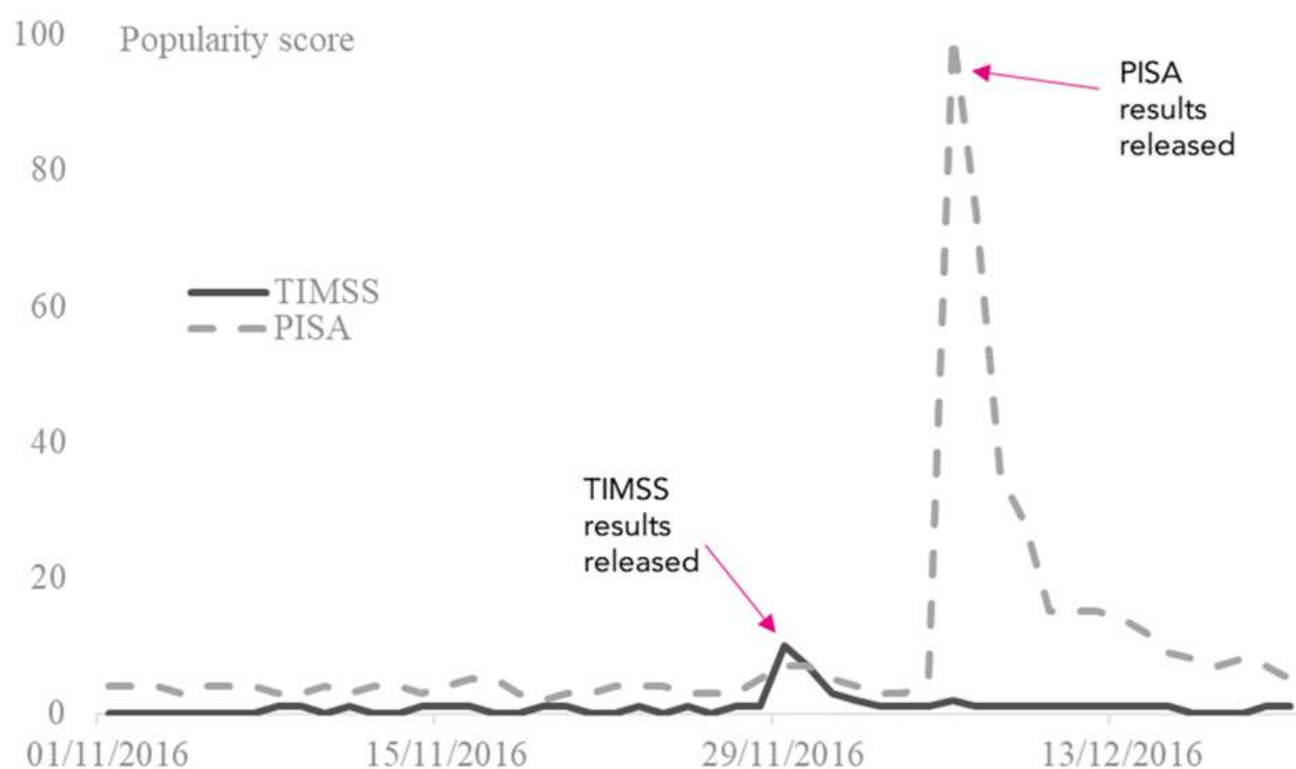
1.5.6. Policymakers and politicians’ care

ILSAs are now big business. Whether readers of this book love them or hate them, their results are closely watched by the media, policymakers and the wider public (Jerrim, 2024). In particular,

international rankings and country comparisons generate substantial attention, shaping national debates about educational quality, standards, inequality and reform. They also influence political narratives and contribute to changes in education policy (Rutkowski et al., 2020). Given this prominence, it is important that the methods underpinning ILSAs are understood by as wide an audience as possible.

As an aside, PISA often dominates policy and political narratives surrounding ILSAs. Its results are particularly widely watched, cited and discussed. Jerrim (2023) illustrates this by showing that PISA attracts around ten times more Google searches than TIMSS when their respective results are released (see Figure 1.5.3). Yet there is no clear methodological reason why PISA should receive such disproportionate attention. It is not inherently more robust than TIMSS; in many ways, it has simply benefited from more effective marketing and greater public visibility. A key message is therefore that users and consumers of ILSAs should, wherever possible, triangulate findings across different studies rather than automatically prioritising evidence from PISA.

Figure 1.5.3. Google searches made for PISA and TIMSS when the results are released



1.6. Challenges with using data from large-scale international assessments.

ILSAs have the potential to provide important insights into key educational issues and into how these vary across countries. However, they also have important limitations, many of which are inherent to their design and difficult to fully overcome. In this section, we provide an overview of some of these key issues.

1.6.1. Cross-national comparability is a spectrum rather than being absolute

Considerable effort is made to ensure ILSAs are as comparable as possible via common assessment frameworks, standardised procedures, translation protocols, sampling rules and technical quality standards. Yet, for some variables, full comparability is simply difficult to achieve. As education systems differ in their educational qualifications, languages, cultures and social contexts, some measurement differences cannot be fully removed via careful study design. Indeed, for most measures cross-national comparability should be seen as sitting along a spectrum rather than simply being “comparable” or not.

A classic example is educational qualifications. In ILSAs, these are often measured through parental education (e.g. PISA, PIRLS and TIMSS) or through adults’ own qualifications (e.g. PIAAC). This may seem a relatively straightforward issue at first glance, since the International Standard Classification of Education (ISCED) framework provides an international classification of education levels across countries. It may thus appear straightforward to then map national qualifications onto this framework. Yet, in practice, this is much more difficult to achieve. Research using other international surveys has shown that the coding of educational attainment into the ISCED schema can be inconsistent across countries, between different studies and even in the same study over time (Ortmanns & Schneider, 2016; Ortmanns, 2020). These inconsistencies can create misleading changes in the apparent distribution of educational attainment over time and produce differences between international studies that reflect coding practices rather than genuine variation. Thus, as Ortmanns and Schneider (2016) put it “educational attainment is one of the most difficult background variables to measure and code in a comparable and coherent way across countries and studies.” There is no simple solution to this problem; some uncertainty surrounding cross-national comparisons of educational qualifications (and thus also differences in ILSA scores and other outcomes across parental education groups) will remain.

Several other examples also exist, including variables that one might previously have taken for granted. Take gender, for instance. Historically, most ILSAs used a binary categorisation (male/female), making comparisons of gender gaps in achievement and other outcomes relatively straightforward. Yet this is now beginning to change. As noted by Jerrim (2026), in TIMSS 2023 some participating systems included response options beyond “boy” and “girl”. For example, in some systems a third category was included, labelled “other” in some countries while being phrased more specifically as “non-binary” or “another gender” in others. There is hence now variation across countries not only in whether more than two gender response categories are provided, but also in how these additional categories are described. Moreover, such differences are likely to reflect cross-national differences in wider cultural and political contexts. This in turn introduces complications for cross-national comparisons of gender gaps which are not straightforward to resolve.

Thus, for applied users and consumers of ILSAs, my advice is to always remember that “comparability” of the data always operates along a scale. Some of the things being measured are inherently more difficult to make comparable across countries and over time than others. So, while this is a central goal of ILSAs, for many measures it can only be partially achieved.

1.6.2. Change in measurement over time is inevitable

Another major purpose of ILSAs is to facilitate the study of trends over time. TIMSS now provides data going back to 1995, PISA to 2000 and PIRLS to 2001. These studies therefore already cover a substantial period and there are few signs they are going away; if anything, participation in several ILSAs has increased over time.

This creates an important trade-off. On the one hand, if the aim is to measure long-term trends in education, then consistency is vital. To measure change, one should not change the measure. On the other hand, the issues facing societies and the technology we use do not stand still. It is thus inevitable that some of the questions included in background questionnaires may become more or less relevant over time, with it also possible that the meaning of apparently stable measures may also change (e.g. recall the example of gender in the section 1.6.1 above). For example, the number of books at home may not capture the same cultural or educational resources today as it did thirty years ago. ILSAs also need to make use of technological innovations, both to improve assessment quality and to reflect changes in how pupils learn.

A good example is the shift to computer-based testing. This brings many advantages. It allows some questions to be scored automatically, makes it possible to include more interactive item formats and may better reflect the fact that young people are increasingly working on digital devices. It also enables adaptive assessment designs which can increase the reliability of measurement without increasing test time (see chapter 6 - section 6.9 - for further discussion). However, this creates a challenge for measuring trends. Taking a test on paper is not the same as taking a test on a computer. This change therefore cuts across the principle that, if we want to measure change over time, we should avoid changing the measure.

Nevertheless, this transition has already happened in several ILSAs. Earlier rounds were typically administered using pen-and-paper tests, while PISA began moving its main assessment to computer-based delivery in 2015. This has led to debate about the extent to which results remain comparable over time. Jerrim et al. (2018), for example, discuss evidence from a randomised field experiment conducted as part of the PISA 2015 field trial. Their analysis indicated that pupils tended to perform less well on the computer-based version of the PISA assessment than on an otherwise equivalent paper-based version. This raises important questions about how far the shift in assessment mode may have affected changes in PISA scores pre/post the 2015 cycle. For instance, the OECD average science score fell by 8 points between 2012 and 2015 (pre/post the introduction of computer-based assessment) compared to a fall of only 2 points between 2018 and 2022 (pre/post the COVID-19 pandemic)⁴.

Yet the move to computer-based assessment was always likely to happen at some point. Similar changes in measurement are also likely in the future. The key is therefore not to pretend that ILSA measures can remain completely unchanged forever; one must recognise that some methodological change is an inevitability. What matters, then, is that such changes are accompanied by robust studies of their impact, with the findings transparently reported. As such, it is also likely that comparisons of longer-term trends will be somewhat more vulnerable to methodological changes than comparisons over shorter time horizons.

1.6.3. Test effort and survey satisficing behaviour

ILSAs are low-stakes assessments and surveys. Pupils and schools have nothing directly riding on the results. Indeed, because of the complex assessment design, pupils and schools typically do not even receive individual results (see chapter 6 for further discussion of why this is the case).

⁴ This figure is based on a consistent set of 35 OECD countries.

Now imagine being a teenager asked to sit a test where nothing depends on the outcome and where nobody (you, your teacher or your parents) is going to find out how you performed. Would you really be motivated to try your absolute hardest? For many pupils, the honest answer may be no. This low-stakes nature of the assessment is therefore a major challenge for ILSAs. Country-level differences in achievement scores may partly reflect genuine differences in young people's knowledge and skills, but they may also reflect differences in test motivation and effort. The problem is that it is then difficult to tease these two explanations apart.

There is also interesting academic evidence supporting this point. Gneezy et al. (2019) conducted an experimental study with pupils from two high schools in the United States and four high schools in Shanghai. Pupils were asked to complete a 25-minute, 25-question mathematics test using items drawn from previous PISA assessments. However, unlike in the actual PISA assessment, some pupils were given a financial incentive to perform well. Specifically, pupils in the treatment group were given the equivalent of \$25 at the start of the test but were told that \$1 would be taken away for each question answered incorrectly (or left unanswered). The intuition is that this raised the stakes of the assessment for this group. Their performance was then compared with that of a randomly assigned control group who had no financial incentive to do well. The authors found that the incentive had a positive and statistically significant effect in the US sample, but not in Shanghai. This is consistent with the idea that part of the difference in PISA results between these jurisdictions may reflect differences in test effort (and not just differences in their knowledge and skills). Indeed, the authors estimate that US mathematics PISA scores would be around 22 to 24 points higher if pupils were motivated to do well⁵.

The low-stakes nature of ILSAs may also impact responses to the accompanying background questionnaires. After completing a long assessment (e.g. two hours in the case of PISA) pupils are then asked to complete these questionnaires. There is again little personal incentive for respondents to exert maximum effort when they respond. Some may therefore engage in survey “satisficing” behaviours, providing answers that are quick and easy rather than carefully considered. This issue has been discussed in ILSA research, including studies of questionnaire response times and suspicious response behaviour such as “straight-lining” (Lundgren & Eklöf, 2023; OECD, 2024b). Moreover, as with test effort, the extent of questionnaire satisficing may vary

⁵ However, the samples of schools and pupils included in the Gneezy et al. (2019) study were small and not nationally representative, meaning this evidence should be interpreted as suggestive rather than definitive.

across countries and over time. This may then create erroneous differences between countries and across cycles in questionnaire-based measures.

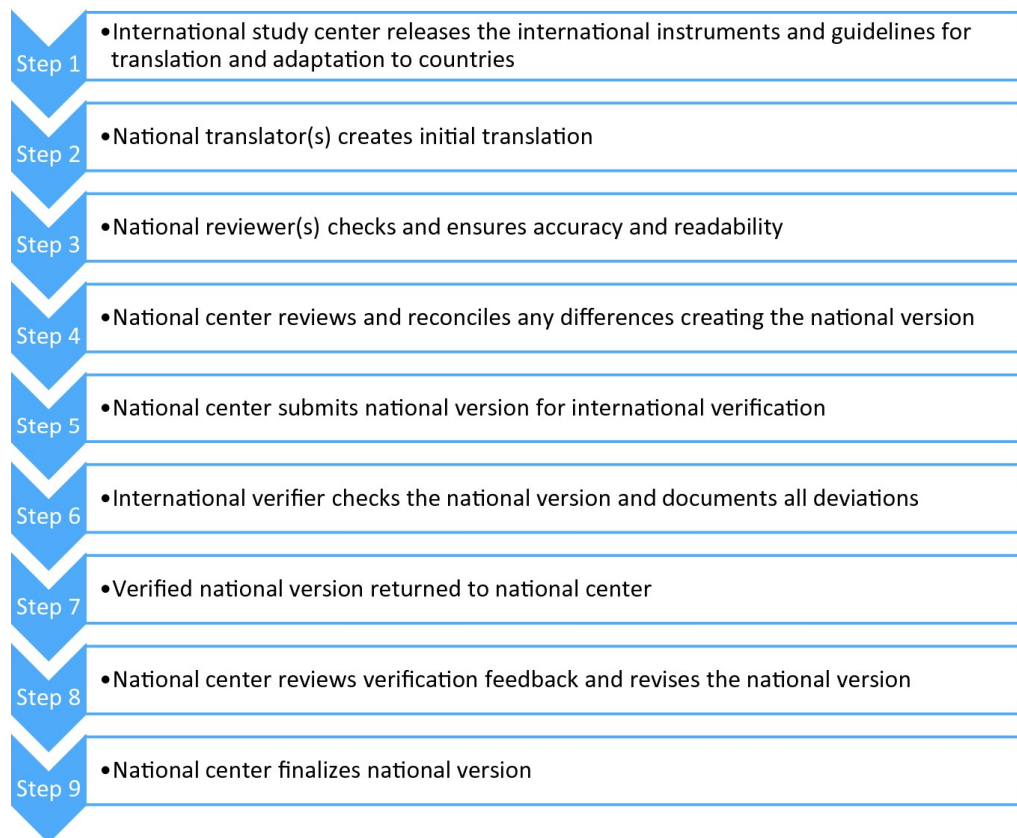
To their credit, the OECD (2023) has discussed this issue at length, recognising the challenge that “low effort” poses for ILSAs. In particular, they acknowledge that changes in PISA test scores over time and differences between countries may partly reflect variation in pupil engagement with the assessment rather than in knowledge and skills alone. For instance, they highlight the case of changes in PISA scores for Albania between 2018 and 2022, concluding that “the decline in performance in Albania – one of the largest ever registered in PISA – reflects, at least in part, the absence of student engagement”.

1.6.4. The challenges with translation

As one would expect in international studies covering many countries, ILSAs invest substantial time and effort in translation and adaptation. The aim is to ensure that assessment and questionnaire items are interpreted consistently and carry comparable meaning across diverse national contexts. Koršňáková et al. (2020) provide an overview of this process for IEA studies, reproduced in Figure 1.6.4.1 below. In summary, the international study organisers develop source versions of the assessment items and questionnaire materials, typically in English and (in some studies) also another source language such as French. These materials are then translated and adapted into national versions, which are reviewed by those responsible for the study in each participating country. The national versions are subsequently submitted to the international co-ordinators for verification against the source versions. Feedback is then returned to national teams who make any required revisions before finalising the instruments to be used in the field.

Yet this process is not perfect, and issues surrounding translation, adaptation and linguistic comparability are likely to remain. An academic literature has documented such issues and the broader topic of whether the assessment and questionnaire items carry equivalent meaning across languages and national contexts. Arffman (2013), for instance, reviews problems that can arise when international educational achievement tests are translated. El Masri et al. (2016) provide a particularly clear example in their comparison of English, French and Arabic versions of PISA science items. They argue that features of particular items may place different cognitive demands on respondents depending on the language they take the test. For example, timelines and numbers on graphs are typically read from left to right, matching the direction of English and French text. In Arabic, however, text is read from right to left, potentially creating additional processing demands.

Figure 1.6.4.1. Process for developing ILSA assessment and survey instruments



Source: Korsnakova et al. (2020, Figure 6.1)

Other studies have raised related concerns. Jerrim, Lopez-Agudo and Marcenaro-Gutierrez (2022) find that pupils in Wales who took PISA in Welsh scored around 30 points (0.3 standard deviations) lower than otherwise similar peers who took the test in English. They suggest that translation and test-language assignment are possible explanations (amongst others). Asil and Brown (2016) find evidence of measurement non-invariance in PISA reading when comparing Australia with other economies, suggesting caution in some cross-national comparisons. In PIRLS, Goodrich and Ercikan (2019) identify differences between English and French versions of the assessment used in Canada, supported by statistical evidence and reviews by four bilingual experts.

Taken together, these studies suggest that (despite the time, money and effort invested in translation within ILSAs) not all issues of linguistic comparability are likely to be resolved. For applied researchers, I therefore again recommend they think of ILSAs as sitting along a spectrum of comparability, rather than treating them as either simply “comparable” or not. Moreover, particularly with respect to language, some comparisons are likely to be further along this spectrum than others. For instance, comparisons between countries using the same assessment language are logically likely to raise fewer translation-related concerns. On the other hand,

comparisons involving languages that differ substantially are likely to require a greater degree of caution with respect to potential translation issues.

1.6.5. Process issues

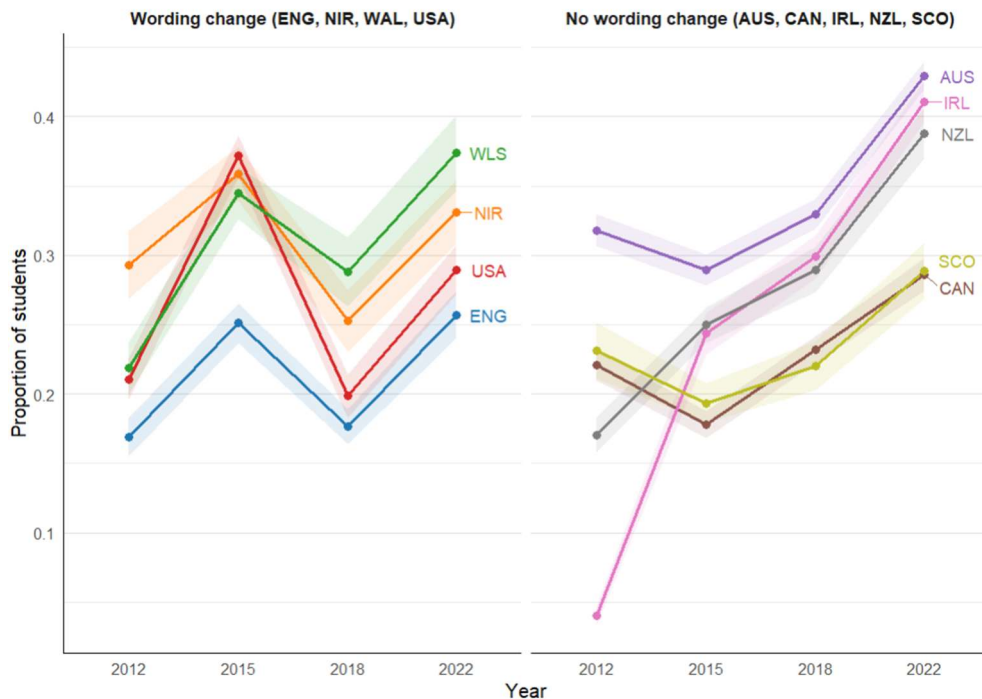
By now it should be clear that ILSAs are large and complicated undertakings. They involve many organisations and individuals, working across multiple countries and languages, often within tight timeframes. Great efforts are made to apply rigorous technical standards and quality-control procedures. Nevertheless, as with any large survey or complex international project, things sometimes go wrong. Such issues may then go on to impact the ILSA data available for analysis and the subsequent inferences that researchers and policymakers may draw.

Jerrim et al. (2026) provide one such case study. They were interested in studying truancy using PISA data over time, particularly in the period before and after the COVID-19 pandemic. However, during their investigations, they discovered a problem. In a selection of countries, the wording of the relevant PISA question had changed in the 2015 cycle. Specifically, in the United States, England, Northern Ireland and Wales, the item changed from “In the last two full weeks of school, how often: I **skipped** a whole school day” to “In the last two full weeks of school, how often: I **missed** a whole school day” [emphasis added by author]. The same change, however, did not occur in other English-speaking countries, such as Scotland and Australia.

Jerrim et al. (2026) went on to note how this materially changed the meaning of the survey question, producing the pattern shown in Figure 1.6.5.1. In the countries where the wording changed (left-hand side), there was a sizeable increase in apparent truancy in the 2015 wave. Indeed, the increase was so large that the further rise in reported truancy after the COVID-19 pandemic in 2022 did not appear especially unusual. Yet a rather different pattern was observed in countries without the wording change, shown on the right-hand side of Figure 1.6.5.1. Here, there was no analogous sharp increase in 2015, making the post-pandemic rise in reported truancy much easier to see. Thus, had this process error in the PISA survey questions not been identified, the important increase in truancy rates in England, Northern Ireland, Wales and the United States would likely have been missed.

For policymakers, government officials and applied researchers, this example highlights the need to remain alert to such process issues in ILSA data. Not everything can be taken on trust. When patterns in the data do not look quite right, it is important to investigate them carefully rather than assuming they reflect genuine changes or differences.

Figure 1.6.5.1. The impact of the wording change to international comparisons of pupil truancy. (Missed versus skipped school).



Source: Adapted from Jerrim et al. (2026, Figure 1).

1.7. A note on software.

Fortunately, a wide range of software packages are now available to handle much of the technical work involved in analysing ILSA data. These include the following packages for the major statistical software environments detailed below. In doing so, we reiterate however that the purpose of this book is not to show readers which buttons to press or commands to run. Rather, it aims to build an intuitive understanding of what ILSA software packages are doing behind the scenes, why particular analytical procedures are required and how the resulting estimates should be interpreted. This is important because, although modern software handles much of the technical work, researchers must still understand the assumptions, limitations and caveats that remain after the analysis has been completed.

1.7.1 STATA.

- **Svysset** (inbuilt command). in The STATA survey commands (svysset) allows for the handling of complex survey designs, including the replicate weights discussed in Chapter 3.
- **Repest**. A flexible command that facilitates the use of both plausible values (see chapter 6, section 6.5) and replicate weights (see Chapter 3). While originally designed for PISA, PIAAC and TALIS, it can also be used with TIMSS and PIRLS. (Avvisati & Keslair, 2014).

- **pv.** A user-written package designed specifically to pool analyses across plausible values in PISA, TIMSS, PIRLS and PIAAC, while allowing a wide range of underlying STATA estimation commands (Macdonald, 2008)
- **pisatools and piaac tools.** A PISA and PIAAC-specific collection covering descriptive statistics, linear and quantile regression, general estimation commands, Oaxaca decompositions and other decomposition analyses, with replicate weights and plausible values handled automatically. (Jakubowski & Pokropek, 2013; Jakubowski & Pokropek, 2019).

1.7.2 R.

- **Rrepest.** The closest R equivalent to STATA's **repest**. It supports BRR, jackknife replicate weights and pooling across plausible values. It provides descriptive statistics, correlations, regressions and user-defined analyses. The package has been designed for use with both OECD and IEA studies. (Ilizaliturri et al., 2026).
- **RALSA.** Imports, merges, prepares and analyses data from assessments including PISA, TIMSS and PIRLS, while applying the appropriate complex-sampling and plausible-value procedures. It includes a graphical interface, making it particularly suitable for users who do not wish to write extensive R code (Mirazchiyski, 2026).
- **intsvy.** Provides study-specific functions for importing, merging and analysing ILSA data. Its functions cover means, frequencies, correlations and regressions while automatically incorporating replicate weights and plausible values. (Caro & Biecek, 2017)

1.7.3 SPSS

- **IEA IDB Analyzer.** A free Windows application that generates syntax (SPSS, R and SAS) for analysing TIMSS, PIRLS, ICCS, ICILS, PISA, PIAAC, TALIS and several other large-scale assessments. It automatically selects the appropriate sampling and replicate weights, combines plausible values and supports means, percentages, percentiles, benchmarks, correlations and linear/ logistic regression. (Gonzalez, 2025).

1.8. Test your knowledge questions

1.8.1 Task 1.

Table 1.8.1.1 provides data for 20 countries.

Table 1.8.1.1 Data on 20 hypothetical countries

Country	Income inequality (gini). 2010.	Socio-economic gap in test scores (age 9). 2010	Socio-economic gap in test scores (age 15). 2016	Socio-economic gap in test scores (age 15). 2025	Annual national assessments in primary school (before age 11)	Annual national assessments in secondary school (after age 11)
A	0.15	65	75	90	No	No
B	0.40	45	50	70	No	No
C	0.18	70	75	87	No	No
D	0.22	40	50	63	No	No
E	0.13	60	80	97	No	Yes
F	0.27	25	35	49	No	No
G	0.22	30	35	50	No	No
H	0.43	65	90	106	No	Yes
I	0.18	45	55	69	No	No
J	0.23	35	45	59	No	No
K	0.13	55	80	93	No	Yes
L	0.42	40	50	66	No	No
M	0.38	40	85	97	No	Yes
N	0.35	30	35	46	No	No
O	0.28	55	60	70	No	No
P	0.21	40	50	63	No	No
Q	0.10	35	70	85	No	Yes
R	0.19	40	60	73	No	Yes
S	0.08	35	55	71	No	Yes
T	0.17	45	55	70	No	No

The variables are defined as follows:

- Income inequality (gini). 2010). A measure of income inequality in the country (the gini coefficient) measured in 2010. Gini coefficients take a value between 0 and 1, with higher values indicating greater levels of income inequality.
- Socio-economic gap in test scores (age 9). 2010. The difference in test scores between children from high and low income backgrounds at age 9. Measured in 2010.
- Socio-economic gap in test scores (age 15). 2016. The difference in test scores between children from high and low income backgrounds at age 15. Measured in 2016.
- Socio-economic gap in test scores (age 15). 2016. The difference in test scores between children from high and low income backgrounds at age 15. Measured in 2025.
- Annual national assessments in primary school (before age 11). A binary indicator of whether the country conducts national assessments of pupils in primary school (prior to age 11).

- Annual national assessments in secondary school (after age 11). A binary indicator of whether the country conducts national assessments of pupils in secondary school (after to age 11).

Task 1. Question 1. Please come up with a research question using these data falling into each of the following categories:

1. Benchmarking
2. Links with macro (country-level) factor
3. The impact of institutional features
4. Generalisability
5. Change over time

Task 1. Question 2. Please produce a simple descriptive analysis (i.e. use descriptive statistics only) to provide a first insight into answering each question.

1.8.2 Task 2.

Table 1.8.2.1 refers to national-specific sample survey data collected within countries. These data have been ex-post harmonised by a researcher for the purpose of comparing the relationship between material education and academic achievement of 15-year-olds across the two countries.

Table 1.8.2.1. Comparisons of the association between maternal education and teenagers test scores from national specific ex-post harmonised data

Mother education	Country A		Country B	
	%	Score	%	Score
Less than high school	15%	350	33%	450
High school diploma	15%	450	33%	500
University degree	70%	520	33%	550

Task 2. Question 1. Please give three reasons why the distribution of mother's education may differ across the two countries. At least two of these should be related to how the data have been collected and processed.

Task 2. Question 2. A policymaker from country A says the following: "socio-economic inequalities in this country are much greater than in country B. The gap between children whose mother holds a degree versus those that did not complete high school is 170 points, compared to only 100 points in country B. We need to invest more in the most disadvantaged children into order to get this gap down." What is your evaluation of this statement?

Task 2. Question 3. Another researcher has access to ex-ante harmonised data across the two countries. While this dataset does not include information on mother education, it does include student reports of the number of books within their home. Results from the analysis they produce are presented in Table 1.8.2.2.

Table 1.8.2.2. Comparisons of the association between books at home and teenagers test scores from ex-ante harmonised ILSA data

Number of books at home	Country A		Country B	
	%	Score	%	Score
0-20 books at home	31%	452	33%	448
21-100 books	35%	498	31%	499
More than 100 books	33%	552	35%	549

Task 2. Question 3a. Please now interpret how socio-economic inequality in educational achievement compares across the two countries.

Task 2. Question 3b. Please give two advantages and two disadvantages of this analysis, as compared to the previous analysis that used information on maternal education.

Chapter 1 references

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educational Measurement: Issues and Practice*, 32(2), 2–14.
<https://doi.org/10.1111/emip.12007>

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, 16(1), 71–93.
<https://doi.org/10.1080/15305058.2015.1064431>

Avvisati, F., & Keslair, F. (2014). REPEAT: Stata module to run estimations with weighted replicate samples and plausible values (Statistical Software Components No. S457918) [Computer software]. Boston College Department of Economics.
<https://ideas.repec.org/c/boc/bocode/s457918.html>

Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2015). *Too many children left behind: The U.S. achievement gap in comparative perspective*. Russell Sage Foundation.

Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software*, 81(7), 1–44.
<https://doi.org/10.18637/jss.v081.i07>

Charalambous, C. Y., Grob, U., Praetorius, A.-K., Köhler, C., & Miao, Z. (2025). Are direct effects of cognitive activation on student outcomes enough? Exploring the role of adaptation and use of opportunities. *Learning and Instruction*, 100, Article 102176. <https://doi.org/10.1016/j.learninstruc.2025.102176>

Chung, J. H. (2010). Finland, PISA, and the implications of international achievement studies on education policy. In A. W. Wiseman (Ed.), *The impact of international achievement studies on national education policymaking* (International Perspectives on Education and Society, Vol. 13, pp. 267–294). Emerald Group Publishing Limited. [https://doi.org/10.1108/S1479-3679\(2010\)0000013013](https://doi.org/10.1108/S1479-3679(2010)0000013013)

Durlauf, S.; Kourtellis, A. & Tan, C-M. (2022). The Great Gatsby Curve. *Annual Review Economics*. 14:571-605. <https://doi.org/10.1146/annurev-economics-082321-122703>

El Masri, Y. H., Baird, J. A., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, 23(4), 427–455. <https://doi.org/10.1080/0969594X.2016.1218323>

Fernández Darraz, M., & Mora Guerrero, G. (2017). Implication of gender in the reading test of the Education Quality Measurement System (SIMCE). Analysis of gender, geographic origin, socioeconomic group, and ethnicity in Araucanía region. *Estudios Pedagógicos*, 43(2), 65–91. <https://doi.org/10.4067/S0718-07052017000200004>

Gillard, J. (2012). Opening remarks to APEC CEO Summit. PM Transcripts. <https://pmtranscripts.pmc.gov.au/release/transcript-18792>

Gonzalez, E. (2025). The IEA's IDB Analyzer. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative digital-based international large-scale assessments* (pp. 587–605). Springer. https://doi.org/10.1007/978-3-031-90951-1_24

Goodrich, S., & Ercikan, K. (2019). Measurement comparability of reading in the English and French-Canadian populations: Special case of the 2011 Progress in International Reading Literacy Study. *Frontiers in Education*, 4, Article 120. <https://doi.org/10.3389/feduc.2019.00120>

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, 1(3), 291–308.

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, 116(510), C63–C76. <https://doi.org/10.1111/j.1468-0297.2006.01076.x>

Herbert, B., Fischer, J., & Klieme, E. (2022). How valid are student perceptions of teaching quality across education systems? *Learning and Instruction*, 82, Article 101652. <https://doi.org/10.1016/j.learninstruc.2022.101652>

Hofmeyr, H. (2022). Why do girls do better? Unpacking South Africa's gender gap in PIRLS and TIMSS. *International Journal of Educational Development*, 94, Article 102648. <https://doi.org/10.1016/j.ijedudev.2022.102648>

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, 62(3), 333–353. <https://doi.org/10.1080/00313831.2016.1258726>

Ilizaliturri, R., Avvisati, F., & Keslair, F. (2026). Rrepest: An analyzer of international large scale assessments in education (Version 1.6.14) [R package]. Comprehensive R Archive Network. <https://doi.org/10.32614/CRAN.package.Rrepest>

International Association for the Evaluation of Educational Achievement. (n.d.). *Early IEA studies*. Retrieved July 16, 2026, from <https://www.iea.nl/studies/iea/earlier>

Jackson, M. (Ed.). (2013). *Determined to Succeed?: Performance versus Choice in Educational Attainment* (1st ed.). Stanford University Press. <https://doi.org/10.2307/j.ctvqsdjrj>

Jakubowski, M. (2010). Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data. In: Dronkers, J. (eds) *Quality and Inequality of Education*. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3993-4_3

Jakubowski, M., & Pokropek, A. (2013). PISATOOLS: Stata module to facilitate analysis of the data from the PISA OECD study. Statistical Software Components S457754, Boston College Department of Economics, revised 06 May 2020.

Jakubowski, M., & Pokropek, A. (2019). piaactools: A program for data analysis with PIAAC data. *The Stata Journal*, 19(1), 112–128. <https://doi.org/10.1177/1536867X19830909>

Jerrim, J. (2014), Unrealistic Educational Expectations. *The Sociological Quarterly*, 55: 196–231. <https://doi.org/10.1111/tsq.12049>

Jerrim, J., Parker, P. D., Chmielewski, A. K., & Anders, J. (2016). Private schooling, educational transitions, and early labour market outcomes: Evidence from three Anglophone countries. *European Sociological Review*, 32(2), 280–294. <https://doi.org/10.1093/esr/jcv098>

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the ‘mode effect’ and what has been done about it? *Oxford Review of Education*, 44(4), 476–493. <https://doi.org/10.1080/03054985.2018.1430025>

Jerrim, J., & Sims, S. (2019). How do academically selective school systems affect pupils’ social-emotional competencies? New evidence from the Millennium Cohort Study. *American Educational Research Journal*, 56(5), 1769–1799. <https://doi.org/10.3102/0002831219830965>

Jerrim, J. (2023). Teacher surveys: The pros and cons of random probability surveys versus teacher panels. *Review of Education*, 11, e3428. <https://doi.org/10.1002/rev3.3428>

Jerrim, J., Parker, P. D., & Shure, N. (2023). Overclaiming. An international investigation using PISA data. *Assessment in Education: Principles, Policy & Practice*, 30(5–6), 351–371. <https://doi.org/10.1080/0969594X.2023.2238248>

Jerrim, J. (2024). Has Peak PISA passed? An investigation of interest in International Large-Scale Assessments across countries and over time. *European Educational Research Journal*, 23(3), 450-476.

Jerrim, J., Prieto-Latorre, C., Marcenaro-Gutierrez, O. D., & Shure, N. (2025). Teacher Self-Efficacy, Instructional Practice, and Student Outcomes: Evidence from the TALIS Video Study. *American Educational Research Journal*, 62(2), 378-413.

Jerrim, J. (2026). School belonging amongst transgender teenagers. International evidence from TIMSS.

Jerrim, J., Palma Caravajal, M., Anders, J., Ladrón de Guevara Rodríguez, M., & Marcenaro-Gutierrez, O. (2026). When technical standards fail: Questionnaire wording and the comparability of PISA truancy data.

Koršňáková, P., Dept, S. & Ebbs, D. (2020). Translation: The Preparation of National Language Versions of Assessment Instruments. In: Wagemaker, H. (eds) Reliability and Validity of International Large-Scale Assessment. IEA Research for Education, vol 10. Springer, Cham. https://doi.org/10.1007/978-3-030-53081-5_6

- Lundgren, E., & Eklöf, H. (2023). Questionnaire-taking motivation: Using response times to assess motivation to optimize on the PISA 2018 student questionnaire. *International Journal of Testing*, 23(4), 231–256. <https://doi.org/10.1080/15305058.2023.2214647>
- Manu, M., Torppa, M., Vasalampi, K., Lerkkanen, M.-K., Poikkeus, A.-M., & Niemi, P. (2023). Reading development from kindergarten to age 18: The role of gender and parental education. *Reading Research Quarterly*, 58, 505–538. <https://doi.org/10.1002/rrq.518>
- Marsh, H. W., & Hau, K.-T. (2003). Big-Fish--Little-Pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, 58(5), 364–376. <https://doi.org/10.1037/0003-066X.58.5.364>
- Macdonald, K. (2008). PV: Stata module to perform estimation with plausible values (Statistical Software Components No. S456951) [Computer software]. Boston College Department of Economics. <https://ideas.repec.org/c/boc/bocode/s456951.html>
- Maehler, D. B., Hernández-Torrano, D., Courtney, M. G. R., Fischer, F. P., Hollricher, L. F., & Gorges, J. (2025). PIAAC Survey of Adult Skills: A review of the research landscape. *International Review of Education*, 71, 685–714. <https://doi.org/10.1007/s11159-024-10123-4>
- Mejía-Rodríguez, A. M., Rožman, M., & Strietholt, R. (2026). Three decades of international studies on computers and education: a scoping review of the literature. *Scandinavian Journal of Educational Research*, 1–18. <https://doi.org/10.1080/00313831.2026.2642590>
- Mirazchiyski, P. V. (2026). RALSA: R analyzer for large-scale assessments (Version 1.6.6) [R package]. Comprehensive R Archive Network. <https://doi.org/10.32614/CRAN.package.RALSA>
- Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading*. Boston College, TIMSS & PIRLS International Study Center. <https://doi.org/10.6017/lse.tpisc.tr2103.kb5342>
- Myoung, E., & Liou, P. Y. (2022). Systematic review of empirical studies on international large-scale assessments of civic and citizenship education. *Scandinavian Journal of Educational Research*, 66(7), 1269–1291. <https://doi.org/10.1080/00313831.2022.2131903>
- Nilsen, T., Stancel-Piątak, A., & Gustafsson, J.-E. (Eds.). (2022). *International handbook of comparative large-scale studies in education: Perspectives, methods and findings*. Springer Cham. <https://doi.org/10.1007/978-3-030-88178-8>
- OECD. (2009). *PISA data analysis manual: SPSS, second edition*. OECD Publishing. <https://doi.org/10.1787/9789264056275-en>

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>

OECD. (2023). Annex A8. How much effort do students put into the PISA test? In *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>

OECD. (2024). *Education at a glance 2024: Sources, methodologies and technical notes*. OECD Publishing. <https://doi.org/10.1787/e7d20315-en>

OECD (2024b). PISA 2022 Technical Report, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/01820d6d-en>.

Oo, C. Z., Eryilmaz, N., & Strietholt, R. (2026). Do findings from a single country generalise internationally? Evidence from an international assessment. *Studies in Educational Evaluation*, 90, Article 101618. <https://doi.org/10.1016/j.stueduc.2026.101618>

Ortmanns, V. & Schneider, S. L. (2016). Harmonization still failing? Inconsistency of education variables in cross-national public opinion surveys. *International Journal of Public Opinion Research*, 28(4), 562–582.

Ortmanns, V. (2020). Explaining inconsistencies in the education distributions of ten cross-national surveys – the role of methodological survey characteristics. *Journal of Official Statistics*, 36(2), 379–409.

Pickett, K. (2024). *The Spirit Level at 15: Technical appendix*. The Equality Trust. <https://media.equality-trust.out.re/uploads/2024/07/Technical-Appendix-20-07.pdf>

President of the Republic of Uzbekistan. (2019, August 23). School education development should turn into a nationwide movement. <https://president.uz/en/lists/view/2820>

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *The American psychologist*, 74(4), 445–458. <https://doi.org/10.1037/amp0000356>

Ruhose, J., & Schwerdt, G. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. *Economics of Education Review*, 52, 134–154. <https://doi.org/10.1016/j.econedurev.2016.02.004>

Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the policy influence of international large-scale assessments in education. In H. Wagemaker (Ed.), *Reliability and*

validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement (pp. 261–277). Springer. https://doi.org/10.1007/978-3-030-53081-5_15

Stiff, J., Lenkeit, J., Hopfenbeck, T. N., Kayton, H. L., & McGrane, J. A. (2023). Research engagement in the Progress in International Reading Literacy Study: A systematic review. *Educational Research Review*, 40, Article 100547. <https://doi.org/10.1016/j.edurev.2023.100547>

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: Difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability*, 33, 139–167. <https://doi.org/10.1007/s11092-020-09346-4>

UAE Government. (2024). 4. Quality education. The Official Portal of the UAE Government. <https://u.ae/en/about-the-uae/leaving-no-one-behind/4qualityeducation>

Veletić, J., Mejía-Rodríguez, A. M., & Olsen, R. V. (2024). A systematic literature review of TALIS secondary research: Trends and future directions. *Review of Education*, 12, e3469. <https://doi.org/10.1002/rev3.3469>

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *TIMSS 2023 Technical Report (Methods and Procedures)*. Boston College, TIMSS & PIRLS International Study Center. <https://timss2023.org/methods>

Washbrook, E., Waldfogel, J., Bradbury, B., Corak, M. & Ghanghro, A.A. (2012), The Development of Young Children of Immigrants in Australia, Canada, the United Kingdom, and the United States. *Child Development*, 83: 1591-1607. <https://doi.org/10.1111/j.1467-8624.2012.01796.x>

Wilkinson, R., & Pickett, K. (2009). *The spirit level: Why more equal societies almost always do better*. Allen Lane.

Xu, W., Chen, J., Chen, X., & Li, X. (2026). A bibliometric and thematic analysis of PISA literature for school improvement and effectiveness over two decades. *International Journal of Educational Research*, 138, Article 103012. <https://doi.org/10.1016/j.ijer.2026.103012>

2. Complex survey designs

2.1 Chapter introduction

Researchers are often interested in describing characteristics of a well-defined population across several countries. For instance, in the context of international large-scale assessments, the aim may be to estimate the average knowledge and skills of pupils of a particular age in a specific subject (e.g. 9/10-year-olds in mathematics). Alternatively, the statistic of interest could be the percentage of pupils who lack basic reading skills or the percentage of teachers who report being dissatisfied with their workload. It is of course not feasible to test or survey every individual of the group of interest in every participating country. Doing so would be prohibitively expensive, administratively complex and logistically unrealistic. Instead, researchers must collect information from a subset of the population. This subset is selected through a random - or more precisely probabilistic - sampling process. One can then use data gathered from this sample to make inferences about the broader population from which it was drawn.

The random/probabilistic nature of the sample is critical for two main reasons. First, in expectation⁶, the selected units will be representative of the target population. This applies not only to observable characteristics (e.g. gender, geographic region, school type) but also to unobservable characteristics that researchers cannot directly measure. Consequently, estimates from the sample (e.g. mean mathematics achievement; the proportion of pupils reporting a positive school climate) can be extrapolated through to the target population within the country as a whole. Second, random sampling provides the foundation for quantifying uncertainty stemming from sampling variation. Specifically, one is likely to produce a different (albeit likely similar) estimate had a different random sample been drawn. Standard errors summarise this sampling variability and therefore indicate the degree of uncertainty surrounding an estimated population value. This in-turn also underpins all widely used forms of statistical inference, including confidence intervals, t-statistics, p-values and tests of statistical significance.

For these reasons, all countries are required to use some form of random/probabilistic sampling to select units from their target populations when they participate in an ILSA. Introductory statistics textbooks and courses introduce the idea of simple random sampling, in which every individual in the population has an equal probability of selection and are selected independently.

⁶ “In expectation” here refers to the fact that – if one were able to do the sampling and collect the data many times – the average of the estimates obtained would be equal to the true population value.

Although this design is conceptually useful, it is rarely practical in large-scale social surveys and data collections (including ILSAs). For instance, it would be inefficient and burdensome to ask a school to organise testing for only one randomly selected pupil. Most social data collections (including ILSAs) thus typically rely on more complex survey designs. These may involve stratification (where the population is divided into subgroups before sampling), clustering (where groups such as schools or classrooms are selected before individuals) and oversampling (where certain groups are sampled at higher rates to ensure adequate representation). The purpose of this chapter is to introduce readers to these ideas, explain how they are used in ILSAs and consider their implications for statistical analysis.

2.2 Definition of the target population.

The target population refers to the group of individuals to whom researchers want the results of a study to generalise. Put differently, it defines the population about which the study aims to make claims. In some cases, this population may be very broad. For example, in the Programme for the International Assessment of Adult Competencies (PIAAC), the target population is non-institutionalised adults aged 16–65 who are usual residents of the participating country⁷. In other settings (i.e. outside of ILSAs), the target population may be much more specific, such as second-generation immigrants aged 8 living in urban areas. The key point is that in high-quality data collections, the target population is clearly defined before sampling begins.

In ILSAs, the target population is usually broad, though typically focused on pupils of a particular age or grade. This distinction – between age and grade – is important. PISA uses an age-based target population, aiming to assess the knowledge and skills of pupils aged between 15 years and 3 months and 16 years and 2 months. TIMSS on the other hand has a grade-based target population, focusing on pupils in the fourth and eighth years of formal schooling (usually Grades 4 and 8). This is not a trivial difference. In countries where grade retention is common (i.e. where lower-achieving pupils may repeat a school year) an age-based sample such as PISA may include pupils spread across several school grades. A grade-based sample, such as TIMSS, instead includes pupils of different ages within the same grade. As different countries use grade repetition to differing degrees and in differing ways, such variation in the definition of the target population may plausibly impact cross-national comparisons.

⁷ Some countries additionally sample adults aged 66–74, but these supplementary groups fall outside the standard PIAAC target population.

The precise definition of the target population can matter greatly. In practice, the “devil is often in the detail”. A useful example comes from PISA. Its target population is defined as “students 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 country, and in grade 7 or higher” (Westat, 2016). This precise wording contains several important points. Most notably, PISA is not designed to represent all 15-year-olds living in a country (despite this being how its results are widely interpreted). Rather, it represents those in this age range who are attending educational institutions and who are in grade 7 or above. As a result, young people who are home-schooled or who have already left formal education are excluded from the target population. In most OECD countries this excluded group is relatively small. However, in some countries it may be larger, meaning that a non-trivial share of 15-year-olds are not represented within PISA estimates.

Data from the PISA 2022 cycle can be used to illustrate this point (OECD, 2024). At the time, Mexico had an estimated 2,193,794 15-year-olds. Of these, 1,592,537 (72.6%) were enrolled in Grade 7 or above and therefore formed the PISA-eligible population. The weighted participating sample represented approximately 1,393,727 pupils (63.5%) of the complete national cohort. Those outside the eligible population included young people who were not enrolled in an educational institution and those still in Grade 6 or below.

In contrast, in countries such as Norway and the Republic of Ireland, only about 1% of 15-year-olds were outside the PISA target population. This difference is entirely consistent with PISA’s “rules” - it is an assessment of pupils enrolled in educational institutions, not all young people of a given age. Nevertheless, it has important implications. It is reasonable to expect that 15-year-olds who are no longer enrolled in school in Mexico are, on average, lower academic achievers than those who remain enrolled. As a result, PISA is likely to overestimate the average achievement of all 15-year-olds in Mexico. It also complicates international comparisons, since Mexico’s results refer to a more selected group of 15-year-olds than in most other OECD countries.

For further case studies of how target population definitions can affect PISA results, interested readers are directed to Jerrim (2017) and Spaul (2019). The former considers the case of Vietnam, a country that has historically performed very well in PISA given its relatively modest level of economic development. Yet it is also a country where a substantial proportion of 15-year-olds have not been enrolled in school. Using PISA 2015 science data, Jerrim (2017) presented a

sensitivity analysis based on the 75th percentile and an assumption that 15-year-olds outside PISA coverage would perform below the national average. Under that scenario, he estimated that Vietnam's comparative position was around 50–60 points lower than suggested by the headline results. Spaul (2019), meanwhile, discusses the case of Turkey. He shows how rising school enrolment rates changed the composition of the PISA target population in this country over time, with lower-achieving pupils entering the assessed population. As a result, PISA may have underestimated the progress Turkey made in increasing the proportion of 15-year-olds reaching a basic level of numeracy and literacy. Together, these examples illustrate how seemingly technical decisions about population coverage can have important consequences for interpreting both country performance and trends over time.

2.3. Sample exclusions 1. What are they and what problems can this cause?

Once the target population has been defined, the next step is to construct a sampling frame. This can be thought of as a comprehensive list of all units that belong to the population from which the sample will be selected. In school-based ILSAs, the first-stage sampling frame normally lists schools. Later-stage frames may list classes or individual pupils, depending on the study design. For example, if the study aims to assess 15-year-old pupils within schools, the frame should include all schools attended by that age group. It is then from this list that the initial sample of clusters (e.g. schools) is drawn.

Yet there are again some important nuances. Countries are usually permitted to exclude a limited proportion of schools and pupils from participation, provided these exclusions meet clearly specified criteria. In TIMSS, for example, school-level exclusions have mainly involved “schools for students with disabilities and very small or remote schools” (LaRoche, & Foy, 2020). Such exclusions are typically made for practical and accessibility reasons; remote schools may be difficult and expensive to reach, while very small schools may be inefficient to include relative to the costs of administering the assessment. Similarly, some pupils with severe disabilities may be unable to access the test materials in a meaningful way. However, exclusions may also reflect features of national education systems. For instance, in TIMSS 2019, Israel excluded ultra-Orthodox schools, as well as those teaching in English or French.

Exclusions can lead to bias when the excluded units differ systematically from the included units in relation to the outcome being estimated. The extent of this bias depends on the interplay between two factors. The first is the proportion of the target population that is excluded; if few schools and pupils are excluded, the impact on population estimates will likely be limited. The second is the degree of selectivity in the exclusions - i.e. how different excluded schools or pupils

are from the broader population. For example, excluding remote schools may introduce little bias into one's estimates if their pupils have similar levels of achievement to pupils elsewhere.

In general, as exclusions become both larger and more selective, the sample becomes less representative of the intended target population. Estimates of population parameters may then be biased. But the implications for cross-national comparisons are somewhat more complex; bias in cross-national differences will depend not only on the level and selectivity of exclusions within each country, but also on how these patterns differ across them. A similar issue arises when studying trends over time, in that bias in trends for a country will depend on how the level and selectivity of exclusions changes over time. Further details on how these issues can be investigated are provided in Section 2.9.

Figure 2.3.1 provides an illustrative example of this problem. The top-left hand panel provides information on a target population that includes a total of 50 units. In each cell is a number between 1 and 10, which one can think of as a measure of that unit's achievement in (for instance) mathematics. In a real study, we would not be able to observe this number for each unit within the target population; rather, it reflects the information we are trying to infer. Within this hypothetical example, the mean score across the 50 units in the target population is 5.12.

Figure 2.3.1. Hypothetical example of exclusions from the target population

Target population (mean = 5.12)					10% random exclusions (mean = 5.11)				
1	3	4	6	8	1	3	4	6	8
1	3	4	6	8	1	3		6	8
1	3	4	6	8		3	4	6	8
2	3	4	6	8	2	3	4	6	8
2	3	4	7	9	2	3	4	7	9
2	3	5	7	9	2	3	5		9
2	4	5	7	9	2		5	7	9
2	4	5	8	10	2	4	5	8	10
3	4	5	8	10	3	4	5	8	
3	4	5	8	10	3	4	5	8	10
10% non-random exclusions (mean=4.67)					20% non-random exclusions (mean=4.33)				
1	3	4	6	8	1	3	4	6	
1	3	4	6	8	1	3	4	6	
1	3	4	6	8	1	3	4		8
2	3	4	6		2	3	4	6	
2	3	4	7		2	3	4	7	
2	3	5	7		2	3	5		
2	4	5	7	9	2	4	5	7	9
2	4	5	8		2	4	5		
3	4	5	8	10	3	4	5	8	10
3	4	5	8		3	4	5	8	

Notes: Figures refer to mathematics knowledge of the unit of the target population measured using a 1-10 scale.

Now say that 10% of units in this hypothetical population are excluded, but these exclusions occur at random (or at least occur in a way that is unrelated to the statistic – e.g. average maths achievement - we are trying to estimate). This situation is reflected in the top-right panel of Figure 2.3.1, where the cells shaded in red have been excluded. However, as these exclusions have occurred randomly, our estimate of the mean score within the target population remains virtually unchanged at 5.11.

The third panel of Figure 2.3.1 in the bottom left also has a 10% exclusion rate from the population, though they are no longer random. Rather, higher-achieving units are disproportionately likely to be excluded. Clearly this now has an impact on our estimate of the mean; it falls to a value of 4.67. This is because 10% of high-achieving units have been systematically excluded from the target population. The bottom right panel then illustrates how the discrepancy becomes larger when the exclusion rate increases from 10% to 20% (while exclusions remain disproportionately concentrated among higher-achieving units). Specifically, the average we would obtain now drops to 4.33. Together, this illustrates how it is the combination of (a) the percentage of exclusions along with (b) their selectivity, which determines the bias this introduces into one's estimates.

ILSAs address the problem of exclusions primarily by limiting their overall scale. In practice, there is often little that can be done about the selectivity of excluded units. For example, pupils attending special schools will likely have (on average) much lower levels of academic achievement than pupils in mainstream schools. Because the selective nature of this group cannot be eliminated, ILSAs instead seek to restrict the proportion of the population that countries are allowed to exclude. For instance, TIMSS formally limits total exclusions to no more than 5% of the national target student population (although Atasever et al., 2024, note 5.5% is the effective threshold arising from rounding). The rationale is that, even if excluded pupils or schools are systematically different from those included, their impact on the representativeness of the data will be limited if they constitute only a small share of the target population. Put differently, the potential for bias increases with both the size and selectivity of exclusions, so controlling the former helps constrain the latter's consequences. It has been suggested that restricting exclusions to around 5% limits bias in mean ILSA test scores to approximately 5 test points, or about 0.05 standard deviations (Jerrim, 2021).

In practice, however, these exclusion criteria are not always met. There have been several instances in which countries have exceeded the thresholds set by the international organisations responsible for conducting ILSAs. Atasever et al. (2024), for example, summarise evidence from

TIMSS and PIRLS and highlight Israel as a notable case. Across ten waves of data collection, Israel had an average total exclusion rate of around 22%, largely driven by school-level exclusions. The United States and Denmark also had average exclusion rates closer to 6% when examined across multiple ILSA cycles. Jerrim and Savaşçı-Smith (2025) further show that exclusion rates increased in several countries during the COVID-19 pandemic, with some exceeding 10% in PIRLS 2021. These levels of exclusion may have a non-trivial impact on estimates of population parameters, including average achievement.

2.4. Stratification. What is it, why is it used and how is it implemented in ILSAs?

Once any permitted school-level exclusions have been removed from the sampling frame, the next step is to draw the sample itself. In ILSAs, this is not done through simple random sampling of individual pupils. Rather, they typically use some form of complex survey design, reflecting the practical realities of collecting data from schools, classrooms, teachers and pupils across many countries. These designs are introduced over the following sections. We begin with one of the most common features of ILSA sampling designs - stratification.

To understand stratification, why it is used in ILSAs and the benefits it can offer, it is useful to first step back and consider simple random sampling. We will do this through an illustrative example. Figure 2.4.1 presents an image of the Union Jack flag. The image is made up of 460,800 pixels, which we can think of as the target population of interest (i.e., a population consisting of 460,800 individual units).

Figure 2.4.1. The Union Jack flag formed of 460,800 pixels.



Suppose we then draw two separate random samples of 46,000 pixels from this population. Each pixel has the same probability of being selected. The two resulting samples are shown in Figure 2.4.2.

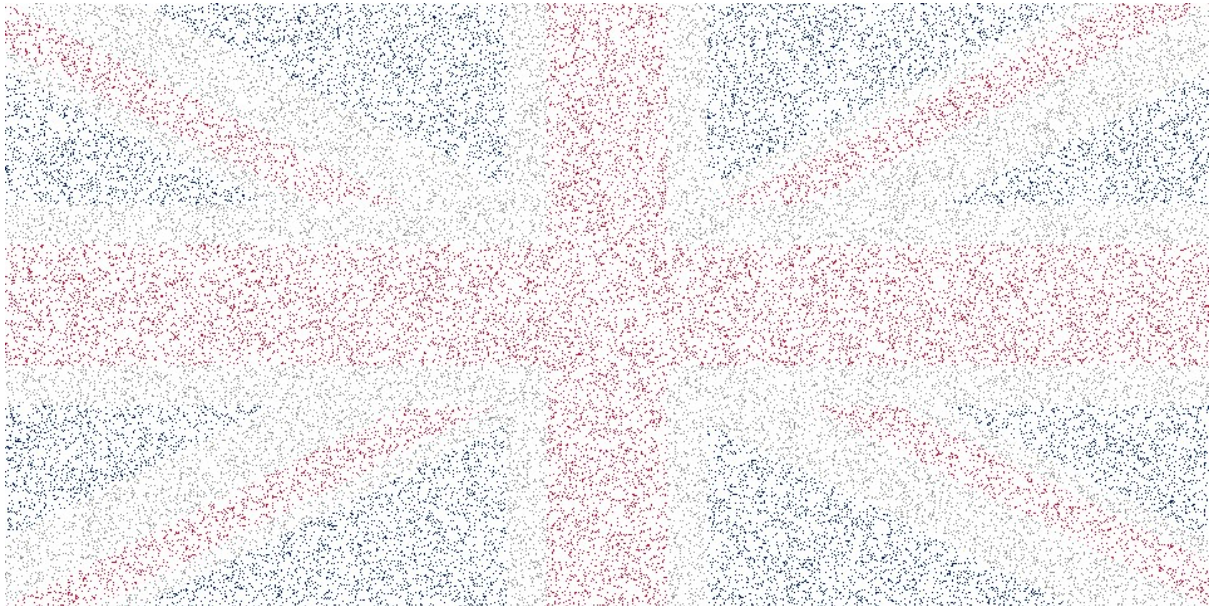
Panel (a) illustrates what happens when simple random sampling works as intended. The selected pixels are spread reasonably evenly across the full image, meaning that the sample contains information from all major parts of the population. Although the resulting picture is somewhat blurry, this is unsurprising - we are observing only 46,000 pixels rather than the full population of 460,800. Nevertheless, the sample is large enough (and sufficiently representative) for the main features of the Union Jack to be clearly visible. We can identify the broad pattern of the flag, including the central cross, the diagonal lines and the contrast between the different coloured areas. This is the basic intuition behind sampling in survey research. We do not need to observe every unit in the population to learn something useful about it. Provided the sample has been selected at random (and it is large enough) it can give a reasonably accurate picture of the population as a whole.

Panel (b), in contrast, shows a deliberately constructed extreme outcome that is nevertheless mathematically possible under simple random sampling. As the sample is selected completely at random, there is always a possibility that any single draw from the population will be “unlucky”. In this example, the sample is not especially representative of the full image. By chance, many of the selected pixels come from the top-right quadrant of the flag, while relatively few come from the left-hand side. As a result, the sample provides an incomplete and uneven picture of the population. We can still make an informed guess that the image is the Union Jack. However, our confidence is lower than in Panel (a). The key lesson is that random sampling generally works well in expectation (i.e. on average across many samples drawn). But we typically only draw and collect data from a single sample - which may not be representative of the population purely due to chance variation.

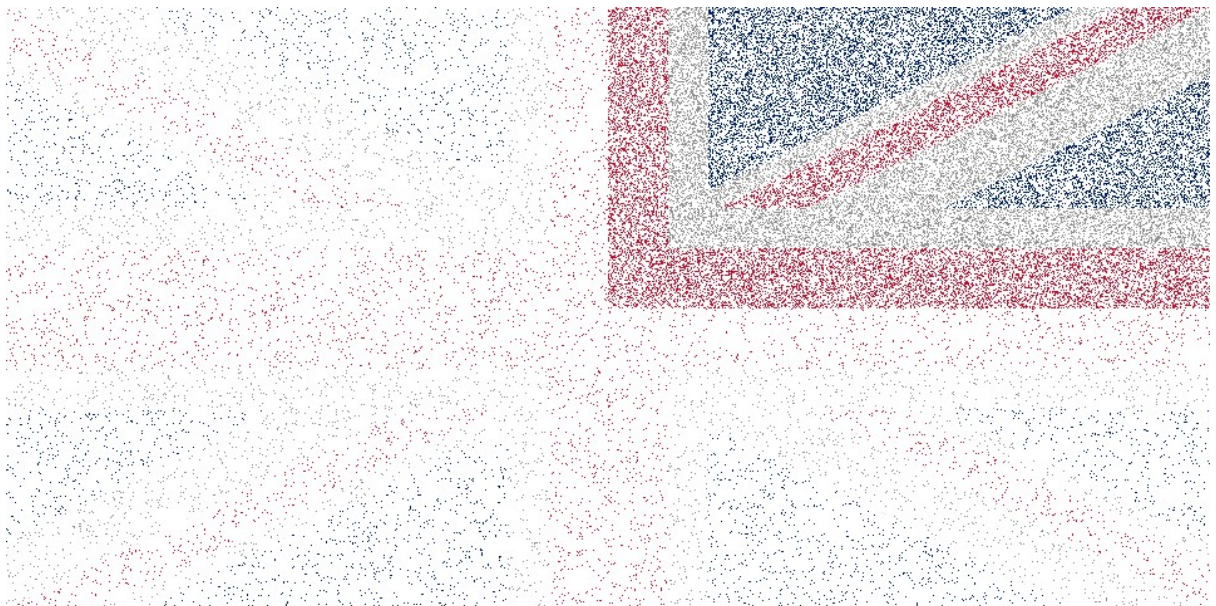
This illustrates a central limitation of simple random sampling; there can be considerable sampling variation from one possible sample to another (again noting how Figure 2.4.2 panel b is a very extreme example that we have contrived). In real studies we usually draw only one sample and there is always a risk that we are “unlucky” in the particular sample selected. We thus may end up with participants who do not accurately reflect the population as a whole.

Figure 2.4.2 Two random samples of 46,000 pixels from the Union Jack flag

(a) A typical simple random sample



(b) A (very) “unlucky” random sample draw

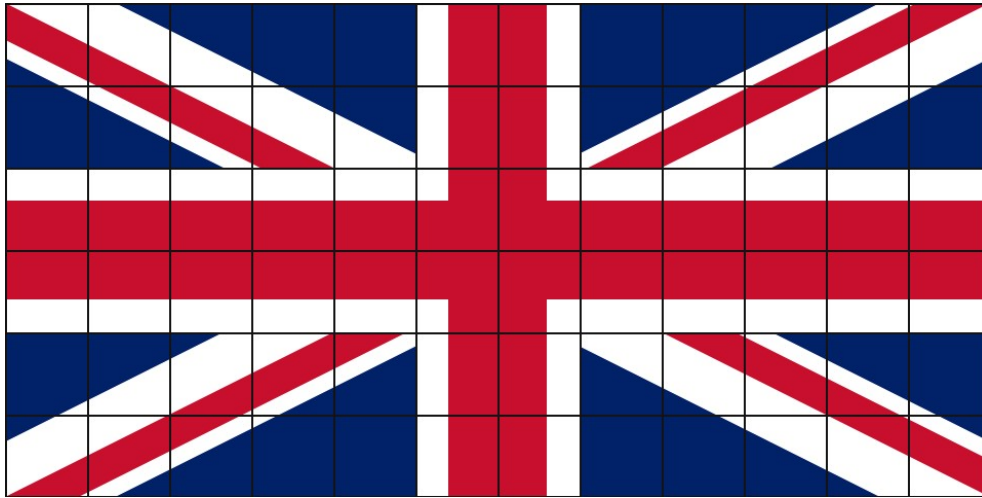


Notes: Figure adapted using AI.

This brings us to stratification. Returning to our illustrative example, suppose that the Union Jack image remains our target population of 460,800 pixels. We divide the image into 72 equal grid squares, arranged in 12 columns and six rows. Each square contains $80 \times 80 = 6,400$ pixels and is treated as a separate group, or stratum. Rather than selecting 46,000 pixels completely at

random from the full population, we select pixels randomly within each of the 72 strata. (approximately 639 pixels per square). Figure 2.4.3 shows the resulting stratification grid.

Figure 2.4.3. The 460,800-pixel Union Jack divided into 72 equal strata of 6,400 pixels.



Notes: Figure adapted using AI.

The implication is that it is no longer possible to end up with a sample drawn almost entirely from one part of the flag, such as the top-right corner. Stratification ensures that every area of the image contributes to the final sample. We have therefore to some extent protected ourselves against an “unlucky” random draw. As a result, we can have greater confidence that the single sample we have drawn covers different parts of the population (as defined by the stratification variables).

In the survey methods literature, stratification is usually seen as offering two main advantages. First, it helps ensure that the sample includes a sufficient number of units from important parts of the population (what counts as “important” is determined by the variables used to construct the strata – see further details below).

Second, stratification can improve the precision of estimates. This is particularly true when units within each stratum are relatively similar to one another. For instance, educational achievement may vary substantially across all secondary schools within a country, but it may vary less within school types or geographic regions. Thus, while simple random sampling will include variation that occurs both within and between strata (e.g. school types), stratified sampling accounts for differences between strata by design. It thereby reduces the sampling error to mainly the variation that occurs within each stratum.

In ILSAs, school-level stratification commonly takes two forms (explicit and implicit stratification). First, schools on the sampling frame are divided into separate groups, known as explicit strata. The variables used to form these groups differ somewhat across countries, but they commonly include characteristics such as geographic region and school type. For example, in the Australian PISA 2022 sample, the explicit stratification variables include the eight states and territories, together with three school types (in addition to one “certainty” stratum). Returning to our prior analogy of the Union Jack flag, these variables can be thought of as forming the “grid” we break the population into. One can thus think of Australia as being divided into a set of state-by-school-type cells, with schools grouped according to the cell to which they belong.

Second, within each explicit stratum, schools are usually ordered according to one or more additional variables. This is known as implicit stratification. For instance, within a particular geographic region, schools might be ordered from lowest to highest according to the percentage of pupils from disadvantaged socioeconomic backgrounds, using administrative data on schools held by the national government. The intuition is that, when schools are then selected systematically from the ordered list, the sample will also tend to be balanced with respect to these additional characteristics. In this way, implicit stratification helps improve balance across the implicit stratification variables (beyond the variables used to define the explicit strata). As with explicit stratification, the variables used for implicit stratification vary across countries, depending on the information available.

This raises an important question - which variables should be used to stratify the sample? Ideally, stratification variables should be strongly correlated with the outcomes of interest. In ILSAs, this means choosing variables that are likely to be related to children’s achievement. In studies such as TALIS (an international survey of teachers) the relevant stratification variables may instead be characteristics associated with teacher outcomes, such as workload, job satisfaction or wellbeing.

The reason follows from the discussion above. Stratification is intended to account for as much between-stratum variation in the outcome of interest as possible. This is achieved by creating strata that are likely to differ from one another on that outcome. For example, if prior evidence suggests that school type, region or the socioeconomic composition of schools is strongly associated with mathematics achievement, then these are promising candidates to stratify the sample by. The key point is that stratification is most useful when the strata differ substantially

from each other in terms of the outcome of interest, while units within each stratum are relatively similar.

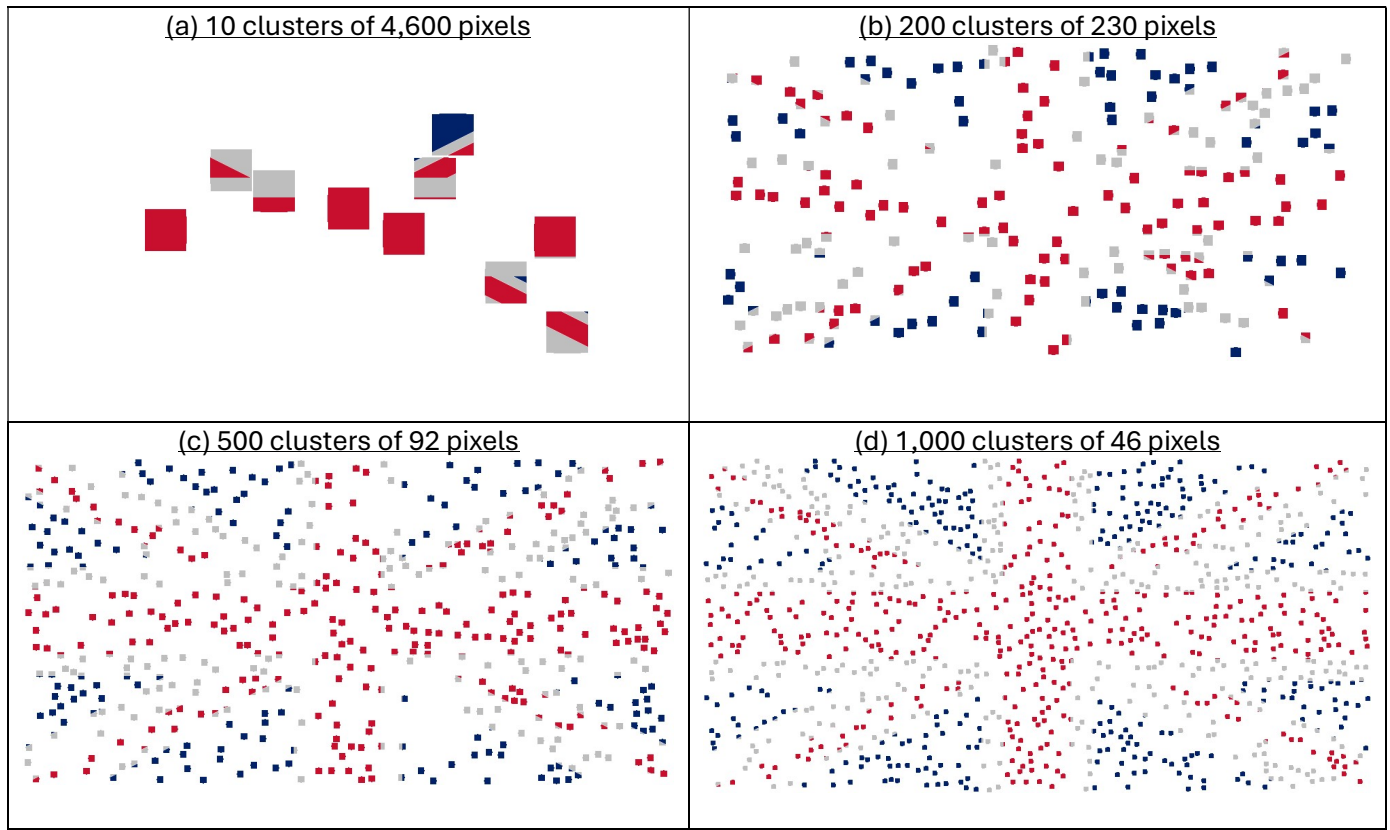
2.5. Clustering. What is it, why is it used and how is it implemented in ILSAs?

Once the sampling frame has been divided into strata, the next step is usually to sample schools. In most ILSAs, researchers do not sample individual pupils or teachers directly from the entire national population. Doing so would be impractical because selected individuals would be spread thinly across a very large number of schools. Instead, ILSAs typically use multistage cluster sampling. In particular, schools are selected first, followed by pupils, teachers or intact classes within those schools. The main advantage is practicality and lower fieldwork cost. For example, it is easier and less costly to recruit 200 schools and assess 20 pupils in each than to test two pupils in each of 2,000.

Yet cluster sampling also has an important statistical cost. Pupils attending the same school tend to be more similar to one another than pupils attending different schools, including in educational achievement. When outcomes are positively correlated within schools, a sample of 20 pupils from each of 200 schools generally contains less independent information than a simple random sample of 4,000 pupils. Clustering then increases standard errors and widens confidence intervals relative to simple random sampling of the same number of pupils. The size of this loss mainly depends on the similarity of outcomes within schools and the number of pupils sampled per school.

We can return to the Union Jack example to illustrate this point. In Figure 2.4.2, we drew a simple random sample of 46,000 individual pixels from a population of 460,800. Figure 2.4.4 instead samples spatial clusters of pixels. Panel (a) uses 10 clusters of 4,600 pixels. Panel (b) uses 200 clusters of 230 pixels. Panels (c) and (d) use 500 clusters of 92 pixels and 1,000 clusters of 46 pixels, respectively. Each panel thus consists of exactly 46,000 sampled pixels – which is also the same as in our simple random sample previously. Selected pixels retain the flag's red or blue colour, while selected white pixels are shown in light grey so that every cluster remains visible.

Figure 2.4.4. Cluster samples of exactly 46,000 pixels from the 460,800-pixel Union Jack flag



Notes: Figure adapted using AI.

There are two important points to note. First, compare Panel (a) of Figure 2.4.4 with Figure 2.4.2 presented previously. Both contain exactly 46,000 sampled pixels. With only 10 large clusters, however, the sample covers relatively few parts of the image. This makes the Union Jack difficult to identify, highlighting the loss of statistical information from sampling clusters of pixels (with the extent of lost precision depending on the degree of similarity within clusters).

Second, Figure 2.4.4 becomes clearer as the number of clusters increases, even though every panel contains the same 46,000 sampled pixels. When nearby pixels are positively correlated, spreading a fixed sample over more clusters generally captures more independent information and reduces sampling variance. This is why, for a fixed number of sampled pupils, a design covering many schools will usually be more precise than one covering only a few schools.

The key analytical implication is that clustering usually increases sampling uncertainty relative to a simple random sample of the same size when outcomes are positively correlated within schools. A sample of 4,000 pupils drawn from a limited number of schools will therefore generally provide less statistical information than a simple random sample of 4,000 pupils drawn

independently from across the country. Standard errors, confidence intervals and tests of statistical significance must therefore reflect this sampling design. Chapter 3 discusses various survey-design methods used for this purpose.

In practice, cluster sampling in ILSAs is commonly combined with stratification, although the exact design varies. When explicit strata are used, the school sampling frame is divided as discussed in the previous subsection, and schools are then often selected systematically within each stratum with probability proportional to size (PPS). Under PPS, larger schools have a higher probability of selection. For instance, in TALIS 2018, countries were allowed to develop their own national stratification strategies rather than imposing one universal set of strata.

This general approach is used across major ILSAs. For instance, PISA 2022 used a two-stage stratified design; schools were first selected systematically with probabilities proportional to size (number of enrolled 15-year-olds), followed by the selection of pupils within sampled schools. In TALIS 2018, countries typically sampled 200 schools and then 20 teachers within each school. TIMSS likewise uses a stratified two-stage cluster design, selecting schools first and one or more intact classes within schools (OECD, 2019, 2025; von Davier, Fishbein, & Kennedy, 2024).

2.6. Selection of pupils. Oversampling and pupil exclusions.

Once schools have been selected in the first stage of the sample design, the second stage involves selecting participants within those schools. The approach differs somewhat across ILSAs. In PISA and TALIS, individual pupils or teachers are typically sampled at random from within each selected school. In TIMSS and PIRLS, the design usually involves selecting one or two intact classes, with all eligible pupils in those classes invited to participate. The exact number of participants selected within each school varies across studies, countries, grades and survey cycles. Nevertheless, the resulting within-school sample commonly includes somewhere between 20 and 45 participants (although class-based samples may be larger).

There are however two additional points to note.

First, a small number of within school pupil-level exclusions are usually permitted, though only for specific reasons. These pupil-level exclusions are typically where functional or intellectual disability prevents participation or where the individual has insufficient proficiency in the assessment language (typically having received less than one year of instruction in that language). As with school-level exclusions (discussed in section 2.3 above), the study organisers

set limits on the maximum percentage of pupils that countries are allowed to exclude, as well as the criteria under which such exclusions can be made. However, as noted in section 2.3, these limits are sometimes exceeded. A further complication is that the definition and classification of disability vary across countries. For this reason, the TIMSS technical report notes that countries are “asked to translate the TIMSS international exclusion standards into the local equivalent because disability criteria vary from country to country” (Siegel & Foy, 2024). This may introduce some cross-national inconsistency in which pupils are excluded from the achieved sample.

Second, some countries choose to oversample pupils with particular characteristics. For example, in Australia, Aboriginal and Torres Strait Islander pupils are a group of particular interest from an educational equity perspective. In PISA, pupils identified as belonging to this group have therefore been included in the sample with certainty within participating schools, rather than being selected at random in the same way as other pupils. This approach is used to ensure that the achieved sample includes enough Aboriginal and Torres Strait Islander pupils to support meaningful subgroup analysis. However, oversampling also means that these pupils have a higher probability of selection. This must be taken into account when producing population estimates. The survey weights supplied with ILSA datasets adjust for these unequal selection probabilities, ensuring that oversampled groups do not contribute disproportionately to national estimates. These weights are discussed in detail in the subsequent chapter (see section 3.2).

It should also be noted that oversampling is sometimes used in international studies for a different purpose - to enable reporting at the subnational level. For example, Canada oversamples schools within provinces so that ILSA results can be reported separately for each individual province. Similarly, in PISA, the United Kingdom oversamples schools in Wales, Northern Ireland and Scotland so that results can be produced for each of its four constituent countries. This type of oversampling is entirely legitimate and does not in itself undermine representativeness (though it can impact the precision). Rather, it is an intentional feature of the sample design. Yet the fact that oversampling has been used must be accounted for by analysts when using data from ILSAs via the application of the survey weights. We again direct the reader to the following chapter (section 3.2 in particular) for further details.

2.7. Non-response.

Non-response is a potential problem in any survey. ILSAs are no exception. Some sampled schools may decline to participate and some sampled pupils or teachers within participating schools may be absent or otherwise fail to take part. This creates two main challenges.

The first, and most important, is that non-response may make the achieved sample less representative of the target population. This can introduce bias into estimates of population parameters, such as average achievement in a country or the proportion of pupils reaching a particular benchmark. As with the discussion of exclusions in section 2.3, the extent of this bias depends on two factors - the level of non-response and the selectivity of non-respondents. If non-response is low, its impact is likely to be limited. Similarly, if non-response occurs essentially at random (or at least independent of the outcome being measure), then it should not systematically impact point estimates (i.e. the achieved sample may be smaller but it remains unbiased in expectation). This is similar to the example of exclusions set out in Figure 2.3.1; if non-response or exclusions occur randomly, this should have minimal impact on our point estimates. However, if non-response is concentrated among particular types of schools or pupils, especially those who differ in achievement or other outcomes of interest, bias into estimates is likely to be introduced.

The second issue is loss of statistical power. Even when non-response occurs randomly, it reduces the amount of information available for analysis (e.g. one can think of it as simply reducing the number of pixels visible in Figure 2.4.2). This means standard errors may be larger and confidence intervals wider than they would have been with full participation. Thus, random non-response may leave point estimates unbiased but still reduce the precision with which population parameters can be estimated.

In ILSAs, non-response can occur at two levels; (a) the school-level and (b) the pupil level. We take these in-turn.

Starting with schools, those that were sampled are approached by the national organisation responsible for delivering the ILSA and invited to participate. The process varies across countries. Some schools may receive incentives, such as payments to cover staff time or administrative costs, while in other countries participation may be required by law or strongly encouraged by

education authorities. Nevertheless, some sampled schools may decline to take part, for reasons such as workload pressures, disruption to teaching, assessment fatigue or competing priorities.

ILSAs attempt to pre-empt school non-response by identifying up to two replacement schools for each originally sampled school. If a sampled school refuses to participate, these replacement schools can be approached to take its place. The replacements are typically the schools located immediately before and after the originally selected school on the sampling frame. This is important because the frame has already been explicitly and implicitly stratified (see section 2.4 above). Replacement schools should therefore be similar to the originally sampled school in terms of the observable characteristics used in the stratification process (e.g. region, school type and socioeconomic composition).

This use of replacement schools can be thought of as a form of near-neighbour school substitution (i.e. a form of donor imputation). Data from a non-responding school are effectively replaced with data from another, similar school. The intuition is that this helps reduce the bias that might arise from selective school non-response, at least with respect to the observed characteristics used to stratify the school sample. It also helps maintain the intended sample size, thereby limiting losses in statistical power. However, replacement schools do not solve all problems associated with non-response. If schools decline to participate for reasons linked to unobserved characteristics (which are also associated with our outcomes of interest), replacement schools may not be equivalent. In such cases, some risk of non-response bias remains.

It should also be noted that, typically, only two replacement schools are included per each originally sampled school. If both replacement schools also refuse to participate, no further schools are then contacted to take their place.

This approach means that most ILSAs report two school response rates – the before replacement rate and the after-replacement rate. The before-replacement rate refers to the proportion of originally sampled schools that agreed to participate. The after-replacement rate also counts participating replacement schools, where these were used in place of initially sampled schools that declined. By definition, the before-replacement response rate will always be equal to or lower than the after-replacement rate (the replacement schools move the numerator in the

response rate calculation upwards, while the denominator remains unchanged)⁸. Conceptually, it can be thought of as the “pre-imputation” response rate, since it reflects participation among the schools originally selected by the sample design. For this reason, if only one school response rate is reported, the before-replacement rate is often the more informative measure.

As noted above, the extent of bias introduced by non-response depends partly on the proportion of sampled units that do not participate. This raises an important question - what level of school non-response is generally considered acceptable in ILSAs? The answer is not entirely straightforward because of the use of replacement schools as discussed above. But, by way of example, the criteria used in PISA are summarised in Figure 2.7.1. (Other ILSAs typically use a similar broad approach, although the specific response rate thresholds required vary). In this figure, the school response rate before replacement is shown on the horizontal axis, while the school response rate after replacement is shown on the vertical axis. The figure is then divided into three broad zones, each corresponding to a different judgment about the acceptability of the achieved school response rate.

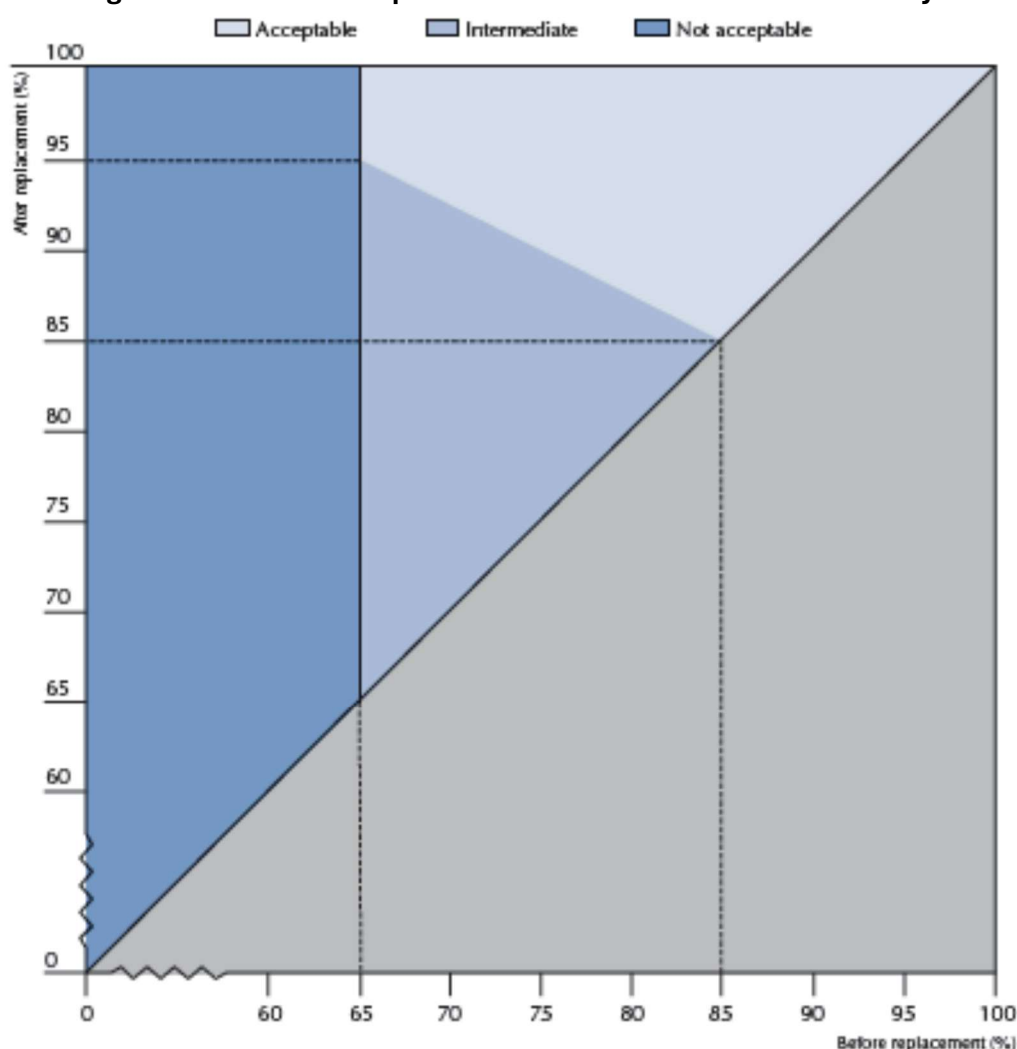
The first zone is where the school response rate is considered unacceptable. A country falls into this category if its before-replacement school response rate is below 65%. This represents the minimum initial response threshold that countries are expected to meet. If the response rate among originally sampled schools falls below this level, then the risk of bias from school non-response is (in theory) considered too great to support reliable estimates of population parameters. It should be noted that several countries often fail to reach the “acceptable” criteria for school response rates. For instance, Anders et al. (2021) noted that seven countries (including five members of the OECD) in the PISA 2015 study failed to reach this threshold.

The second zone is labelled acceptable. This is where the school response rate is judged to be sufficiently high that the risk of bias from school non-response is relatively low. Countries automatically fall into this zone if they achieve a before-replacement school response rate of at least 85%. Countries with a before-replacement response rate between 65% and 85% may also be classified as acceptable, provided their after-replacement response rate is high enough. Importantly, this operates on a sliding scale. A country with a before-replacement school

⁸ Say there were 100 initially sampled schools, 80 of which participated, leading to a before replacement response rate of $80/100 = 80\%$. Now say of the 20 non-participating originally sampled schools, 10 replacement schools additionally agreed to participate. These 10 schools would be included in the numerator, but not the denominator. Thus, the after-replacement school response rate calculation is $(80+10) / 100 = 90\%$. (Note this is an unweighted example; in ILSA weighted values are typically used).

response rate of 65% would need an after-replacement response rate of 95% for its school sample to be deemed acceptable. On the other hand, if the before-replacement response rate is higher (e.g. 80%), then a lower after-replacement response rate (around 88%) is deemed sufficient.

Figure 2.7.1. School response rate criteria used in the PISA study



Source: Adapted from OECD (2024, Figure 6.1).

The third and final zone is labelled “intermediate”. This applies when a country’s before-replacement school response rate lies between 65% and 85%, but its after-replacement response rate does not meet the variable threshold required for the sample to be classified as acceptable. In these cases, the school response rate is not low enough to be automatically deemed unacceptable, but neither is it high enough to be treated as unproblematic. Countries in this zone are therefore required to conduct a non-response bias analysis. The purpose of this analysis is to assess whether non-participating schools differ systematically from participating schools on observable characteristics (and hence whether school non-response may have

introduced bias into estimates of population parameters). These bias analyses are discussed in further detail in the following section.

Turning to within-school non-response, the situation is somewhat simpler. Rather than using a before- and after-replacement framework, ILSAs typically set a minimum participation rate that countries are expected to achieve among the sampled units within participating schools. In PISA, for example, at least 80% of sampled pupils are expected to participate. In TALIS, the corresponding figure is 75% of sampled teachers. In TIMSS, countries are expected to achieve participation from at least 95% of sampled classes and 85% of sampled pupils. The rationale for why the thresholds vary across studies is not particularly clear.

It should also be noted that the survey weights included in ILSA datasets, discussed in more detail in the following chapter (see section 3.2), usually contain a component that adjusts for non-response. At the school level, non-response adjustments in the weights are typically based on the variables used for stratification. These adjustments are useful, but they are not magic. Their effectiveness depends on whether the adjustment variables are related to both the probability of school non-response and the outcome of interest. For example, if schools in particular regions or school sectors are less likely to participate (and these same characteristics are also associated with mathematics achievement) then applying weights when using these variables may help reduce non-response bias. However, if non-responding schools differ from participating schools in ways not captured in the stratification variables (e.g. leadership quality, school climate or attitudes toward testing) the adjustment may do little to remove bias. This also means that the effectiveness of school-level non-response adjustment is likely to vary across countries, given the stratification variables used differ across national contexts. Consequently, some countries may be better equipped than others to adjust for selective school non-response.

At the pupil or teacher level, the non-response adjustments included within the weights are often much more limited. In PISA, for example, within-school non-response adjustments are largely based on characteristics such as grade and gender (Anders et al., 2021). While useful, these variables are unlikely to capture much of the selectivity in individual pupil or teacher non-response in many countries. Readers are directed to the following chapter (section 3.2 in particular) for further details on survey weights.

2.8. Non-response bias analysis

When a country fails to achieve the required response rates discussed in the previous section, it is typically required to conduct a non-response bias analysis (NRBA) of its sample. The purpose of an NRBA is to examine whether non-participating schools, pupils or teachers differ systematically from those who did participate, and hence whether non-response may have introduced bias into estimates of population parameters. In practice, however, details of these analyses have often been remarkably limited. They have not routinely been made public by the international organisations that lead ILSAs, such as the OECD. Instead, publication has generally been left to individual countries, and most have not released detailed information on the methods used, the data analysed or the conclusions reached.

However, the nature of NRBAs has been discussed in the academic literature, most notably by Anders et al. (2021) and Jerrim (2021). These studies suggest that NRBAs have most often been conducted at the school level. Typically, they compare the characteristics of schools that participated in the study (the final achieved school sample) with the characteristics of all schools originally selected for participation. These comparisons are usually based on observable characteristics available on the sampling frame, often including the variables used to stratify the sample. The analysis is generally conducted by individual countries and then submitted to the international organisation responsible for the study. A panel of experts appointed by the international organisation then reviews the evidence and judges whether the achieved sample is sufficiently representative, or whether school non-response is likely to have introduced bias into estimates of population parameters. The difficulty, of course, is that such analyses are limited by the information available on non-participating schools.

Several examples of non-response bias analyses are presented in the supplementary material to Anders et al. (2021). However, Table 2.8.1 provides a further example. It is based on an abbreviated version of the analysis conducted for the US component of TALIS 2018 (Kastberg et al., 2021) where, of the 214 originally sampled and eligible schools, 125 participated (before replacement schools were added). The analysis compares all schools in the original eligible sample with those that ultimately agreed to participate in terms of school control, location, census region, poverty level, grade structure and school size. A separate table (not reproduced here) also considered school enrolment, the percentage of pupils from different ethnic groups and the percentage eligible for reduced-price lunch.

Table 2.8.1. Comparison of eligible and participating schools in the original, before-substitution U.S. TALIS 2018 school sample

	Eligible	Participating	Difference	P-value
School control				
Public	89.3	87.3	-1.9	0.437
Private	10.7	12.7	1.9	
Locale				
Central city	31.3	29.8	-1.5	0.218
Suburb	37.8	34.3	-3.6	
Town	9.7	8.9	-0.8	
Rural	21.1	27.0	5.9	
Census region				
Northeast	19.9	14.1	-5.9	0.088
Midwest	20.9	23.5	2.6	
South	41.6	46.8	5.2	
West	17.5	15.6	-1.9	
Poverty level				
High	37.2	36.9	-0.3	0.936
Low	62.8	63.1	0.3	
Grade structure				
Grades 6–8	48.1	48.0	-0.1	0.976
Grades 9–12	25.6	25.1	-0.5	
Other	26.3	26.9	0.6	
School size				
Small	38.6	43.5	4.8	0.077
Medium	32.6	33.9	1.2	
Large	28.7	22.7	-6.1	

Source: Kastberg et al. (2021, Table D-2). Sample size is 214 eligible schools and 125 participating schools. Note: Bias is the participating-school estimate minus the eligible-school estimate. Biases were calculated from unrounded weighted estimates and may therefore differ slightly from calculations using the displayed percentages. Detail may not sum to 100% because of rounding.

As Table 2.8.1 illustrates, this particular NRBA reported some modest differences between eligible and participating schools. For instance, the participating sample contained a somewhat higher proportion of small schools and schools in the South of the country. However, none of the observed differences was statistically significant. Yet, as Anders et al. (2021) and Jerrim (2021) note, there are several limitations to this approach.

First, NRBA's are typically conducted at the school level, where sample sizes are often quite small (usually around 200 schools). Yet discussion of the results is frequently framed around whether differences between the originally sampled schools and the participating schools are “statistically significant”. This is problematic because, with such limited sample sizes, these tests are likely to be underpowered. In other words, the absence of statistically significant differences should not be interpreted as strong evidence that non-response bias is absent.

Second, NRBAAs often consider only a small number of observable school characteristics. Ideally, participating and non-participating schools would be compared using variables that are strongly related to the outcomes of interest. In PISA, TIMSS, and PIRLS, for example, one would ideally compare schools using independent measures of prior academic achievement, such as national assessment results. Or, in the example of TALIS in Table 2.8.1 above, one would ideally like to examine differences between the eligible and achieved samples in terms of variables that are likely to correlate with outcomes such as teachers' working hours, job satisfaction and participation in professional development. However, such information is rarely available (or at least rarely used) in published NRBAAs. Instead, analyses often rely on more basic sampling-frame variables, which may only weakly capture differences in the outcomes of interest (as per Table 2.8.1 above).

Third, individual-level non-response is potentially more important than school-level non-response, as suggested by previous analyses of PISA data (Micklewright et al., 2012; Durrant & Schnepf, 2018). Yet relatively few pupil-level NRBAAs have been conducted (particularly in earlier ILSA waves), despite the fact they would often have much greater statistical power because of the larger number of sampled pupils.

Finally, as noted above, very little detail is usually published about the content of NRBAAs or how final judgments are reached. This lack of transparency makes it difficult for secondary analysts to independently assess the likely extent of non-response bias.

The consequence is that, in many respects, NRBAAs conducted as part of ILSAs have limited impact. While they are in principle designed to provide evidence as to whether non-response has compromised the representativeness of a country's sample, many countries required to conduct an NRBA still have their data and results published, even when the technical standards set by the international organisation have not been met. For example, Anders et al. (2021) note that, out of 24 NRBAAs conducted as part of PISA, only three led to the conclusion that a country's data were not sufficiently representative and should be excluded from the main results. Indeed, there have been instances in which countries failed to meet even the minimum 65% before-replacement school response threshold yet still had their results published. This raises questions about the extent to which NRBAAs function as a meaningful quality-control mechanism, rather than as a procedural requirement that allows countries with low response rates to remain in the published international database.

Moreover, as noted in previous sections, non-response is only one of several reasons why an ILSA sample may fail to represent the broader population it is intended to cover. Its effects may be compounded by, for instance, high levels of school and/or pupil exclusions. What matters for bias in the final sample is therefore not any one of these issues in isolation, but their combined effect. The following section thus draws on case studies that have examined the representativeness of the final ILSA samples achieved.

2.9. Investigating the likely bias in the final ILSA samples. A case study of PISA data from Canada and the UK.

A small number of academic studies have investigated the representativeness of achieved ILSA samples, considering the cumulative effect of the issues discussed in previous sections. Anders et al. (2021), for example, examine potential problems with the PISA 2015 data for Canada. They note that Canada either failed to meet, or only narrowly met, several of the technical standards set by the OECD for exclusions and response rates. Taken together, they argue that these issues meant that the Canadian PISA sample covered only around half of the 15-year-old population, compared with coverage of over 90% in countries such as Finland, Japan and South Korea. Under a series of selected middle-range assumptions, Anders et al. (2021) then indicate that Canada's estimated mean reading score is likely to be inflated by around 10 to 20 PISA test points (equivalent to approximately 0.10 to 0.20 standard deviations). Jerrim (2021) conducts a similar analysis using PISA 2018 data for England and Wales, concluding that sample exclusions and non-response likely produced an upward bias in average scores of around 10 to 15 test points (0.10 to 0.15 standard deviations).

To illustrate how these conclusions are reached, we provide a simple walk-through example below, based on PISA data for Wales. The same general approach can be applied more broadly to investigate likely bias in sample surveys. The starting point is to assume that auxiliary information is available for both the achieved ILSA sample and the wider target population. This may be possible through links between ILSA data and administrative records for example (see chapter 7, section 7.3, for further discussion of such linkages). Ideally, this auxiliary information should be strongly correlated with the quantity we are trying to estimate. For instance, if one's aim is to estimate average PISA mathematics achievement, then the auxiliary measure should be closely related to PISA mathematics scores. A good candidate would be pupils' performance on an independent mathematics assessment, preferably taken within a year or so of the ILSA. We

also need to know the distribution of this auxiliary measure in the relevant population as a whole. By comparing the distribution of the auxiliary variable in the achieved PISA sample with its distribution in the full population, we can begin to assess how representative the sample is and how much bias exclusions and non-response may have introduced.

An example of this approach is shown in Table 2.9.1 using PISA 2018 data from Wales. For the moment, focus on columns (1) to (4). In Wales, pupils aged 15/16 in school Year 11 sit national examinations in mathematics, known as GCSEs, within a few months of the PISA assessment. These examinations result in grades ranging from A* (highest grade) to U (lowest grade) as shown in column (1). These GCSE grades have been linked to individual pupils in the PISA sample for Wales. This allows the GCSE grade distribution of the achieved PISA sample to be calculated (see column 2). Moreover, because GCSE mathematics is a national examination, the corresponding grade distribution is also known for Wales as a whole, as shown in column (3). The difference between the GCSE grade distribution in the PISA sample and the national population can therefore be calculated, as shown in column (4).

Table 2.9.1. National examination grade distribution in Wales. PISA sample compared to population values

1	2	3	4	5
GCSE grade	PISA sample	National distribution	Difference (percentage points)	Average PISA score by grade
A*	11.1%	8.7%	2.4%	577
A	12.0%	9.2%	2.8%	541
B	19.4%	17.1%	2.3%	504
C	26.2%	22.9%	3.3%	474
D	13.5%	13.9%	-0.4%	437
E	7.9%	10.4%	-2.5%	413
F	3.7%	5.9%	-2.2%	387
G	3.3%	5.9%	-2.6%	382
U	2.9%	6.1%	-3.2%	359

What does this analysis show? The achieved PISA sample for Wales overrepresents pupils with high levels of mathematics achievement and underrepresents those with lower levels of achievement, as measured by GCSE mathematics grades. For example, 23% of pupils in the PISA sample achieved one of the two highest GCSE grades (A* or A) compared with 18% in the national grade distribution. Conversely, only 6% of pupils in the Welsh PISA sample achieved one of the two lowest grades, compared with 12% of pupils in the population distribution. This suggests that the final PISA sample was academically stronger than the wider population of Year 11 pupils in

Wales. Consequently, estimates of average mathematics achievement based on the PISA sample are likely to be biased upward.

How can this be translated into the likely magnitude of bias in average ILSA test points? This again relies on linking the ILSA sample to auxiliary information, such as national examination grades. In the Welsh example, the linked data made it possible to estimate the average PISA mathematics score achieved by pupils in each GCSE mathematics grade category, as shown in column (5). For instance, pupils who achieved an A* in GCSE mathematics and also completed the PISA assessment obtained an average PISA mathematics score of 577. Those who achieved an A scored 541, and so forth. The likely bias can then be estimated by multiplying the difference between the PISA sample and the national population in each GCSE grade category (column 4) by the average PISA score associated with each grade (column 5), then summing together the respective values. Concretely, the calculation in the example above is:

$$\begin{aligned} &2.4\%*577 + 2.8\%*541 + 2.3\%*504 + 3.3\%*474 - 0.4\%*437 - 2.5\%*413 - 2.2\%*387 - 2.6\%*382 - \\ &3.2\%*359 \\ &= 14.2 \end{aligned}$$

This puts the upward bias in the mean PISA mathematics score for Wales at around 14 test points (around 0.14 standard deviations)⁹. They note that, using a similar methodology, the upward bias in England's scores around be around 11 test points.

The approach described above relies, of course, on suitable auxiliary information being available. But what can researchers and officials do when they do not have this luxury? This was the situation faced by Anders et al. (2021) in their analysis of PISA 2015 data for Canada, and by Andersson and Sandgren Massih (2023) in their analysis of Swedish PISA 2018 data. In such cases, one can use information from published PISA reports and the international PISA database to estimate the proportion of the target population not covered by the assessment because of exclusions and non-response. For example, Anders et al. (2021) estimate that 47% of the Canadian target population was not covered in PISA 2015. They then explored how Canada's average PISA score would change under different assumptions about how these uncovered pupils would have performed had they taken the test.

⁹ An important implicit assumption being made is that, within each GCSE category, participating and nonparticipating pupils would have had the same mean PISA mathematics score.

Table 2.9.2 illustrates this approach, broadly following Anders et al.'s analysis. The average reading score among Canadian pupils who completed PISA was 527. The left-hand column then presents alternative assumptions about the average score of the uncovered 47%. For instance, if they are assumed to have scored 520 points on average, the revised national estimate is:

$$\begin{aligned}
 &53\% \cdot 527 + 47\% \cdot 520 \\
 &= 0.53(527) + 0.47(520) \\
 &= 523.71 \\
 &\approx 524.
 \end{aligned}$$

Based on the formula:

$$\hat{u}_{all} = q \cdot \bar{S}_{observed} + (1 - q) \cdot \bar{S}_{uncovered} \quad (\text{Equation 2.9.1})$$

Where: $q = \frac{W}{N}$

N = Size of the relevant population (e.g. all 15-year-olds in a country)

W = Weighted number represented by assessed participants

S_{obs} = Observed mean score of pupils that participated in PISA.

S_{ass} = Assumed mean for everyone not represented within the data

With the observed bias then equal to $(1 - q) \cdot (\bar{S}_{observed} - \bar{S}_{uncovered})$.

Table 2.9.2. Example of simulating potential bias in average ILSA scores when no auxiliary information related to the outcome is available

Assumed score of non-participants	Revised country average score
527 (same as participants)	527
520	524
510	519
500	514
490	510
480	505
470	500

Each row of Table 2.9.2 then shows how the revised estimate of Canada's average PISA score changes as progressively lower scores are assumed for pupils not covered by the assessment. The logic is the same as in Section 2.3 and Figure 2.3.1; the implied upward bias increases as the assumed performance gap between assessed participants and uncovered pupils increases. If non-participants would have performed similarly to participants, the effect on the national

average is small. If, however, non-participants would have scored substantially lower, the upward bias in the published estimate becomes larger. Of course, we do not know which (if any) of these simulated scenarios is correct. Nevertheless, the exercise is useful because it allows officials and researchers to examine a plausible range of potential bias under different assumptions.

An important limitation of both approaches outlined above is that ILSAs are usually used to produce comparative estimates (either between countries or within a country over time). Evidence of bias in one country's estimate does not automatically tell us the extent of bias in the comparison of interest. For example, returning to the case of Wales, suppose that average PISA scores are estimated to be upwardly biased by around 14 points. This does not necessarily mean that the difference between Wales and another country is biased by 14 points. If similar upward bias is also present in the comparator country, the bias in the difference between the two countries may be smaller. In this sense, two wrongs can (partially) make a right.

A similar point applies to trends over time. Non-random exclusions and non-random non-response may bias estimates in a given survey cycle. However, the estimated change in performance between cycles may be less affected if the size and direction of the bias have remained broadly stable over time.

2.10. Test your knowledge questions

2.10.1 Task 1.

Table 2.10.1.1 provides information on the “exclusion rate” for 10 countries. This includes both school level exclusions and pupil level exclusions.

Table 2.10.1.1 School and pupil exclusion rates across 10 hypothetical countries.

	School level exclusions %	Pupil level exclusions %
Country A	4	4
Country B	0	0
Country C	5	5
Country D	3	8
Country E	0	6
Country F	2	9
Country G	4	4
Country H	10	0
Country I	2	9
Country J	2	5

Task 1 Question 1. A policymaker says the following to you “*School level exclusions are far too high in country H. There is no way that the data for this country can be fairly compared to the data from other countries.*” What is your response?

Task 1. Question 2. How would your response to the policymaker change in the following three scenarios:

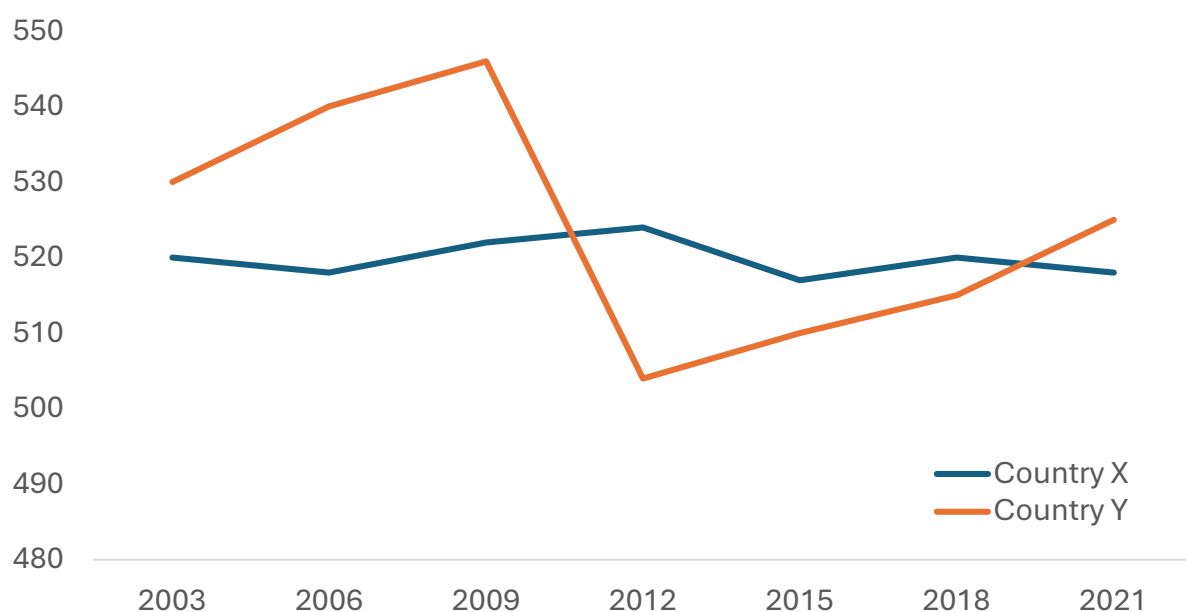
- Scenario (a): There was an administrative error in country H, which led to some school being randomly left off the list of schools that could be sampled.
- Scenario (b): The school exclusions in Country H were all in a geographically remote location. The demographic composition of this location is much like the rest of the country.
- Scenario (c): Schools that specialise in teaching pupils with Special Educational Needs have been excluded in Country H.

Task 1. Question 3. What additional data would you like to have access to in order to make further inferences about how these exclusions may have impacted the sample?

2.10.2. Task 2.

Figure 2.10.2.1 displays average scores for two countries (X and Y) on a mathematics test between 2003 and 2021. The standard deviation of the test score is around 100 points. The target population of the test is pupils who are aged 15-years-old and currently enrolled in school.

Figure 2.10.2.1. Trends in average ILSA test scores across two hypothetical countries



Task 2. Question 1. Please comment on the trend in the test scores across these two countries over time.

Task 2. Question 2. You are made aware of a policy change in country Y that occurred in 2010. This raised the minimum age young people could leave school from 14 to 17 years of age. Please comment on the results for country Y in light of this new information.

Task 2. Question 3. How do you think the above may impact the distribution of test scores in country Y (e.g. the standard deviation of achievement; the 10th percentile?)

Task 2. Question 4. A senior official from country Y wants to issue a press release with the headline “*our education policies have led to a sustained increase in educational performance over the last 20 years.*” Do you think this claim can be justified based on the above?

2.10.3. Task 3.

Table 2.10.3.1 provides information from a sample drawn for an international survey that has been linked to administrative records. The three columns in this table include the following information:

- National exam grade = The grades achieved in national examinations, where A refers to the top/best grade and J the worse/bottom grade.
- Sample = The percent of the sample participating in the international survey that obtained each grade in the national examination.
- Population= The percent of pupils in the target population that achieved each national examination grade.

Table 2.10.3.1. Comparisons of an ILSA sample to the population in national examination grades (hypothetical data)

National exam grade	Sample	Population
A (top grade)	4.0%	3.7%
B	8.4%	7.2%
C	10.9%	9.5%
D	12.7%	11.5%
E	20.2%	18.2%
F	22.5%	21.4%
G	11.3%	12.7%
H	6.7%	8.6%
I	3.2%	5.4%
J (bottom grade)	0.0%	2.0%

Task 3. Question 1. Please state as many reasons as you can think of for why the distribution of national examination grades may differ across the ILSA sample and the population.

Task 3. Question 2. What do you think the above implies if one were to use the ILSA sample data to estimate average mathematics scores in the population of interest.

Task 3. Question 3. You are now provided with Table 2.10.3.2. This includes details on how average scores on the ILSA test vary by national examination grade. Please try to use this information to estimate the likely bias in average test scores based on the ILSA sample.

Table 2.10.3.2. Average ILSA test scores of pupils achieving different national examination grades

National exam grade	Average test scores
A (top grade)	620
B	585
C	561
D	532
E	506
F	467
G	434
H	396
I	364
J (bottom grade)	337

Chapter 2 References

Anders, J., Has, S., Jerrim, J., Shure, N., & Zieger, L. (2021). Is Canada really an education superpower? The impact of non-participation on results from PISA 2015. *Educational Assessment, Evaluation and Accountability*, 33, 229–249. <https://doi.org/10.1007/s11092-020-09329-5>

Andersson, C., & Sandgren Massih, S. (2023). PISA 2018: did Sweden exclude students according to the rules? *Assessment in Education: Principles, Policy & Practice*, 30(1), 33–52. <https://doi.org/10.1080/0969594X.2023.2189566>

Atasever, U., Jerrim, J., & Tieck, S. (2024). Exclusion rates from international large-scale assessments: An analysis of 20 years of IEA data. *Educational Assessment, Evaluation and Accountability*, 36, 405–428. <https://doi.org/10.1007/s11092-023-09416-3>

Durrant, G. B., & Schnepf, S. V. (2018). Which schools and pupils respond to educational achievement surveys? A focus on the English Programme for International Student Assessment sample. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4), 1057–1075. <https://doi.org/10.1111/rssa.12337>

Jerrim, J. (2017, July 19). *Why does Vietnam do so well in PISA? An example of why naive interpretation of international rankings is such a bad idea*. FFT Education Datalab.

<https://ffteducationdatalab.org.uk/2017/07/why-does-vietnam-do-so-well-in-pisa-an-example-of-why-naive-interpretation-of-international-rankings-is-such-a-bad-idea/>

Jerrim, J., & Sims, S. (2019). *The Teaching and Learning International Survey (TALIS) 2018* (Research Report DFE-RR938). Department for Education. https://assets.publishing.service.gov.uk/media/5f6484c2e90e075a01d2f4ce/TALIS_2018_research.pdf

Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *Review of Education*, 9(3), Article e3270. <https://doi.org/10.1002/rev3.3270>

Jerrim, J., & Savaşçı-Smith, D. (2025). Data collection during a pandemic: How did COVID-19 impact the sample composition in PIRLS 2021? In Ž. Džumhur, P. Koršňáková, & S. Meinck (Eds.), *Dinaric perspectives on PIRLS 2021* (IEA Research for Education, Vol. 17, pp. 179–195). Springer. https://doi.org/10.1007/978-3-031-88002-5_8

LaRoche, S., & Foy, P. (2020). Sample implementation in TIMSS 2019. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 9.1–9.261). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch9-sample-implementation.pdf

Micklewright, J., Schnepf, S. V., & Skinner, C. (2012). Non-response biases in surveys of schoolchildren: The case of the English Programme for International Student Assessment (PISA) samples. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(4), 915–938. <https://doi.org/10.1111/j.1467-985X.2012.01036.x>

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>

OECD. (2024). *Annex 13.A: Sampling outcomes* [Data set]. In *PISA 2022 technical report*. OECD Publishing. <https://stat.link/mou0ea>

OECD (2024), *PISA 2022 Technical Report*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/01820d6d-en>.

Siegel, P., & Foy, P. (2024). *TIMSS sample design*. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report: Methods and procedures* (pp. 3.1–3.30). TIMSS & PIRLS International Study Center, Boston College. <https://doi.org/10.6017/lse.tpisc.timss.rs7952>

Spaull, N. (2019). Who makes it into PISA? Understanding the impact of PISA sample eligibility using Turkey as a case study (PISA 2003–PISA 2012). *Assessment in Education: Principles, Policy & Practice*, 26(4), 397–421. <https://doi.org/10.1080/0969594X.2018.1504742>

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *TIMSS 2023 Technical Report (Methods and Procedures)*. Boston College, TIMSS & PIRLS International Study Center. <https://timss2023.org/methods>

Westat. (2016). *Sampling in PISA*. OECD. <https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/pisa-database/survey-implementation-tools/pisa-2018/SAMPLING-IN-PISA.pdf>

3. Accounting for complex survey designs

3.1 Chapter introduction

The previous chapter described several complex features of ILSA survey designs. This included the definition of the target population, the use of stratification and clustering when selecting participants, oversampling, exclusions and survey non-response. Each of these features has important implications for subsequent analysis. They affect how representative the achieved sample is likely to be, how much uncertainty surrounds estimates and how survey weights and standard errors should be calculated. Consequently, ILSA data cannot be analysed as if they came from a simple random sample. The design of the study must be taken into account when producing population estimates and drawing statistical inferences.

This chapter describes how these complexities are handled in the analysis of ILSAs. The focus is primarily on intuition and simple numerical examples, rather than on formal statistical derivations. It begins by introducing survey weights, explaining how they are derived and the role they play in producing estimates that are representative of the target population. It then provides an intuitive introduction to replication methods for estimating sampling variation under complex survey designs. The chapter also considers the advantages and limitations of these approaches relative to other commonly used alternatives.

3.2 Sample weights

The previous chapter noted that some schools, regions or groups may be oversampled to support reliable subgroup estimates, such as estimates for Indigenous pupils in Australia or provincial-level results in Canada. It also explained that individual schools often have unequal probabilities of selection because schools are sampled with probability proportional to size. In addition, non-response means that some originally selected schools, pupils or teachers do not appear in the final sample. Together, these features mean that different sampled units may represent different numbers of individuals in the target population.

Survey weights are derived to account for these unequal probabilities of selection and participation. Their purpose is to try and ensure that estimates from the achieved sample can be generalised appropriately to the population of interest. Although weights are important in ILSAs, they are not unique to them (they are widely used across social surveys more generally). The technical reports for each ILSA provide detailed explanations of how weights are constructed

(e.g. OECD, 2024, Chapter 10; Siegel & Foy, 2024). Here, however, the focus is on the intuition behind weighting rather than the technical details.

Broadly speaking, ILSA weights have three components.

The first component is the base weight. At the school level, the base weight is essentially the reciprocal of the probability that a given school is selected into the sample. For example, if a school has a one-in-100 chance of selection, its base weight would be 100. At the individual level, the base weight reflects the reciprocal of the probability that an individual is selected, conditional on their school already having been chosen. These selection probabilities are known before data collection begins, because they follow directly from the sample design. Base weights therefore account for design features such as probability proportional to size sampling and any deliberate oversampling. In effect, they indicate how many population units each sampled unit is meant to represent.

The second component is the non-response adjustment. These adjustments are designed to compensate for the fact that some sampled units do not participate in the study. At the school level, non-response adjustments are usually based on the variables used to stratify the sample (as discussed in chapter 2). For example, participating schools may be reweighted to represent non-participating schools within the same region or school type. At the individual level, however, non-response adjustments in ILSAs tend to be more limited, often relying on basic characteristics such as school, gender and/or grade. The effectiveness of these adjustments depends on whether the variables used to construct them are predictive of both non-response and the outcome being estimated. At the school level, their effectiveness is therefore likely to vary across countries, since countries differ in the stratification variables used. At the individual level, the adjustments may do relatively little to correct for selective non-response, given the limited set of variables typically available (Anders et al., 2021).

Finally, weights may be trimmed, both at the school and individual level. This means placing an upper limit on the size that any weight is allowed to take. In doing so, one avoids a situation where a small number of schools or individuals have a disproportionate influence on population estimates. For instance, extremely large weights can make estimates unstable, increase standard errors and reduce statistical power. Trimming is therefore used to balance representativeness against precision.

The final individual sample weight is then essentially the product of these components (the base weight, the non-response adjustment and any trimming applied).

Table 3.2.1 provides some hypothetical data that allows us to illustrate how weights are used in the context of a simple analysis. Here we have a sample of 10 individuals, along with their sample weight and a test score.

Table 3.2.1. Hypothetical data to illustrate how weights should be used when calculating population averages

ID	Weight	Score
1	25	561
2	20	523
3	10	544
4	15	556
5	20	537
6	5	448
7	10	473
8	10	439
9	5	469
10	2	452

If we don't apply the weight - and thus produce the sample average - one would reach an average score of:

$$\frac{561 + 523 + 544 + 556 + 537 + 448 + 473 + 439 + 469 + 452}{10} = 500.2$$

In other words, in such an analysis, one is implicitly assuming that everyone in the sample represents the same number of individuals in the population.

In contrast, to get the weighted average, the calculation becomes:

$$\frac{(25 * 561 + 20 * 523 + 10 * 544 + 15 * 556 + 20 * 537 + 5 * 448 + 10 * 473 + 10 * 439 + 5 * 469 + 2 * 452)}{(25 + 20 + 10 + 15 + 20 + 5 + 10 + 10 + 5 + 2)}$$

$$\frac{63,614}{122} = 521.4$$

Thus - in the weighted average - different individuals in the sample represent a different number of individuals in the broader population.

Note that the difference between the unweighted and weighted averages in this example is substantial. This occurs because individuals 6 to 10 have both lower test scores and lower sample weights. If the weights are ignored, these individuals have too much influence on the analysis relative to the number of people in the population that they represent. As a result, the unweighted estimate places excessive emphasis on lower-scoring individuals and pulls the average downward.

Table 3.2.2 illustrates how weighting can affect substantive estimates, using PISA data for the United Kingdom. The UK is divided into four constituent countries (England, Scotland, Northern Ireland and Wales). England is by far the largest, accounting for around 83% of the relevant UK population. However, to produce separate PISA estimates for each constituent country, schools in Scotland, Northern Ireland and Wales are oversampled. As a result, pupils in England make up only 34% of the UK PISA sample, despite representing 83% of the population.

If the weights are ignored, the average UK PISA score would be calculated as:

$$\frac{(4081 \times 495) + (2613 \times 499) + (2197 \times 494) + (3270 \times 473)}{12,179} = 489.8$$

On the other hand, when the survey weights are applied, the average becomes:

$$\frac{(570,080 \times 493) + (54,884 \times 499) + (23,151 \times 492.2) + (35,264 \times 472.4)}{683,379} = 492.4$$

The difference is driven largely by Wales, which has a notably lower PISA score than the other UK countries (a difference of around 20 points). Without weights, the oversampling of Wales is not accounted for, so Welsh pupils receive too much influence in the UK estimate. Specifically, Welsh pupils end up contributing 27% of the sample in the unweighted estimates rather than the 5% contribution implied by its population share.

Table 3.2.2. Weighted and unweighted averages using PISA data from the UK

	With weights			Without weights		
	Estimated target-population	% of total	Mean	Sample size	% of total	Mean
England	570,080	83.4%	493	4,081	33.5%	495
Scotland	54,884	8.0%	499	2,631	21.6%	499
Northern Ireland	23,151	3.4%	492.2	2,197	18.0%	494
Wales	35,264	5.2%	472.4	3,270	26.8%	473
Total (Whole UK)	683,379	100	492.4	12,179	100	489.8

For applied researchers, the key message is that survey weights should be applied wherever possible. This applies not only when estimating final statistics or regression models, but also during intermediate stages of analysis. For example, weights should be used when standardising a variable, such as transforming it to have a mean of zero and a standard deviation of one, or when dividing a continuous variable into groups, such as socioeconomic status quartiles. Most statistical software makes this straightforward, as discussed in section 1.7, but analysts still need to remember to do so. As the examples above illustrate, weights are likely to make the greatest difference when groups are an outlier in terms of the outcome of interest and have particularly high or low weights relative to others (e.g. the example of Wales in Table 3.2.2).

Readers should also note that there may be circumstances when one wishes to rescale the weights supplied in the international database. We return to this issue in chapter 4 (section 4.6) in our discussion of so-called “house” and “senate” weights.

3.3 Replication methods to calculate standard errors

We next turn to how stratification and clustering are accounted for in one’s statistical analysis. Unlike survey weights, these features of ILSA sample designs do not usually affect the point estimate itself. Rather, they affect the uncertainty surrounding that estimate (standard errors, confidence intervals, p-values and tests of statistical significance).

ILSAs typically use replication methods to obtain appropriate estimates of sampling variance and associated standard errors under complex survey designs. These methods are not usually covered in most undergraduate statistics courses and are often absent from postgraduate social science training as well. The basic intuition is that replication methods provide an empirical way of estimating sampling uncertainty. They use information about how the sample was designed - together with the data actually collected - to assess how much estimates might vary if a different sample had been drawn.

Because replication methods may be unfamiliar, I begin by explaining the intuition behind one widely used approach - the bootstrap. It is important to note that ILSAs do not generally use bootstrapping itself. Rather they use related replication methods (e.g. jackknife and balanced-repeated-replication) that rely on a broadly similar logic. The purpose of beginning with the bootstrap is therefore pedagogical. It is one of the best-known replication methods in the social

sciences and hence provides a useful entry point for understanding how replication approaches estimate sampling uncertainty.

To begin, I discuss bootstrapping in the context of a simple random sample. This deliberately abstracts from the complexities of ILSA survey designs, allowing us to focus first on the underlying intuition. Table 3.3.1 presents a hypothetical target population of 100 individuals, represented by 100 rectangles. This is the population about which we wish to make inferences. From this population, 10 individuals are randomly selected and labelled A to J. For these 10 sampled individuals, data are collected and a statistic of interest (e.g. the mean) is calculated. A hypothetical example of these observed data and the resulting statistic is shown in Table 3.3.2.

Table 3.3.1. Hypothetical population of 100 individuals with a simple random sample of 10 individuals drawn.

					F				
	B						I		
				E					
A						G			J
			C			H			
				D					

Table 3.3.2. Data collected from the sample of 10 individuals on our statistic of interest

Main sample	Outcome value
A	2
B	4
C	5
D	6
E	8
F	2
G	4
H	4
I	2
J	3
Sample mean	4

Next comes the conceptually tricky part. From the 10 individuals for whom data have been collected (A to J) we draw another sample of 10 individuals. Crucially, this is done with replacement. In other words, we take a new random sample from the original sample, allowing the same individual to be selected more than once. This new sample is called the first bootstrap resample (or replicate sample).

Table 3.3.3 provides an illustrative example. In this bootstrap resample, individual A appears three times. This is possible because, after A has been selected once, they are returned to the pool and can be selected again. Similarly, individual F appears twice. In contrast, individuals B, C and G do not appear in this particular resample at all. This is a standard feature of bootstrap resampling; within any given resample some individuals will be selected multiple times while others won't be selected at all.

Using this bootstrap resample, we can calculate an alternative estimate of the sample mean. In this example, this produces a value of 3.4. This is then my first bootstrap replicate estimate. It can be interpreted as one plausible estimate of the population value that might have been obtained had a different sample been drawn from the population.

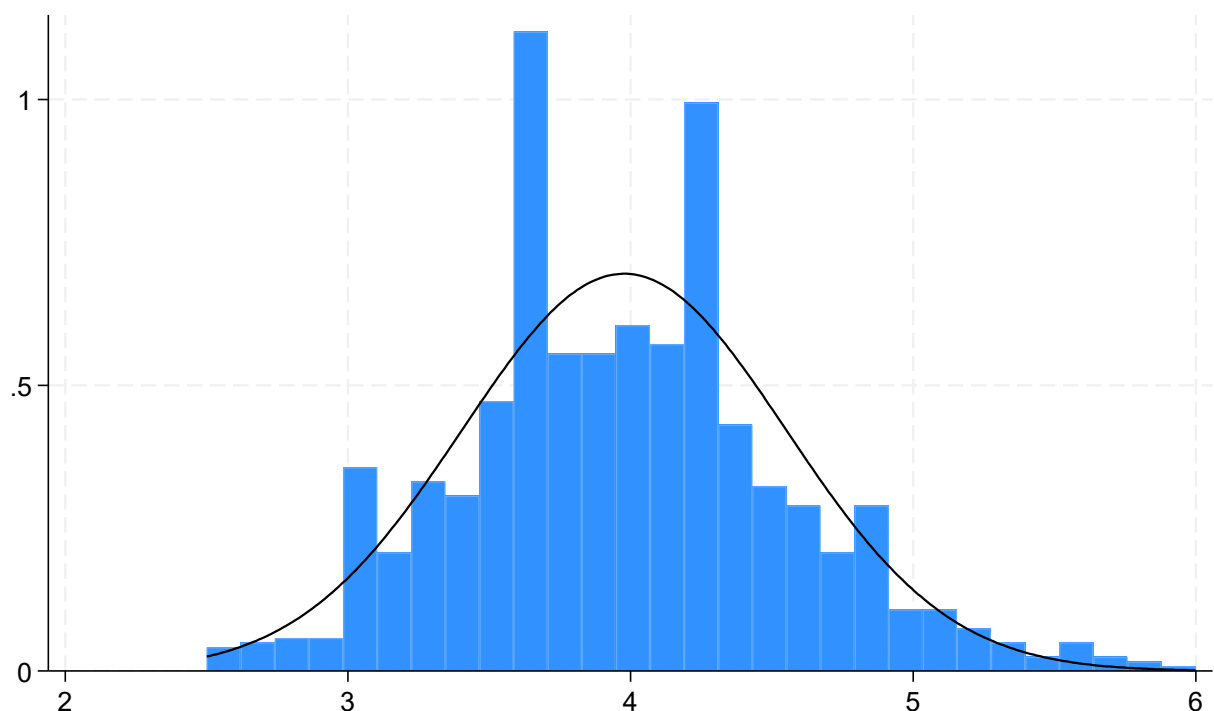
Table 3.3.3. The first bootstrap resample from our statistic of interest

Resample draw	Replicate 1 sample	Outcome value
Random selection 1	D	6
Random selection 2	F	2
Random selection 3	J	3
Random selection 4	F	2
Random selection 5	E	8
Random selection 6	A	2
Random selection 7	H	4
Random selection 8	A	2
Random selection 9	A	2
Random selection 10	J	3
Replicate 1 estimate		3.40

We do not carry out this resampling process only once. Instead, we repeat it many (hundreds or even thousands) of times. Each repetition creates a new bootstrap resample (similar to another version of Table 3.3.3) with each resample producing its own alternative estimate of the statistic of interest. We therefore end up with one main estimate (calculated from the original sample in Table 3.3.2) and many alternative estimates (calculated from the many bootstrap resamples).

If we plot these replicate estimates, the resulting distribution will look something like the histogram shown in Figure 3.3.1. It will be centred close to the estimate from the original sample (4) and will be approximately normally distributed. This distribution approximates the range of estimates that might have been obtained if different samples had been drawn from the population. The bootstrap estimate of the sampling standard error is then simply the standard deviation of this distribution (i.e. the standard deviation of the statistic of interest across the many bootstrap resamples).

Figure 3.3.1. The distribution of 1,000 bootstrap re-estimates of the statistic of interest from our hypothetical population



The information contained in the bootstrap resamples can also be presented in a different way. Return to Table 3.3.3, which showed the first bootstrap replicate sample. An alternative presentation of this same information is given in Table 3.3.4. This table starts with the original sample data from Table 3.3.2, but adds an additional column labelled “Replication Weight”. This replicate weight records how many times each individual from the original sample appears in the first bootstrap resample. For example, individual A has a replicate weight of 3, because they were selected three times in the bootstrap resample. Individual F has a replicate weight of 2 for the same reason. In contrast, individuals B and C have replicate weights of 0, indicating that they were not selected into this particular bootstrap resample at all.

Table 3.3.4. Illustration of the first bootstrap replicate sample presented in the form of a replication weight

Replicate 1 sample	Replication Weight (# times selected into the resample)	Outcome value
A	3	2
B	0	4
C	0	5
D	1	6
E	1	8
F	2	2
G	0	4
H	1	4
I	0	2
J	2	3

We can now use the bootstrap replicate weight to produce our first bootstrap re-estimate of the statistic of interest:

$$\frac{(3 \times 2) + (0 \times 4) + (0 \times 5) + (1 \times 6) + (1 \times 8) + (2 \times 2) + (0 \times 4) + (1 \times 4) + (0 \times 2) + (2 \times 3)}{10} = 3.4$$

This is the same estimate as presented in Table 3.3.3. The only difference is that the information about the bootstrap resample is now expressed as a set of replicate weights. It therefore follows that information from multiple bootstrap resamples can be represented using multiple sets of replicate weights. These replicate weights can then also be used to produce the distribution shown in Figure 3.3.1 and thus the associated standard error.

The example above illustrates how replication methods work under a simple random sample. But what happens when we move to the more complex survey designs used in ILSAs, as discussed in the previous chapter? The general logic is the same; alternative versions of the sample are created, the statistic of interest is recalculated for each version and the variation across these replicate estimates is used to estimate sampling uncertainty.

The key difference lies in how the replication samples are constructed. Crucially, the resamples must reflect the way the original sample was drawn. In most ILSAs, schools are the primary sampling units in a clustered design. This means that replication resamples should be created by selecting entire schools (rather than individual pupils) with replacement. Similarly, where the original design involved stratification, schools should be resampled within their respective strata.

Table 3.3.5 provides an illustrative example of a hypothetical stratified cluster sample. In this example, the sample design contains two strata (A and B). Within each stratum, 10 schools are selected, giving a total sample of 20 schools.

Table 3.3.5. A hypothetical sample that has been drawn with two strata (A and B) each with 10 schools (1 – 10)

Strata A. School 1	Strata B. School 1
Strata A. School 2	Strata B. School 2
Strata A. School 3	Strata B. School 3
Strata A. School 4	Strata B. School 4
Strata A. School 5	Strata B. School 5
Strata A. School 6	Strata B. School 6
Strata A. School 7	Strata B. School 7
Strata A. School 8	Strata B. School 8
Strata A. School 9	Strata B. School 9
Strata A. School 10	Strata B. School 10

Table 3.3.6 illustrates a stratified cluster bootstrap, in which schools are resampled with replacement within strata. This preserves the clustering and stratification in the hypothetical design. In each resample, 10 schools are selected with replacement from stratum A, with a further 10 schools selected with replacement from stratum B. For example, in the first replication resample, School 1 from stratum A is selected twice (meaning all pupils in this school enter the resample twice). Schools 2 and 3 from stratum A are each selected once, and so forth. In stratum B, Schools 1 and 2 are not selected in the first resample, while School 3 is selected three times etc.

Otherwise, the process and intuition are the same as under simple random sampling discussed above. The key point is that, in more complex survey designs involving stratification and clustering, the resamples must be drawn in a way that reflects this complexity. Only then will the resulting replicate estimates provide an appropriate basis for estimating sampling uncertainty.

Table 3.3.6. Hypothetical resamples of the schools (clusters) within strata

Strata	Cluster	Resample 1	Resample 2	
A	A1	2	0	
A	A2	1	0	
A	A3	1	2	
A	A4	0	2	
A	A5	3	2	
A	A6	0	1	
A	A7	0	1	
A	A8	1	1	
A	A9	1	0	
A	A10	1	1	
B	B1	0	3	
B	B2	0	0	
B	B3	3	2	
B	B4	2	1	
B	B5	1	1	
B	B6	1	0	
B	B7	0	1	
B	B8	2	1	
B	B9	1	0	
B	B10	0	1	

Notes: Figures refer to the number of times the school has been selected into the resample.

3.4 Balanced repeated replication and jackknife replication methods.

The previous section introduced the intuition behind replication methods using the example of bootstrapping. However, bootstrapping (BS) is only one member of a wider family of replication approaches. Other methods include the jackknife (JK) and balanced repeated replication (BRR). It is these approaches that are actually used in ILSAs. The broad logic is similar across these approaches; the statistic of interest is repeatedly re-estimated using alternative versions of the sample, with the variation across these estimates used to calculate the standard error. The main difference between them thus lies in how the alternative samples (and thus the replicate weights) are constructed. While full details are outside the scope of this book, a short summary of how these approaches differ in the construction of the replicate weights is as follows:

- Bootstrap. The data are repeatedly resampled with replacement.
- Jackknife. One (or more) observations or groups are left out of the data at a time.
- Balanced repeated replication. Repeatedly splits the data into half-samples¹⁰.

¹⁰ A common variant is “Fay’s adjustment” to BRR, where the half-splits of the data are up/down weighted rather than simply retained/dropped. This is done to improve the stability of the weights across the replicates.

PISA and TALIS both use Fay's BRR-based approaches, while TIMSS and PIRLS use jackknife repeated replication. In PIAAC, both jackknife and BRR approaches have been used. The preference for one replication approach over another is usually driven by technical features of the survey design. Readers interested in these details are directed to Foy and Almaskut (2023) for discussion of the jackknife and to OECD (2009, chapter 4) for balanced repeated replication. The key point for the purposes of this book is that each approach can be understood as generating a set of replicate weights similar to the bootstrap example shown in Table 3.3.4. These replicate weights are then applied repeatedly to the same dataset, with the statistic of interest recalculated each time. The variation across these replicate estimates provides an approximation of sampling variability, while accounting for complex design features such as stratification and clustering.

We conclude this subsection with a hypothetical numerical example showing how individual weights (discussed in Section 3.2) and replicate weights (discussed in this section) are used together to produce both the final estimate and its associated sampling variance. In this example, we assume that the replicate weights have been created using balanced repeated replication (BRR) without Fay's adjustment¹¹. The final point estimate is first calculated using the main individual sample weight. The same statistic is then recalculated multiple times using each set of BRR replicate weights. The sampling variance is obtained by comparing each replicate estimate with the full-sample estimate, using the following formula¹²:

$$\sigma^2 = \frac{\sum_{r=1}^R (\widehat{\theta}_r - \widehat{\theta})^2}{R} \quad (\text{Equation 3.4.1})$$

Where:

σ^2 = The sampling variance of the estimate

$\widehat{\theta}_r$ = The estimate of the statistic of interest applying replicate weight R.

$\widehat{\theta}$ = The estimate of the statistic of interest applying the main sample weight.

R = Number of replication weights.

¹¹ While the specific formula differs slightly for other replication approaches, such as the jackknife, the underlying logic and much of the mechanics remain very similar.

¹² If the BRR weights are created using Fay's adjustment, the formula becomes $\sigma^2 = \frac{\sum_{r=1}^R (\widehat{\theta}_r - \widehat{\theta})^2}{R(1-k)^2}$. Where k is the adjustment factor. Fay's adjustment is used to create the PISA and TALIS BRR weights, with k set to 0.5.

Ignoring other features of the ILSA test design (to be discussed in chapter 6) the square root of this quantity gives the standard error.

Note that this formula should look reasonably familiar. It closely resembles the formula for the standard deviation, being based on the average of squared differences. This makes intuitive sense, given the discussion of Figure 3.3.1 in the previous section.

The data we use are presented in Table 3.4.1. Score represents the students' scores on an ILSA assessment. Sample weight refers to the main sample weight. R1 to R4 refer to four replicate weights.

Table 3.4.1. Hypothetical data to illustrate the use of final weights and replication weights

ID	Score	Final weight	R1	R2	R3	R4
1	503	8	16	0	16	0
2	513	6	12	0	12	0
3	489	8	0	16	0	16
4	538	8	16	16	0	0
5	446	12	24	24	0	0
6	489	16	0	0	32	32
7	539	14	0	0	28	28
8	500	10	20	0	0	20
9	486	12	0	24	24	0
10	532	6	0	12	12	0

We first calculate our point estimate of interest applying the final sample:

$$\frac{((503 * 8) + (513 * 6) + (489 * 8) + (538 * 8) + (446 * 12) + (489 * 16) + (539 * 14) + (500 * 10) + (486 * 12) + (532 * 6))}{100}$$

Thus $\hat{\theta} = 500.64$. This is our estimate of the population mean score.

Next, we repeat the same calculation four further times, once using each replicate weight:

$$\begin{aligned}\hat{\theta}_1 &= \frac{((503 * 16) + (513 * 12) + (489 * 0) + (538 * 16) + (446 * 24) + (489 * 0) + (539 * 0) + (500 * 20) + (486 * 0) + (532 * 0))}{16 + 12 + 0 + 16 + 24 + 0 + 0 + 20 + 0 + 0} \\ \hat{\theta}_2 &= \frac{((503 * 0) + (513 * 0) + (489 * 16) + (538 * 16) + (446 * 24) + (489 * 0) + (539 * 0) + (500 * 0) + (486 * 24) + (532 * 12))}{0 + 0 + 16 + 16 + 24 + 0 + 0 + 0 + 24 + 12} \\ \hat{\theta}_3 &= \frac{((503 * 16) + (513 * 12) + (489 * 0) + (538 * 0) + (446 * 0) + (489 * 32) + (539 * 28) + (500 * 0) + (486 * 24) + (532 * 12))}{16 + 12 + 0 + 0 + 0 + 32 + 28 + 0 + 24 + 12} \\ \hat{\theta}_4 &= \frac{((503 * 0) + (513 * 0) + (489 * 16) + (538 * 0) + (446 * 0) + (489 * 32) + (539 * 28) + (500 * 20) + (486 * 0) + (532 * 0))}{0 + 0 + 16 + 0 + 0 + 32 + 28 + 20 + 0 + 0}\end{aligned}$$

Which leads to values of:

$$\hat{\theta}_1 = 494.5$$

$$\widehat{\theta}_2 = 491.1304$$

$$\widehat{\theta}_3 = 508$$

$$\widehat{\theta}_4 = 505.875$$

These values can then be plugged into the formula presented in Equation 3.4.1:¹³

$$\frac{(494.5 - 500.64)^2 + (491.13 - 500.64)^2 + (508 - 500.64)^2 + (505.875 - 500.64)^2}{4} = 52.43$$

Which gives a sampling variance of 52.43. Ignoring any further features of the ILSA test design (covered in chapter 6), one can take the square root of this value to get a standard error of 7.24.

Hence, we estimate the population mean to be 500.64, with the sampling standard error of 7.24.

The construction of the replicate weights and the variance scaling factor differ across replication methods, although the general principle of comparing replicate estimates with the full-sample estimate remains the same.

3.5 Other methods that can be used to account for complex survey designs.

While the previous section discussed replication methods for accounting for the complex survey designs used in ILSAs, these are not the only approaches available. Alternative methods may be more familiar to some readers, since they are commonly used in social science research and are more frequently taught in postgraduate research methods courses. One such approach is to estimate standard errors using analytic formulas that make appropriate corrections for key aspects of the survey design.

For example, if the sample design is clustered (e.g. schools selected as the primary sampling units as in many ILSAs) then Taylor linearization can be used to approximate the sampling variability under a complex design (Williams, 2008). In practice, such adjustments can be implemented relatively easily in standard statistical software. For instance, the “cluster” and “svy” options are available in STATA, while survey design packages are also available in R (e.g. “survey”, “srvyr”). This approach can also be used to adjust standard errors to reflect stratification.

¹³ With Fay replicate weights, both the replicate estimates and the variance scaling differ. Once the replicate estimates have been calculated from the appropriate Fay weights, the variance divisor is $R(1 - k)^2$.

Another common way to account for clustering in survey data is to estimate a multilevel (i.e. random effects) model. In the context of ILSAs, pupils would typically form the level 1 units and schools the level 2 units. The multi-level model would then include a school-level random intercept term. We return briefly to the issue of estimating multilevel models with ILSA data in the following section.

Readers may naturally ask why ILSAs therefore rely on replication methods to estimate sampling uncertainty rather than these alternative approaches? Kreuter and Valliant (2007) provide a useful discussion of this issue.

First, replication methods do not require cluster and stratification variables to be included in the public-use dataset (although many ILSAs do). Instead, analysts only need the final individual and replicate weights supplied with the data. This can be advantageous because replicate weights are often less sensitive than detailed design variables and can help preserve respondent anonymity. By avoiding the release of detailed information about the sampling structure, replication approaches reduce the risk that particular schools or individuals could be identified from the data.

Second, replication methods do not require analysts to know every detail of the survey design. The relevant information needed to estimate appropriate standard errors (taking account all the features of the complex survey design) should already be embedded in the replicate weights. Thus analysts can simply apply the supplied replicate weights to obtain design-corrected standard errors. This should also help ensure the complex survey design is taken into account consistently across analyses.

Third, replication methods can be used to estimate standard errors for a wide range of statistics using the same general procedure. For example, the BRR formula in Equation 3.4.1 can be applied to any statistic that can be re-estimated using each set of replicate weights (e.g. means, totals, proportions, regression coefficients, group differences). In contrast, Taylor linearization requires statistic-specific approximations; different linearised variance estimators are needed for different quantities (e.g. means, totals, ratios and regression parameters). This makes replication weights a very flexible approach. It also means that replication methods should produce consistent values for the standard errors across different statistical software packages. Kreuter

and Valliant (2007) discuss these issues in their comparison of variance estimation methods for stratified and clustered samples.

Finally, some of the other approaches commonly used to analyse data from ILSAs account for only certain aspects of the complex survey design. For example, a school-level random-intercept (i.e. multi-level) model can model between-school heterogeneity and the dependence of observations within schools. However, unless the sampling weights and other relevant design features are explicitly incorporated, such a model does not automatically account for unequal selection probabilities, stratification or other aspects of the survey design. In this respect, replication methods are often better suited to producing standard errors that reflect the full survey design, rather than just the clustering of individuals within schools alone.

Replication methods do, of course, also have some limitations. In particular, they require statistics to be estimated multiple times (once using each set of replicate weights) and therefore often have longer run times than alternative approaches. In addition, when replicate weights are supplied directly in the dataset (e.g. as is the case in PISA and TALIS) they increase the size of the data file. These are, however, arguably relatively minor issues compared to the advantages set out above.

3.6 A note on multi-level (random effect) models.

As noted in the previous subsection, the use of replicate weights is generally recommended when analysing ILSA data, particularly when the aim is to produce standard errors that reflect the complex survey design. Despite this, multilevel models are frequently used in applied research with ILSAs. For example, Scherer (2022) rapid systematic review included 132 studies. This likely reflects the fact that multilevel models are much more familiar to many education and social science researchers. They are widely used for analysing clustered data and are commonly taught in postgraduate research methods courses. Replication methods, on the other hand, tend to receive less attention in standard statistical training. Multilevel models may also be more familiar to journal reviewers, who may expect to see them used when data have a hierarchical structure and may be less aware of replicate-weight procedures. As a result, analysts may default to multilevel modelling even in situations where replication methods would provide a more direct way of accounting for the sample design.

There are also disadvantages to using multilevel models in the analysis of ILSAs, although these are not always widely discussed. It is therefore useful to make these points explicit, not least because applied researchers are often asked by peer reviewers to use multilevel models simply because the data are clustered. Two points are particularly important.

First, standard multilevel models account for clustering by modelling the nesting of pupils within schools. However, they do not automatically account for other aspects of the complex survey design, particularly stratification. (Although methods have been developed to allow for multilevel models to additionally account for such features – see Asparouhov & Muthén 2006 for further details). If the primary goal is to obtain standard errors that appropriately reflect the full sample design, replication procedures may thus be preferable.

Second, the use of survey weights becomes more complicated in a multilevel modelling framework. Final individual ILSA weights are usually designed for single-level population estimation, whereas multilevel models may require weights to be specified separately at the school and pupil levels. In practice, this often involves decomposing the final student weight into a school-level weight and a conditional within-school student weight. The latter may also need to be rescaled, for example so that the level 1 (individual) weights sum to the observed or effective cluster size. These decisions can affect results and there is no consensus on the best approach. See Atasever et al. (2025) for further discussion.

In my view, for most analyses of ILSAs, the advantages of using multilevel models over replication approaches are hence not always clear. I therefore recommend using multilevel models selectively, where they are particularly well-suited to addressing the research question of interest. This includes:

- Where one aims to decompose variation in an outcome into individual and school-level components (i.e. estimating intra-cluster correlations).
- When one is specifically interested in estimating random slopes, where the association between an individual-level covariate and the outcome is allowed to vary across schools.
- When one's goal is to estimate school-level residuals with shrinkage, so that noisy school-level estimates are pulled towards the overall mean.

Yet such questions are rarely the central focus of most applied analyses of ILSAs. Rather, researchers are primarily interested in estimating population-level associations or group

differences. For such purposes, the benefits of multilevel modelling are likely to be more limited, with replication procedures therefore being more appropriate.

3.7 A case study comparing the different approaches

Given there are different approaches to estimating standard errors, it is reasonable to ask how much difference this choice makes in practice. The answer is, inevitably, analysis specific.

However, to provide some indicative evidence on this issue, Table 3.7.1 presents the sampling standard errors for average PISA scores from the 2022 cycle, calculated using different approaches. For simplicity, the illustration uses only the first “plausible value” (the meaning of “plausible values”, how they should be incorporated into analysis and the failure to do so are discussed later in the book - see chapter 6 section 6.5). The purpose here is more limited; to show how estimates of sampling uncertainty can vary depending on one’s approach to accounting for the complex survey design.

Five different approaches are presented:

- No adjustment. No adjustment for either clustering or stratification. This treats pupils as independent observations while still applying the final pupil weights.
- Taylor linearization to account for clustering.
- Taylor linearization accounting for both clustering and stratification.
- Two-level multilevel model, with pupils nested within schools, to account for clustering.
- BRR weights with Fay’s adjustment (the recommended approach).

In all five cases, the final survey weights are applied. Table 3.7.1 presents the estimated sampling standard errors under each approach.

A few notable results emerge. First, ignoring the complex survey design (i.e. treating the data as if it were drawn using a simple random sample) is a serious error. Doing so leads to substantially underestimated sampling standard errors, often less than half the size of those obtained using the recommended BRR approach. This has clear implications for inference, as it would make confidence intervals too narrow and increase the risk of overstating the statistical significance of cross-country differences.

Table 3.7.1. Estimated sampling standard errors using different approaches to account for complex survey designs. (PISA 2022).

	No adjustment	Linearization (clustering)	Linearization (clustering + stratification)	Multi-level model	BRR weights
Germany	1.27	4.25	4.25	4.24	3.01
UK	1.26	3.07	2.64	3.08	2.21
Mexico	0.94	2.64	2.49	2.67	2.22
Netherlands	1.59	7.27	3.44	7.30	3.82
Panama	1.14	3.96	3.26	3.97	2.76
Chinese Taipei	1.65	5.69	4.42	5.65	3.74
Vietnam	1.19	4.41	4.54	4.40	3.88

In contrast, the Taylor linearization and multilevel approaches generally produce standard errors that are larger than those obtained using BRR. These methods may therefore suffer from the opposite problem, producing estimates of the sampling standard errors that are too conservative. This appears to be partly because they do not fully account for stratification in the sampling design, although this does not seem to be the only factor. The results also highlight how particular features of the survey design matter more in some countries than others. For example, in Chinese Taipei and Vietnam, clustering has a substantial impact on the estimated standard errors. On the other hand, when stratification is not taken into account in the Netherlands, the standard errors roughly double (e.g. 3.8 under BRR versus approximately 7.3 under a multilevel model or a Taylor linearization approach that accounts only for clustering).

The results for Germany are particularly noteworthy – specifically the estimates based on Taylor linearization. Note that the values are identical regardless of whether stratification is accounted for or not. This is because the stratification variable available in the international database for Germany does not contain the relevant information; these details have not been disclosed. Consequently, for countries such as Germany, it is not possible to directly account for stratification when estimating sampling standard errors and the associated inferential statistics using the publicly available survey design variables. In practice, when using the publicly available PISA data, the only way to incorporate this aspect of the survey design in Germany is therefore to use the supplied BRR replicate weights. As discussed in section 3.5, this illustrates one of the key advantages of replication-based approaches - they do not require all survey design variables (such as stratification groups) to be disclosed.

3.8 Test your knowledge questions

3.8.1 Task 1

The data in the table below illustrates mathematics test scores for a sample of ten children in England and Wales. The columns are as follows:

- ID = The student ID
- Country = Whether the student lives in England or Wales.
- Weight = The final student weight of the country. The total of these weights is equal to the size of the student population.
- Score = The score that children achieved on the mathematics test

ID	Country	Weight	Score
1	England	120000	561
2	England	120000	523
3	England	120000	544
4	England	120000	556
5	England	120000	537
6	Wales	10000	448
7	Wales	10000	473
8	Wales	10000	439
9	Wales	10000	469
10	Wales	10000	452

Task 1. Question 1. Please calculate the average test score across all pupils in the sample. Please show your working.

Task 1. Question 2. Please calculate the average test score for (a) pupils in England and (b) pupils in Wales. Please show your working.

Task 1. Question 3. Please produce an estimate of the average test score for the population of pupils across England and Wales. (Do this either by hand or using Excel, rather than using statistical software such as R or STATA). Please show your working.

Task 1. Question 4. Why do your answers to questions 1 and 3 differ? Please give the intuition behind why the values differ.

Task 1. Question 5. If I was interested in producing an estimate of the average score in Wales, does it matter if I were to apply the weights or not?

Task 1. Question 6. Could you re-scale the weights in the table above to create “senate weights” – where each country would be given the sample emphasis in the analysis? (See chapter 4 – section 4.7 – if you need assistance).

3.8.2. Task 2.

You are provided with the following dataset, which contains mathematics test scores for one country. It includes the following columns:

- ID = Student ID.
- Score = The score the pupil achieved on the test.
- Weight = The final pupil weight.
- R1 to R4 = Five Balanced Repeated Replication (BRR) weights

ID	Score	Weight	R1	R2	R3	R4
1	532	3	6	0	6	0
2	510	4	8	0	8	0
3	498	5	0	10	0	10
4	557	10	20	10	0	0
5	425	2	4	4	0	0
6	478	6	0	0	12	12
7	521	3	0	0	6	6
8	512	5	10	0	0	10
9	499	4	0	8	8	0
10	511	3	0	6	6	0

Task 2. Question 1. Please produce an estimate of the average test score in this country. Please do this either by hand or using Excel (not R or STATA).

Task 2. Question 2. Now please create five alternative estimates of the population mean, once using each of the replication weights. Please do this either by hand or using Excel (not R or STATA).

Task 2. Question 3. Now please create an estimate of the standard error for the average score, based on the information provided.

Task 2. Question 4. Please replicate the analysis using either STATA or R.

3.8.3. Task 3.

A survey has been conducted which has a complex survey design. Schools are first selected as the primary sampling unit, with 30 pupils then randomly selected from within each school.

An analyst estimates a population mean reading score of 500 when applying the final respondent weight. Four BRR replicate weights are produced using Fay's adjustment (set to 0.5). Applying these BRR weights produces the following four alternative estimates:

Replicate	Estimate
1	494
2	506
3	498
4	502

Task 3. Question 1. Explain the different purposes of the final respondent weight and the replicate weights.

Task 3. Question 2. Calculate the BRR sampling variance and standard error. Show all main stages of the calculation.

Task 3. Question 3. Construct a 95% confidence interval using the BRR standard error.

Task 3. Question 4. Assume another researcher treats the data as if simple-random-sampling was used. When doing so, they produce a standard error for the mean reading score of 5. Please produce an alternative confidence interval using this alternative standard error.

Task 3. Question 5. Compare the confidence instances from questions 3 and 4 above. Explain why they are likely to differ.

Task 3. Question 6. A public dataset supplies replicate weights but does not release school identifiers or detailed stratification variables. Explain why valid standard errors can still be calculated. Please state one advantage and one disadvantage of supplying only the replication weights.

References

- Asparouhov, T., & Muthén, B. (2006). *Multilevel modeling of complex survey data*. In *Proceedings of the Joint Statistical Meeting in Seattle, ASA Section on Survey Research Methods* (pp. 2718–2726). American Statistical Association. <https://www.statmodel.com/download/SurveyJSM1.pdf>
- Atasever, U., Huang, F. L., & Rutkowski, L. (2025). Reassessing weights in large-scale assessments and multilevel models. *Large-scale Assessments in Education*, 13, Article 9. <https://doi.org/10.1186/s40536-025-00245-y>
- Bryan, M. L., & Jenkins, S. P. (2016). Multilevel modelling of country effects: A cautionary tale. *European Sociological Review*, 32(1), 3–22. <https://doi.org/10.1093/esr/jcv059>

Foy, P. and Almaskut, A. (2023). Estimating Standard Errors in the PIRLS 2021 Results. https://pirls2021.org/wp-content/uploads/2023/08/P21_MP_Ch13-standard-errors.pdf

Kreuter, F., & Valliant, R. (2007). A survey on survey statistics: What is done and can be done in Stata. *The Stata Journal*, 7(1), 1–21. <https://doi.org/10.1177/1536867X0700700101>

Organisation for Economic Co-operation and Development. (2009). *PISA data analysis manual: SPSS, second edition*. OECD Publishing. <https://doi.org/10.1787/9789264056275-en>

Organisation for Economic Co-operation and Development (2024), *PISA 2022 Technical Report*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/01820d6d-en>.

Scherer, R. (2022). Analyzing international large-scale assessment data with a hierarchical approach. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education* (pp. 1–55). Springer. https://doi.org/10.1007/978-3-030-38298-8_59-1

Siegel, P., & Foy, P. (2024). TIMSS sample design. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report: Methods and procedures* (pp. 3.1–3.30). TIMSS & PIRLS International Study Center, Boston College. https://timss2023.org/wp-content/uploads/2024/10/T23_TR_Chap3.pdf

Williams, R. (2008). Taylor series linearization (tsl). In *Encyclopedia of survey research methods* (Vol. 0, pp. 877-877). Sage Publications, Inc., <https://doi.org/10.4135/9781412963947.n572>

4. The statistical nuts and bolts of conducting international comparisons

4.1. Chapter introduction

The previous two chapters focused on the complex survey designs used in ILSAs and how these should be taken into account in one's analysis via the application of sampling and replicate weights. This chapter turns to a different set of challenges; the practical issues that arise when making international comparisons.

We begin with a deceptively simple question - which countries should I include in my analysis? Many ILSAs have grown substantially over time, with over 90 education systems participating in PISA 2025, and now include a heterogeneous set of participants. Including every available country may sometimes be appropriate but, in other cases, may produce comparisons that are difficult to interpret or of limited relevance to the research question. We thus discuss instances when broad international coverage is particularly useful, before contrasting this to examples when a focus on a smaller subset of countries may be more appropriate.

The chapter then considers how one can establish whether differences between countries (and change within a country over time) is statistically significant at the conventional 5% threshold. We begin with informal comparisons based on confidence intervals before introducing formal two-sample t-tests. This leads naturally to the problem of multiple hypothesis testing (i.e. when multiple countries are being compared, the chances of obtaining at least some statistically significant differences increases purely by chance). We therefore discuss common corrections for multiple comparisons, including their trade-offs between limiting false-positive findings and retaining sufficient statistical power. The chapter also briefly considers the common practice of pooling data across countries and estimating a three-level multilevel model (typically with individuals nested within schools nested within countries). In particular, we caution that such models involve important assumptions which are often unlikely to be met in the context of cross-national comparative research.

The final part of the chapter addresses several additional analytic choices that often arise in the analysis of international data. We first consider how to construct "international statistics" such as the average achievement score across a group of countries. This introduces the distinction between so-called house and senate averages (i.e. whether countries contribute according to their population size or are given equal influence in the international estimate). Our discussion includes an explanation of how senate weights can be derived and subsequently used in practice.

The chapter then turns to the standardisation of variables to have a mean of zero and a standard deviation of one (i.e. to calculate effect sizes / standardised mean differences). In international research, this involves the additional (and sometimes non-trivial) decision of whether national or international values of the mean and standard deviation should be used. We then conclude with a cautionary discussion of translation and the need to consult national as well as international documentation.

4.2. Which countries should I include in my analysis?

One of the most difficult decisions when beginning an international study is determining which countries to include in the analysis. It is, of course, possible to use data from every participating education system. However, with more than 90 countries now taking part in some ILSAs, including all of them may not always be the most effective choice. As discussed in chapter 8 on data visualisation, a larger number of countries means more data points, making findings harder to present clearly and communicate to readers. Interpretation may also become more difficult when attention is spread across so many comparators. Researchers may therefore prefer to focus on a carefully selected subset of participating countries that is most relevant to the question being investigated.

Some approaches I have used (or have seen used) include:

- Including all countries that participated in the ILSA (e.g. the widely reported “PISA rankings”).
- Comparing countries within a clearly defined group (e.g. the OECD or European Union).
- Including countries that share a common language, such as English-speaking countries (Anders et al., 2026).
- Comparing a focal country with the highest-performing education systems (e.g. Micklewright et al., 2014).
- Selecting countries according to a particular characteristic or factor of interest.
- Including countries primarily because the required data are available.

However, there is often no single correct answer when deciding which countries to include in an international analysis. In most cases, the choice involves a degree of judgment and depends on the purpose of the study. The most defensible approach is to link the selection of countries explicitly to the research question being addressed and to explain the rationale clearly. Returning to the discussion in Chapter 1, Section 1.5, the following broad guidance may help researchers make this decision in a transparent and theoretically informed way.

4.2.1 Benchmarking

Benchmarking often begins with a focal country whose performance we wish to compare with that of other education systems. The key question then becomes - which countries provide the most meaningful benchmarks? Consider Australia. One option would be to compare the outcome of interest there with every other country for which data are available. This would provide a comprehensive international picture, but not every comparison would necessarily be particularly informative. Differences between Australia and countries such as Ghana or Honduras may, for example, be difficult to interpret because their economic, institutional and educational contexts are so different. A broad international comparison can therefore be useful for describing where Australia sits globally, but this leads to a large number of data points (and including many comparators) of limited practical relevance.

An alternative is to focus benchmarking exercises on a more clearly defined group of countries (e.g. OECD members) which provide a more contextually comparable set of benchmarks. Even within this narrower group, however, some comparisons are likely to be more meaningful than others. For instance, countries such as England and New Zealand are likely to be particularly relevant benchmarks for Australia due to their linguistic, cultural, institutional and economic similarities. On the other hand, comparisons with other OECD nations such as Chile or Mexico may be less relevant (e.g. including differences in the percent of pupils included in the ILSA target population – see section 2.2 for discussion of this issue). Thus, while a large group of comparator countries is often chosen, it is often helpful to identify a smaller number of especially relevant benchmarks (explaining clearly why these were chosen).

4.2.2 Generalisability

When the aim is to assess whether a finding generalises across countries, the selection of cases should follow directly from the hypothesis being tested. Researchers should ask themselves, “in which countries do we expect this relationship to hold”? If the theoretical claim is that it applies universally, then all countries with suitable data should be included. If the expectation is limited to economically developed countries, restricting the analysis to OECD members may be more appropriate. Similarly, if the argument is that the relationship should emerge only where women enjoy a certain level of social, political or economic rights, countries could be selected using such a clearly defined external indicator. The key point is that it becomes much harder to challenge the chosen sample of countries when the inclusion criteria are explicitly connected to the underlying theory and research question.

4.2.3 Impact of institutional structures

As discussed in Section 1.5, when examining the possible effects of institutional structures (e.g. between-school tracking) the relevant variation occurs across countries. When the explanatory variable varies only between countries, its effect is identified primarily by the number of countries rather than by the number of pupils¹⁴. This creates an important limitation because only a finite number of countries participate in ILSAs, meaning that the country-level sample will never be especially large. At the same time, because the institutional feature of interest varies between national education systems, researchers generally want to maximise the amount of available cross-country variation. There is therefore a strong incentive to include all countries for which the necessary data are available.

A prominent example is Hanushek and Wößmann (2006); an influential study with more than 2,000 citations. Using data from PIRLS and PISA, the authors examined the relationship between between-school tracking and educational inequality across countries. However, their analysis included only around 20 countries. This raises an important question - is such a small country-level sample large enough to provide reliable estimates when the institutional variation of interest also occurs between countries? In my view, the study provides a useful illustration of the inherent limitations of cross-national research on education-system structures. Even when very large numbers of pupils are included, the effective sample size for institutional comparisons remains the number of countries, which is often modest.

4.2.4 Macro-forces

Much the same argument as presented for institutional structures applies when studying broader macro-level factors. As these factors vary mainly across countries, the available sample size and the amount of relevant variation will always be limited. The initial temptation is therefore to maximise the country sample by including all those where the relevant data are available. However, country selection should still be guided by substantive knowledge and theory rather than by sample size alone. Some countries may differ in ways that make the relationship under investigation less comparable, while the quality of the relevant data may also vary across settings.

¹⁴ The pupil and school sample sizes nevertheless affect the precision with which country outcomes and lower-level relationships are estimated.

For instance, consider again the Great Gatsby Curve discussed in section 1.5 and illustrated in Figure 1.5.1. This examines the association between income inequality and intergenerational mobility. Andrews and Leigh (2009) argued that former communist countries should be excluded from this type of analysis. Their justification was partly substantive; the generations being studied had experienced major political and social upheaval when growing up, meaning that the same relationship might not be expected to hold. It was also methodological; they argued that historical measures of income inequality for these countries have low levels of reliability. Both arguments are reasonable, making the decision to include or exclude these cases far from straightforward. In their analysis of the Great Gatsby Curve relationship, Jerrim and Macmillan (2015) therefore presented estimates both with and without five former communist countries included in their analysis. This reduced their sample size from 23 to 18 countries, with a simultaneous increase in the estimated correlation between inequality and intergenerational persistence from around 0.40 to 0.85. This example thus illustrates how country selection is often a delicate balance. It is also an example of where different choices can meaningfully change the magnitude of the associations produced.

4.3. Testing for statistical significance across countries.

Chapters 2 and 3 introduced the complex survey designs used in ILSAs and explained how replicate weights can be applied to estimate the sampling variance of a statistic of interest. Once the sampling variance has been estimated, it can be converted into a standard error (in most instances, this is simply the square root of the estimated sampling variance)¹⁵. This, in turn, allows one to construct confidence intervals around their estimates and to test whether apparent differences between countries are statistically significant.

To illustrate, suppose that we have estimated the socioeconomic gap in pupils' test scores in three countries, together with the appropriate standard errors. The results are presented in Table 4.3.1. The substantive question is whether the socioeconomic gaps is significantly different across the three countries.

¹⁵ We have not yet considered plausible values, which make an additional contribution to the standard error of ILSA test scores. This issue is discussed in detail in chapter 6 (section 6.5). Nevertheless, the methods described here continue to apply.

Table 4.3.1. Hypothetical estimates of the socio-economic gap in test scores and estimated standard errors across countries

	Socio-economic gap in test scores	Standard error
Country A	50	16
Country B	80	7
Country C	100	7

We begin with an informal approach. A common way to present results is a bar chart, where each bar shows the point estimate for a country and a thin horizontal line indicates the corresponding 95% confidence interval. For each country, the 95% confidence interval is then calculated using the standard formula¹⁶:

$$\beta \pm 1.96 * SE$$

Where:

β = The point estimate of interest (e.g. the magnitude of the socio-economic gap in this instance).

SE = The standard error.

Thus, using the figures provided in Table 4.3.1, one reaches 95% confidence intervals of:

Country A: 18.64 to 81.36

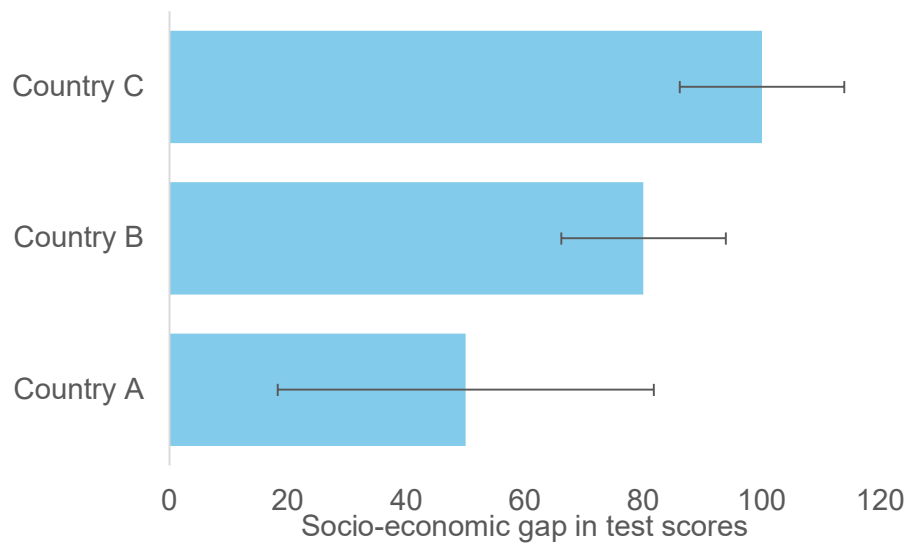
Country B: 66.28 to 93.72

Country C: 86.28 to 113.72

These are now plotted in Figure 4.3.1 below.

¹⁶ A critical value of 1.96 is often used because it corresponds to the large-sample normal approximation. With replication weights, however, the relevant design degrees of freedom are more appropriately based on the number of replicates minus 1 (e.g. with 80 replicate weights, there are actually 79 degrees of freedom, meaning the 5% critical value should be closer to 1.99). However, the exact degrees of freedom depends on the specifics of the survey design and how the replication weights are produced.

Figure 4.3.1. Plot of the socio-economic gap in each country with its 95% confidence interval



From such plots, it is often possible to infer whether differences between countries are statistically significant. Specifically:

- **Where the confidence intervals for two countries do not overlap.** The difference between the countries is statistically significant.
- **The confidence interval for either country includes the other country's point estimate.** The difference between the countries is not statistically significant.
- **The confidence intervals overlap, but neither includes the other country's point estimate.** The result is ambiguous and a formal test is needed to determine whether the difference is statistically significant.

Applying these rules to Figure 4.3.1 illustrates how confidence intervals can be used to assess differences between countries.

First, the confidence intervals for countries A and C do not overlap. This indicates that the difference in the estimated socio-economic gap between these two countries is statistically significant at the 5% level.

The comparison between countries A and B leads to a different conclusion. The upper limit of the confidence interval for country A is 81.36, which includes the point estimate for country B of 80. Consequently, the observed difference in the socio-economic gap between countries A and B is not statistically significant at the 5% level.

However, for countries B and C the conclusion cannot be determined from the graph alone. The upper confidence limit for country B is 93.72, while the lower confidence limit for country C is 86.28, meaning that the two confidence intervals overlap. Nevertheless, the upper confidence limit for country B remains below country C's point estimate of 100. Similarly, the lower confidence limit for country C (86.28) is above country B's point estimate (80). Thus, although the confidence intervals overlap, neither interval contains the other country's point estimate. A formal two-sample significance test is therefore required.

Such a two-sample t-test can be conducted using the following formula:

$$T - statistic = \frac{\beta_b - \beta_c}{\sqrt{SE_b^2 + SE_c^2}} \quad (\text{equation 4.3.1})$$

Where:

β_b = The point estimate for country B (e.g. the socio-economic gap in country B).

β_c = The point estimate for country C (e.g. the socio-economic gap in country C).

SE_b = The standard error of the estimate in country B.

SE_c = The standard error of the estimate in country C.

Plugging the figures from Table 4.3.1 in, we get:

$$\frac{\beta_b - \beta_c}{\sqrt{SE_b^2 + SE_c^2}} = \frac{80 - 100}{\sqrt{7^2 + 7^2}} = \frac{-20}{\sqrt{49 + 49}} = \frac{-20}{\sqrt{98}} = \frac{-20}{9.899} = -2.02031$$

As the absolute value of the calculated t-statistic (2.02) exceeds the critical value of 1.96, we conclude that the difference between Countries B and C is statistically significant at the 5% level. (See, however, section 4.5 for further discussion in relation to the issue of multiple hypothesis testing).

Although the example above focuses on differences between countries, a similar same approach can also be used to test whether estimates differ significantly across two survey cycles. (Though readers are also directed to the discussion of “link errors” in chapter 6 - section 6.8 - for a further technical detail).

4.4. A note about including a country level within a multi-level model

We note that some researchers have analysed ILSA data using a three-level multilevel model, with individuals at level 1, schools at level 2 and countries at level 3. Such models are often used when the substantive interest lies in country-level characteristics, such as features of education systems, national policy contexts or broader macro-economic factors (e.g. income inequality).

Yet such an approach introduces additional methodological challenges. In particular, modelling country-specific effects as random is most defensible when the countries included can be regarded as exchangeable units drawn from a broader population. This assumption is difficult to justify in ILSAs, where participating countries and education systems are not randomly sampled but choose to participate.

A further concern is the limited number of country-level units available for analysis. This creates particular difficulties for estimating country-level effects. Bryan and Jenkins (2016) show that, even when individual-level sample sizes are large, a small number of countries can lead to unreliable estimates of country-level associations. Estimated effects of country-level variables may also have large standard errors and may be sensitive to the inclusion/exclusion of particular countries or country-level controls. The small number of countries also limits the number of macro-level covariates that can be included without overfitting the model.

These concerns are even more acute for cross-level interactions involving country-level variation. For example, researchers may wish to examine whether the association between an individual-level characteristic (e.g. socio-economic status or immigrant background) and achievement varies according to national policy or institutional arrangements. Such questions require sufficient information at the country level to estimate differences in slopes reliably. Simulation evidence suggests that estimates and confidence intervals can be too narrow when the number of countries is small, especially in models including cross-level interactions (Stegmueller, 2013), though others have argued against this concern (Elff et al., 2021). A danger, therefore, is that specifying country as a third level in a multilevel model may give the appearance of a sophisticated treatment of the data structure, while masking the fragility of the country-level estimates on which the substantive conclusions depend.

Consequently, pooled ILSA multilevel models with countries at the highest level should be used cautiously (and – in my view – sparingly). If analysts do choose to estimate such a model, it is vital they state the intended population of inference, report the number of countries included, keep country-level specifications parsimonious, use appropriate small-sample inference and examine sensitivity to influential countries.

4.5. Multiple hypothesis testing

When conducting international comparisons, researchers often include many countries (and potentially several time points) in their analysis. Recall the discussion in chapter 1 (section 1.5.1) concerning the use of ILSAs to benchmark a focal country against its international peers. In practice, the focal country is rarely compared with only one other education system. Instead,

comparisons are usually made with a broad group of countries. Yet the more comparisons we conduct, the greater the probability that at least one will produce a statistically significant result purely by “chance” (random variation due to sampling/measurement error). This issue is commonly known as the multiple-comparisons (or multiple-testing) problem.

Return to Table 4.3.1, for example. Suppose that Country C is our focal country and that we wish to compare it with Countries A and B. We are therefore testing two hypotheses; (1) whether the socioeconomic achievement gap in Country C differs from that in Country A; (2) whether the socioeconomic achievement gap in Country C differs from that in Country B. Assuming a 5% significance level and independent tests¹⁷, the probability of obtaining at least one false positive result (i.e. declaring at least one statistically significant difference when in reality there is no difference) can be calculated as:

$$1 - (1 - 0.05)^2 = 0.0975$$

In other words, there would be around a 9.75% chance of incorrectly concluding that the socioeconomic achievement gap in at least one of Countries A or B differs from that in Country C.

The more comparator countries included in the analysis, the greater this risk becomes. For example, if we test 100 hypotheses (e.g. compare our focal country with 100 others) which were all independent, then there is around a 99.4% chance that at least one false positive will occur¹⁸.

A common response to the multiple-testing problem is to apply a correction that uses a more conservative threshold for statistical significance. One of the most widely used approaches is the Bonferroni correction (Haynes, 2013). Suppose, for example, that we conduct two significance tests. Rather than declaring a result statistically significant whenever the p-value is below 0.05, we divide this conventional threshold by the number of hypotheses being tested. The adjusted threshold would therefore be $0.05/2 = 0.025$. A result would then only be considered statistically significant if the p-value falls below 0.025.

As the number of significance tests increases (i.e. as more comparator countries are included) the Bonferroni correction becomes increasingly conservative. For example, with 100 comparator countries, a difference would be declared statistically significant only if the p-value falls below

¹⁷ In this example, the tests are not actually independent, as both involve country C. This is likely to make the Bonferroni correction even more conservative. Moreover, with dependence between the tests, the probability of finding at least one difference in the example above may differ slightly from 9.75% (were this taken into account).

¹⁸ With a correlation between tests (e.g. testing one focal country against several others), this percentage may differ (depending on the covariance between the tests). We abstract from this here for simplicity.

0.0005. This is a widely recognised limitation of the approach. Although the correction reduces the risk of false positives, it can make genuine differences harder to detect. In other words, it can increase the likelihood of false negatives.

To see this in action, return once more to the example we presented in section 4.3. There, we tested three hypotheses:

- Whether country A differs from country B
- Whether country A differs from country C
- Whether country B differs from country C

If we adjust for multiple hypothesis testing using a Bonferroni correction, p-values for the differences between countries must now fall below $0.05/3 = 0.0167$ to be declared statistically significant.

The t-statistic calculated for the comparison between countries B and C at the end of section 4.3 can be converted into a p-value using Excel's T.DIST.2T() function¹⁹. In this example, the p-value for the difference between Countries B and C is approximately 0.045. Without any correction for multiple hypothesis testing, this result would be declared statistically significant because the p-value falls below the conventional 0.05 threshold. However, with three comparisons, the adjusted significance threshold is ($0.05/3 = 0.0167$). Because 0.045 is greater than 0.0167, the difference between Countries B and C would no longer be considered statistically significant after adjusting for multiple testing.

More generally:

- If a difference is not statistically significant before a Bonferroni correction is applied, it will not become significant afterwards.
- If the p-value is very small (i.e. the t-statistic very large) and only a small number of comparisons are being made, applying a Bonferroni correction is unlikely to change the conclusion of whether the difference is statistically significant or not.
- Results that are only marginally significant before correcting for multiple hypothesis testing are unlikely to remain statistically significant after the correction is applied. This is particularly likely to be true when the total number of comparisons being made is large.

¹⁹ The same calculation can also be performed easily in most statistical software packages and various online calculators.

Despite this concern, formal corrections for multiple hypothesis testing remain relatively uncommon in analyses of ILSAs. One reason is that many ILSAs now include a large number of countries, leading to a huge number of possible pairwise comparisons. Applying a Bonferroni correction would hence produce an extremely stringent significance threshold. Although this would reduce the likelihood of false positives, it could also generate many false negatives by making genuine differences difficult to detect. Nevertheless, the underlying problem does not disappear. Even when no correction is used, researchers should still recognise that some statistically significant findings may arise purely by chance when many comparisons are being conducted.

4.6. House versus senate averages

When analysing ILSAs, researchers often want to report not only estimates for individual countries but also summary statistics for a wider international group, such as the OECD average. However, two different types of international statistic are commonly used in the literature, with subtle but important differences in their calculation and interpretation. These are known as “house” and “senate” statistics, terminology borrowed from voting arrangements in the United States political system. A house statistic gives greater weight to countries with larger target populations, meaning that pupils from more populous education systems contribute more to the international estimate. Senate statistics, on the other hand, gives each country equal weight (regardless of its population size). The distinction therefore depends on the quantity of interest; whether the aim is to describe the average pupil across the combined international population or the average performance of the participating countries. Although the two approaches may produce similar results, they can differ meaningfully when large and small countries have systematically different outcomes. A summary is therefore:

- House statistics. International estimates are weighted by each country’s target population size. More populous countries therefore receive greater influence when the overall statistic is calculated.
- Senate statistics. Each country receives equal weight, regardless of its target population size. Every participating country therefore contributes equally to the international statistic.

Table 4.6.1 provides a simple example illustrating the difference between house and senate statistics using hypothetical data from five countries. Our aim is to calculate both types of international average for the percentage of young people attending private schools. To calculate

the house average, we follow the weighting procedure introduced previously in chapter 3 section 3.2:

$$\frac{(20,000 * 0.1 + 15,000 * 0.05 + 25,000 * 0.05 + 20,000 * 0.15 + 250,000 * 0.4)}{(20,000 + 15,000 + 25,000 + 20,000 + 250,000)} = \frac{107,000}{330,000} = 32\%$$

In contrast, the senate average can be simply calculated as:

$$\frac{(0.1 + 0.05 + 0.05 + 0.15 + 0.4)}{5} = 15\%$$

There is therefore a substantial difference between the house and senate averages in this hypothetical example. Moreover, as the calculation illustrates, another way of thinking about the difference between house and senate averages is whether we calculate weighted or unweighted statistics at the country level.

Table 4.6.1. Hypothetical data on the percentage of pupils attending private school in five countries

Country	Population size	% attending private school
A	20000	10%
B	15000	5%
C	25000	5%
D	20000	15%
E	250000	40%

The hypothetical example in Table 4.6.1 also helps illustrate when house and senate international statistics are likely to differ. In particular, the gap will tend to be larger when:

- A small number of very populous countries are included alongside many smaller countries.
- Those large countries are also outliers on the statistic being examined.

This pattern can be seen clearly in Table 4.6.1. Country E has a target population of 250,000. This is approximately three times larger than the combined population of the other four countries. It also has a much higher percentage of children attending private school (40% compared to between 5%-15% elsewhere). For the house international percentage, Country E receives substantially greater weight because of its much larger population. Its unusually high private-school participation rate therefore pulls the house international percentage upwards. On the other hand, Country E contributes no more to the senate international percentage than any other

country. The difference between the two is therefore driven by the combination of Country E's large population and high percentage of pupils attending a private school.

Choosing between house and senate statistics is often a matter of judgment and should be guided by the research question. Suppose, for example, that the topic of interest is private schooling in South America. A senate average would support statements such as "in the average South American country, X percent of children are enrolled in private schools." The focus is thus on the typical country within the region (South America). In contrast, a house average would support statements such as "Across South America, X percent of children are enrolled in private schools". In this case, the focus is on the combined target population living in the region (South America).

4.7. How to create and use senate weights

ILSAs are usually supplied with "house" versions of both the final individual sampling weight and the replication weights. Within each country, the sum of the individual sampling weights therefore approximates the size of the target population. For example, if the hypothetical data in Table 4.6.1 came from an ILSA, the values in the "population size" column would be obtained by summing the supplied individual weights within each country. To calculate a house average, together with its associated estimate of sampling uncertainty, analysts can therefore use the sampling and replication weights supplied following the procedures described in Chapter 3. Indeed, the "with weights" calculation presented in Table 3.2.2 is in fact the house average for the United Kingdom.

International senate statistics are straightforward to calculate for some quantities, such as international means (e.g. the OECD average). As illustrated in Table 4.6.1, one can first estimate the statistic separately for each country, before then simply averaging these across countries.

However, the matter may become more complicated when:

- The statistic of interest is something other than the mean.
- The sampling variation of the international senate statistic must also be estimated (e.g. to formally test whether the estimate for a country is significantly different from the international estimate).

For these reasons, analysts may need to convert the "house" versions of the final individual and replication weights supplied in ILSAs into corresponding "senate" weights. This conversion

essentially involves rescaling the house weights separately within each country so that national target population size no longer determines its influence on the combined estimate.

The relevant scale factor for each country can be calculated by:

$$\text{Scaling factor} = \frac{\text{Constant}}{\text{Sum of house weight}}$$

Where:

Scaling factor = The value an analyst will need to multiply the supplied “house” versions of the weights by to convert them into equivalent senate weights.

Constant = A constant value one wishes the sum of the senate weights in each country to equal (typically set to 500 or 1,000).

Sum of house weights = The total value of the house weights when summed across each observation in the ILSA for the country (roughly equal to the target population size).

The supplied house weights are then simply multiplied by this scaling factor to convert them into their equivalent senate versions.

Table 4.7.1 provides a simple numerical example using hypothetical data from two countries (Y and Z). The dataset contains information for ten individuals in each country, including the score each person achieved on a test. It also includes the final individual sampling weight in its original “house” form (labelled “house weight”). Within each country, these weights sum approximately to the size of the target population represented by the sample. The table also presents a corresponding “senate weight” for every individual. The calculations below illustrate how these senate weights are produced by rescaling the supplied house weights separately within countries Y and Z.

Table 4.7.1. Hypothetical individual level data from two ILSAs

Country	ID	Score	House weight	Senate weight
Y	1	503	20	76.9
Y	2	513	40	153.8
Y	3	489	10	38.5
Y	4	538	10	38.5
Y	5	446	30	115.4
Y	6	489	30	115.4
Y	7	539	50	192.3
Y	8	500	10	38.5
Y	9	486	50	192.3

Y	10	532	10	38.5
Z	11	543	6	76.9
Z	12	563	12	153.8
Z	13	534	3	38.5
Z	14	564	3	38.5
Z	15	476	9	115.4
Z	16	528	9	115.4
Z	17	584	15	192.3
Z	18	525	3	38.5
Z	19	523	15	192.3
Z	20	579	3	38.5

Let us begin by calculating the scaling factor for country Y. First, sum the supplied house weights across all individuals in the country:

$$20 + 40 + 10 + 10 + 30 + 30 + 50 + 10 + 50 + 10 = 260$$

Now we can get the scale factor as:

$$1000/260 = 3.846$$

Note that a value of 1,000 is used here, but any value could be chosen (the key point is that the same value must be used for each country).

We then multiply everyone's house weight in country Y by the scaling factor of 3.846 to obtain the corresponding senate weight.

If we do the same for country Z, we get the total value of the weights as:

$$6 + 12 + 3 + 3 + 9 + 9 + 15 + 3 + 15 + 3 = 78$$

Which gives a scale factor of:

$$1000/78 = 12.821$$

Then, in country Z, we multiply the "house weight" by this scale factor of 12.821 to produce the corresponding senate weight.

At the end, we can check we have done this process correctly by verifying that the sum of the weights in each country is equal to the same value.

So, in country Y:

$$76.9 + 153.8 + 38.5 + 38.5 + 115.4 + 115.4 + 192.3 + 38.5 + 192.3 + 38.5 = 1000$$

And in country Z:

$$76.9 + 153.8 + 38.5 + 38.5 + 115.4 + 115.4 + 192.3 + 38.5 + 192.3 + 38.5 = 1000$$

The same rescaling process must then be repeated for every replicate weight in the dataset, converting each from its house to its senate form²⁰. This is important to ensure that the sampling error is estimated correctly.

A further important point to note is that house weights should be converted into senate weights only after all restrictions to the analytic sample have been applied. Suppose, for example, that the intended analysis focuses only on children whose parents have a low level of education. The sample should first be restricted to this group, with the house and replicate weights then rescaled into their senate versions. The same holds true with respect to instances where observations with missing data are removed through listwise deletion (noting this is not considered best practice in the handling of missing data - Huang & Keller, 2025). If this is not done (i.e. analytic sample selection taking place after rescaling of the weights) then the senate weights may no longer sum to the same total within each country. This, in turn, would mean that each country would not necessarily contribute equally to the final senate estimate of the international statistic.

A final point to note is with respect to the use of t-tests to compare an individual country's estimate with an international average. This is due to the fact that the country being tested will have also contributed to the international average. The formula for the t-statistic comparing individual countries to senate international statistics thus becomes (Foy & LaRoche, 2020):

$$T - statistic = \frac{\beta_k - \beta_{int}}{\sqrt{SE_{int}^2 + \frac{N-2}{N} SE_k^2}}$$

Where:

β_k = Estimate of the statistic of interest in country k.

β_{int} = Estimate of the statistic of interest across countries – i.e. the international statistic (e.g. the OECD average).

SE_k = Standard error of the estimate in country k.

²⁰ OECD (2009, p. 167-169) apply the same scaling factor (calculated from the final individual weight) to the final weight and each replicate weight. It is also possible to calculate a separate scaling factor for each replication weight and use these (so that the sum of each replicate weight within a country will also equal the same value). It is not clear which of these two approaches should be preferred – though my hunch is that the difference between using one over the other will be trivial.

SE_{int} = Standard error of the international statistic (e.g. standard error of the OECD average).

N number of countries included in the international statistic.

Otherwise, the approaches and issues discussed throughout this chapter (in the context of comparing estimates across two individual countries) remain unchanged.

4.8. National versus international standardisation

In education and the broader social sciences, variables are often standardised to have a mean of zero and a standard deviation of one. This is usually done to make results easier to interpret and compare. For example, a difference of 100 points on a test is difficult to interpret in isolation; whether this is a large or small difference depends entirely on the scale. Standardisation places variables on a common metric (standard deviation units) which provides a more intuitive sense of magnitude. It also makes it easier to compare results across outcomes measured using different scales. Indeed, even when variables are not standardised directly, findings are often discussed in terms of effect sizes, standardised mean differences or standardised regression coefficients.

The usual way to standardise a variable is to subtract the mean and divide by the standard deviation:

$$Z_i = \frac{(\hat{X}_i - \hat{\mu})}{\hat{\sigma}} \quad (\text{Equation 4.8.1})$$

Where:

Z_i = The continuous variable for individual i in its standardised form.

$\hat{X}_i$ = The continuous variable for individual i in its original metric.

$\hat{\mu}$ = The mean of the continuous variable in its original metric.

$\hat{\sigma}$ = The standard deviation of the continuous variable in its original metric.

The added complication in international comparative research is deciding whether to use national or international values for the mean and standard deviation.

National standardisation uses each country's own mean and standard deviation in the calculation. The variable is therefore standardised separately within each country, producing a national distribution with a mean of zero and a standard deviation of one. This approach focuses on relative differences within countries. In doing so it abstracts from cross-country differences in both the mean and variation of the variable being standardised.

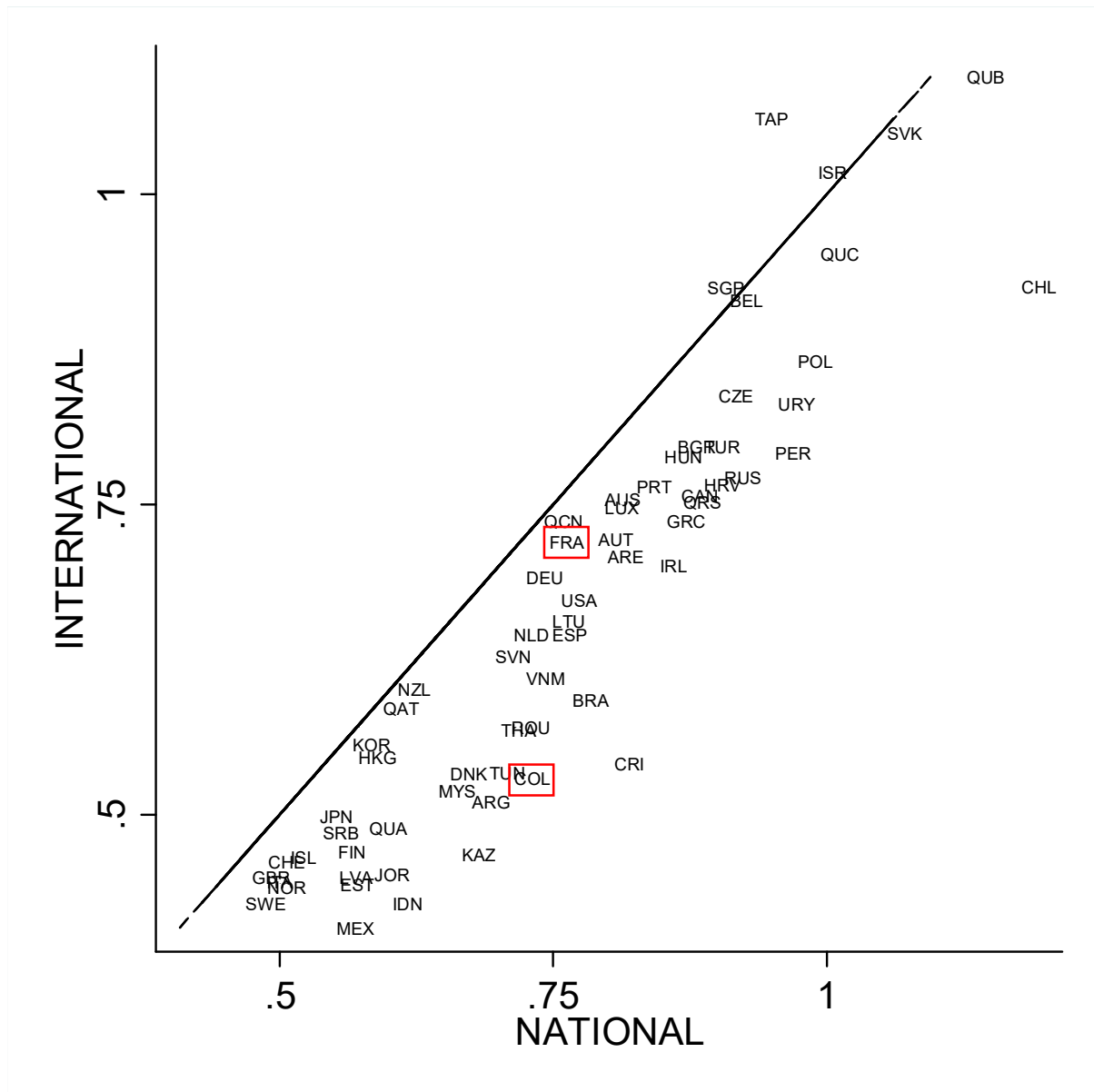
On the other hand, international standardisation uses a common international mean and standard deviation in equation 4.8.1. The data from all countries are therefore standardised simultaneously via a single application of the formula. Unlike national standardisation, this does not force each country's distribution to have a mean of zero and a standard deviation of one. The data hence continue to include cross-national differences in the mean and standard deviation of the variable being standardised. Cross-country differences in the variation are thus still incorporated into subsequent estimates using the standardised variable.

Recall that standardisation is often used so that results can be interpreted in terms of how a one-standard-deviation change in a variable is associated with an outcome (which may have also been standardised). The choice between national and international standardisation therefore depends on whether the analysis should abstract from cross-country differences in the variability of the measure. Unfortunately, there is no universal rule about which approach should be preferred. Rather, the appropriate choice depends on the research question and the interpretation one wishes to place on the resulting estimates.

The choice between national and international standardisation can make a non-trivial difference to the results. Figure 4.8.1 illustrates this using PISA data on the association between parental education and pupils' test scores across countries. The horizontal axis presents estimates based on nationally standardised achievement scores. These figures therefore abstract from cross-country differences in the overall dispersion of test performance. The vertical axis meanwhile presents estimates based on internationally standardised scores. The diagonal 45-degree line then illustrates where the figures are equal.

There is clearly a strong correlation between the two sets of results. However, almost every country lies below the 45-degree line, indicating that the estimated parental education gap is usually larger when test scores are standardised nationally rather than internationally. The intuition is that a national standard deviation captures only variation in achievement within a country, whereas an international standard deviation reflects both within-country and between-country variation. The international standard deviation is therefore typically larger. Because the standard deviation enters the denominator of the standardisation formula, dividing by the smaller national value produces a larger standardised estimate of the parental education gap in test scores.

Figure 4.8.1. Comparison of parental education gaps in test scores across countries using national versus international standardisation of the test scores.



Notes: Figures refer to the difference in standardised test scores between children from low and high parental education backgrounds. Figures on the horizontal axis are where test scores have been nationally standardised. Figures on the vertical axis are where they have been internationally standardised.

A further notable point is that results for some individual countries change substantially depending on the standardisation approach used. Consider France and Colombia (highlighted in red rectangles in Figure 4.8.1) for example. When test scores are nationally standardised, the parental education gap is around 0.75 standard deviations in both countries. Under international standardisation, however, the estimated gap in Colombia falls to around 0.50 standard deviations, while remaining close to 0.75 in France. This is due to national standardisation

abstracting from Colombia’s relatively low variation in PISA test scores, while the international standardisation does not. Consequently, the conclusion drawn about whether France and Colombia have similarly sized parental education gaps depends heavily on the standardisation approach used.

Similar choices arise at other intermediate stages of analysis of international data. One example is the practice of dividing a continuous variable into categories (e.g. quartiles or quintiles). In ILSAs, researchers must decide whether these groups should be constructed separately within each country or across an international sample. For instance, creating socio-economic quartiles from a socio-economic status index within individual countries allows comparisons between pupils in the most and least advantaged quarter of the population within every country. On the other hand, creating socio-economic quartiles using a pooled international sample means that some countries will contain a smaller/larger proportion of “advantaged” and “disadvantaged” pupils than others. Again, neither approach is inherently better than the other. The appropriate choice depends on whether the research question concerns relative position within national societies or absolute position across a wider international population.

As with national versus international standardisation, this choice changes the interpretation of the estimates and can produce nontrivial differences in the results (as per Figure 4.8.1). Applied researchers should therefore state clearly which approach they have used and explain why it is appropriate for their research question. Ideally, results should also be presented using more than one approach to show how sensitive the substantive conclusions are to the analytical choice made.

4.9. Test your knowledge questions

4.9.1 Task 1.

Table 4.9.1.1 provides average test scores for five countries along with the estimated standard error. Your interest is in comparing the results for country A to the other countries in the dataset. The data has been collected using a random probability sample design, with separate samples drawn for each of the five countries. The sample size in each country was 10,000 individuals.

Table 4.9.1.1. Average test scores across five countries

Country	Mean	Standard error
A	535	5
B	525	5
C	520	5
D	515	5
E	510	5

Task 1. Question 1. Please calculate (either by hand or using Excel) the nations which country A are significantly different from at the 5% level. Please show your working. (Hint: You can convert a t-statistic to a p-value in Excel using the T.DIST.2T function).

Task 1. Question 2. Please explain what is meant by a “Bonferroni correction” and situations in which it might be applied. How would your answers to question 1 above change if you make a Bonferroni correction?

4.9.2 Task 2.

Figure 4.9.2.1 below illustrates the percentage of pupils that repeat a grade based on sample survey data. The thin black line running through the centre of each bar reflects the 95% confidence interval. Based on the information presented in the graph alone, you need to report whether differences between each country are statistically different at the 5% level or not, or if further information is needed. Please use Table 4.9.2.1 to report your answers for each pairwise comparison, explaining your reasoning for each.

Figure 4.9.2.1. Hypothetical data on the percent of pupils that repeat a grade

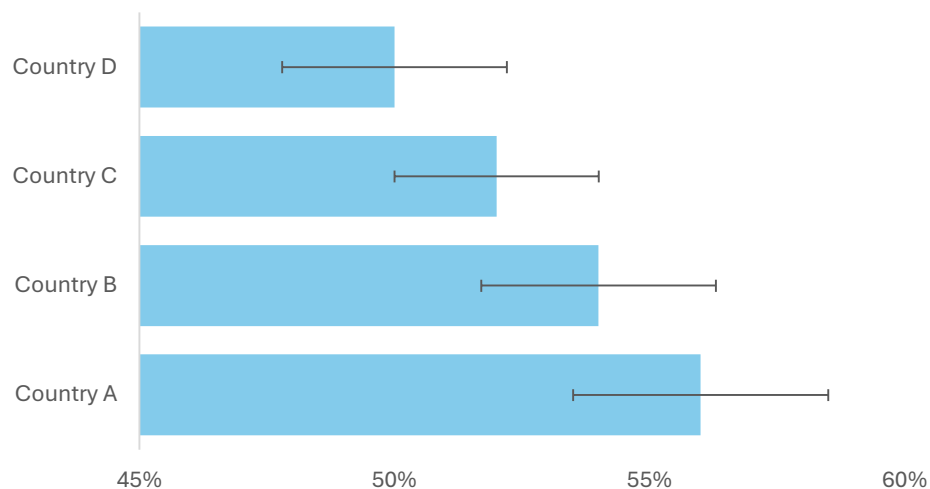


Table 4.9.2.1. Self-completion table to report whether differences between pairwise comparisons of countries are statistically significant

Countries to compare	Difference significant or not	Why? Please state your reasoning here.
A vs B		
A vs C		
A vs D		
B vs C		
B vs D		
C vs D		

4.9.3. Task 3

Table 4.9.3.1 presents average test scores for three countries, each of which has a sample of 10 pupils. The columns refer to the following:

- Country = Country in which the pupil studies.
- ID = The pupil ID.
- Score = The score the pupil achieved on the test
- Weight = The final pupil weight in the sample.
- REP_1 to REP_5 = Five replication weights that have been created using the BRR (Fay's) approach.

Table 4.9.3.1. Hypothetical pupil achievement data across countries with weights and BRR replication weights.

Country	ID	Score	Weight	REP_1	REP_2	REP_3	REP_4	REP_5
England	1	503	20	60	20	0	40	0
England	2	513	40	0	80	0	0	0
England	3	489	10	10	0	10	20	0
England	4	538	10	0	0	10	0	0
England	5	446	30	60	0	30	60	0
England	6	489	30	0	60	0	0	60
England	7	539	50	0	100	0	100	100
England	8	500	10	20	30	20	0	20
England	9	486	50	50	0	150	0	100
England	10	532	10	0	0	0	20	20
Wales	11	453	2	4	2	0	2	0
Wales	12	482	4	0	4	0	0	0
Wales	13	451	1	1	0	1	3	0
Wales	14	506	1	0	0	1	0	0
Wales	15	409	3	9	0	3	3	0
Wales	16	452	3	0	9	0	0	6
Wales	17	507	5	0	10	0	15	5
Wales	18	451	1	1	3	3	0	3
Wales	19	449	5	5	0	10	0	5
Wales	20	482	1	0	0	0	2	3
Scotland	21	543	6	18	12	0	12	0
Scotland	22	563	12	0	12	0	0	0
Scotland	23	534	3	3	0	3	6	0
Scotland	24	564	3	0	0	6	0	0
Scotland	25	476	9	9	0	18	18	0
Scotland	26	528	9	0	9	0	0	9
Scotland	27	584	15	0	30	0	15	15
Scotland	28	525	3	9	3	6	0	6
Scotland	29	523	15	15	0	45	0	15
Scotland	30	579	3	0	0	0	3	6

Task 3. Question 1. Please calculate the “house” average of the test scores in England. How does this compare to the “house” average of the test scores in Scotland and Wales?

Task 3. Question 2. What is the house mean of the test scores when you pool the data across the three countries into a single sample?

Task 3. Question 3. Please derive an estimate of the sampling error of the house average for England. Then please do the same for Scotland, and then for Wales.

Task 3. Question 4. Please derive an estimate of the sampling error of the house average for the data pooled across all three countries.

Task 3. Question 5. Please calculate the “senate” average of the test scores in England. How does this compare to the “senate” average of the test scores in Scotland and Wales?

Task 3. Question 6. What is the senate mean of the test scores when you pool the data across the three countries into a single sample? Can you give the intuition behind why this is higher or lower than the value obtained in task 3 question 2?

Task 3. Question 7. Please derive an estimate of the sampling error of the senate average using the data pooled across the three countries into a single sample.

Task 3. Question 8. Please repeat your analysis in the steps above but now using either STATA, R or other statistical software to produce the results on your behalf.

References

Anders, J., Jerrim, J., Ladrón de Guevara Rodríguez, M., & Marcenaro-Gutierrez, O. D. (2026). The rise in teenagers skipping school across English-speaking countries following the COVID-19 pandemic: Evidence from PISA. *Educational Assessment, Evaluation and Accountability*, 38, 33–59. <https://doi.org/10.1007/s11092-025-09470-z>

Andrews, D., & Leigh, A. (2009). More inequality, less social mobility. *Applied Economics Letters*, 16(15), 1489–1492. <https://doi.org/10.1080/13504850701720197>

Bryan, M. L., & Jenkins, S. P. (2016). Multilevel modelling of country effects: A cautionary tale. *European Sociological Review*, 32(1), 3–22. <https://doi.org/10.1093/esr/jcv059>

Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2021). Multilevel Analysis with Few Clusters: Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference. *British Journal of Political Science*, 51(1), 412–426. doi:10.1017/S0007123419000097

Foy, P., & LaRoche, S. (2020). Estimating standard errors in the TIMSS 2019 results. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 14.1–14.60). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch14-estimating-standard-errors.pdf

Haynes, W. (2013). Bonferroni Correction. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_1213

Huang, F., & Keller, B. (2025). Working with missing data in large-scale assessments. *Large-scale Assessments in Education*, 13, Article 13. <https://doi.org/10.1186/s40536-025-00248-9>

Jerrim, J., & Macmillan, L. (2015). Income inequality, intergenerational mobility, and the Great Gatsby Curve: Is education the key? *Social Forces*, 94(2), 505–533. <https://doi.org/10.1093/sf/sov075>

Micklewright, J., Jerrim, J., Vignoles, A., Jenkins, A., Allen, R., Ilie, S., Bellarbre, E., Barrera, F., & Hein, C. (2014). *Teachers in England's secondary schools: Evidence from TALIS 2013* (Research Report No. DFE-RR302). Department for Education. https://assets.publishing.service.gov.uk/media/5a7eebbe5274a2e87db2cfd/RR302_-_TALIS_report_NC.pdf

Stegmueller, D. (2013), How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science*, 57: 748-761. <https://doi.org/10.1111/ajps.12001>

5. A crash course introduction to classical test and item-response theory

5.1 Chapter introduction.

This chapter introduces the main ideas behind how responses to test questions are converted into measures of achievement. While readers do not need to understand this topic in great depth in their work with ILSA, having some basic background knowledge and understanding is helpful. This chapter has therefore been designed to keep things minimal and as non-technical as possible, providing the intuition behind what are quite complex ideas and techniques. Readers interested in learning more about the issues raised in this chapter are directed to the excellent work of Wu et al. (2017).

The chapter begins with a brief overview of classical test theory (CTT). This is a relatively simple and intuitive approach that has been widely used in educational measurement. However, we also discuss some of the limitations of this approach, motivating why more advanced methods are used in ILSAs. This leads us to introduce readers to Item Response Theory (IRT). Rather than a detailed, technical discussion we focus on explaining IRT at a basic, intuitive level. This includes an initial discussion of a “Rasch” (one-parameter) IRT model, before moving onto the issues of item difficulty, discrimination and guessing in the context of two and three parameter models. The chapter then concludes with a discussion of differential item functioning (“DIF”), its links to measurement comparability, methods of detecting DIF and ultimately what should be done about it.

5.2 Classical test theory (CTT) and raw test scores.

We begin with a discussion of classical test theory because of its conceptual simplicity and the fact the key ideas of how test scores are constructed when using it are likely to be familiar to many readers. Specifically, CTT can be regarded as a test-centric approach to measuring individuals’ knowledge and skills as it focuses primarily on total test scores, with the resulting estimates dependent on the particular test administered (Hambleton & Jones, 1993). The focus of CTT is often on developing and administering a particular test and then measuring individual performance mainly through raw or total test scores (e.g. number or percentage of questions answered correctly). Under CTT, many of the key quantities of interest are tied to the specific assessment and sample used. For instance, a pupil’s observed performance in the subject being tested is represented by their total test score. Likewise, the characteristics of the test questions (e.g. their difficulty) are based on the particular sample that took the test.

The “raw scores” that are often derived under CTT will be familiar to readers. For instance, suppose an individual takes a test with 50 questions, each marked as correct or incorrect. If the test taker answers 25 questions correctly, they will receive a raw test score of 50%. This score then becomes our observed measure of their achievement.

Of course, no test measure is perfectly reliable. All tests are, to some extent, affected by measurement error. In other words, if the same individual took another similar version of the test on a different day, their score would probably not be exactly the same. For instance, in the example above the test taker might score 23 or 28 out of 50, rather than 25. Such random measurement error may arise for several reasons, including the particular set of questions included on the test, chance variation in how well those questions match the pupil’s knowledge or fluctuations in how the pupil feels and performs on the day of the assessment. Within CTT, there is also a close relationship between test length and reliability. All else equal, asking pupils more test questions will usually increase reliability and make an average or percentage score more precise (Ellis & Sijtsma, 2024).

Thus, within CTT, a distinction is made between each individual’s observed score (X) and their true score (T). The observed score is the score they receive on a particular administration of the test. On the other hand, their true score is the score they would achieve on average if they were able to sit many similar versions of the test under similar conditions. The difference between these two quantities (the observed score and the true score) is treated as random measurement error. This relationship can therefore be expressed using the simple formula:

$$X = T + \varepsilon \quad (5.2.1)$$

Where:

X = Individual's observed score, expressed on a stated scale. (In this example, the number of questions answered correctly).

T = Individual's "true" score (i.e. the score they would achieve - on average - if they were able to sit similar versions of the test many times).

ε = Random error.

Because measurement is less than perfectly reliable, there is uncertainty around an individual’s true score when it is estimated using a single observed test score. Within CTT, this uncertainty is captured by the standard error of measurement. The standard error of measurement is equivalent to the standard deviation of the random error term (i.e. the standard deviation of ε). In

other words, it describes the typical amount by which an observed score might differ from the individual's true score because of measurement error. It can be estimated using the following formula:

$$\text{Standard error of measurement} = \sqrt{(1 - \rho) \cdot \text{Var}(X)}$$

Where:

ρ = The reliability of the test.

$\text{Var}(X)$ = The variance of observed scores.

For instance, suppose that in our example 50-question test the variance of observed scores from one administration was 11 raw score points and the reliability of the assessment was 0.8. We would estimate the standard error of measurement to be close to 1.5. For our hypothetical individual who scored 25 out of 50, we could then construct an approximate 95% confidence interval for their true score, running from around 22 to 28 out of 50 (i.e. across repeated test administrations, about 95% of such intervals would contain the corresponding true score). Note that in CTT the standard error of measurement is usually treated as the same for all individuals who took the test (i.e. regardless of whether they are at the bottom, middle or top of the achievement distribution). Thus, for all test takers, the confidence interval would have the same width (e.g. approximately six points in our exemplar test).

It is also possible to estimate question-level CTT statistics from each administration of a test. Here, we focus on two properties of particular interest - item difficulty and item discrimination. Both will be developed further in the discussion of IRT in the sections that follow.

Item difficulty refers to how challenging a question was for test takers. Within CTT, it is typically measured as the proportion of test takers who answered the question correctly within a specific administration of the test. For example, suppose our hypothetical 50-question test was taken by 100 children. Question 1 was answered correctly by 95 children, while Question 50 was answered correctly by 20 children. The CTT item difficulty statistic would therefore be 0.95 for Question 1 and 0.20 for Question 50. Values on this statistic range from zero to one, with lower values indicating more difficult questions. Otherwise, this is an intuitive measure of question difficulty; it is simply the proportion of children who answered the question correctly.

In contrast, item discrimination essentially provides an indication of how well a question is measuring the construct we are trying to capture (and thus its ability to distinguish between more

and less able individuals on this construct). For example, on our hypothetical 50-question maths test, item discrimination captures whether pupils who perform well overall on the test are more likely to answer a specific question correctly. In CTT, item discrimination can be measured as the correlation between responses to a test question and the total test score²¹. Higher values indicate stronger discrimination and are usually taken as evidence that the item is measuring something similar to the rest of the test. In contrast, a low correlation typically indicates the question is measuring something different, that there is an administrative/scoring error or that the wording is confusing test takers. For instance, just say our maths test included the question “my favourite food is cheese”. Responses to this question are likely to correlate poorly with the overall maths test score (i.e. it will have a low “item discrimination” value) as it is measuring something different from the rest of the test. Questions with low discrimination values thus require careful scrutiny. In the extreme, one may wish to exclude them when calculating a test taker’s overall raw score.

From the discussion above, CTT may appear to be a powerful approach, despite its simplicity. It can not only be used to generate total test scores for individual test takers, but also construct confidence intervals around them, estimate the reliability of the test and examine the properties of individual test items (at least within the sample that took that particular administration of the test). And indeed, CTT remains widely used, including before (or alongside) more complex approaches such as item response theory. However, CTT also has important limitations which explain why it is not the central approach used in the scaling of ILSA data. It is, in particular, worth highlighting the following four issues with CTT:

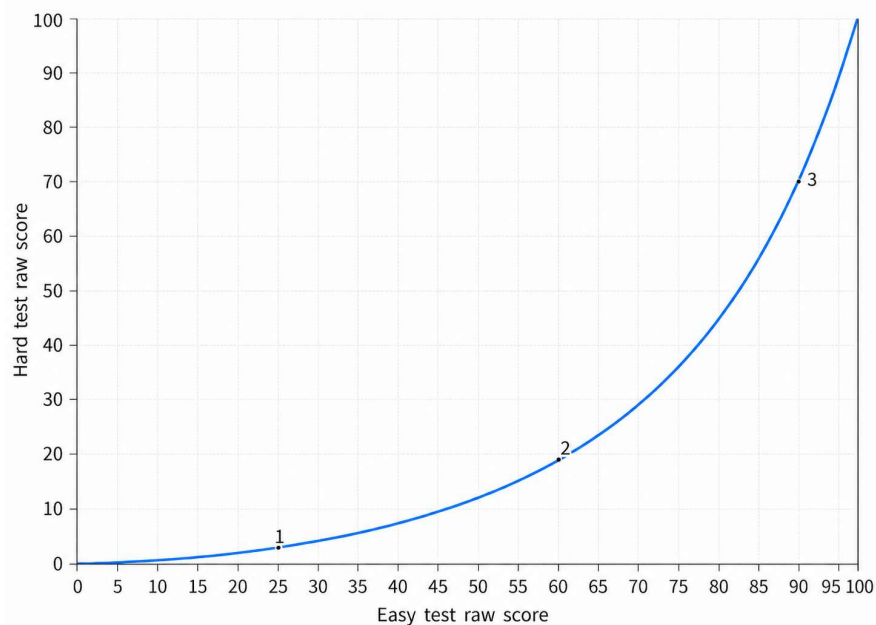
- Item statistics depend on the sample. As noted above, under CTT, question difficulty for a correct/incorrect question is the proportion of respondents who answer it correctly. As this proportion is calculated directly from the responses of the group taking the test, it thus depends on the characteristics of this group. For example, if a question is administered to a class of generally lower-achieving pupils, the proportion answering correctly is likely to be lower, making the item appear more difficult. Conversely, when the same item is administered to a higher-achieving group, the proportion correct is likely to be higher, making the item appear easier. Consequently, CTT item statistics describe how an item performs within a particular group of test takers. Thus, in CTT, item difficulty and discrimination are not fixed, sample-invariant characteristics of the test question.

²¹ Or as the item-total correlation that excludes the item itself from the total score

- Raw scores depend on the specific test administered. Raw test scores (e.g. percentage of items answered correctly) depend not only on the examinee's proficiency but also on the difficulty and composition of the test. The same person may obtain a higher percentage on an easier test and a lower percentage on a harder one. Consequently, a raw score such as percentage correct does not alone separate the examinee's performance from the properties of the particular items included on the test.
- Measurement error is the same across score levels. As noted above, CTT can be used to estimate the standard error of measurement and (in turn) confidence intervals around observed test scores. In its conventional form, however, CTT uses a single standard error of measurement for the whole test. This means that measurement precision is treated as constant across the distribution, with the same width of confidence interval applied to low, average and high-achieving test takers. In practice, a test may measure some individuals more precisely than others, particularly when its items are concentrated around certain levels of achievement. Ideally, this variation in measurement precision should be reflected in different standard errors of measurement for different score levels, indicating how the test provides more precise information for some test takers than for others.
- Lack of interval level of measurement. Under CTT, distances between test-takers' raw scores are not necessarily preserved across different tests. Indeed, the distances may not even carry the same meaning across different regions of the same raw score scale. Hence a 10-point difference on one version of a test may not correspond to a 10-point difference on another version of the test. Likewise, a 10-point difference near the bottom of the raw score scale may not represent the same underlying ability difference as a 10-point raw score difference near the top. This point is easiest to see through an example. Consider Figure 5.2.1 (adapted from Wu et al., 2017, Figure 6.2) which shows the scores of three test takers (1, 2 and 3) on two maths tests. The first is an easier test (plotted on the horizontal axis) and the second is a harder test (plotted on the vertical axis). Each test taker's performance on the two tests is marked on the curve. The figure shows that the three test takers appear in the same rank order on both tests. However, the distances between test takers are not preserved across the easier and harder assessments. For example, on the easy test, test taker 2 is slightly closer to test taker 3 than to test taker 1 in their raw scores. But, on the hard test, the gaps are rather different. Specifically, the raw score of test taker 2 is now much closer to that of test taker 1 than it is to test taker 3. This illustrates how equal differences in raw scores across tests do not necessarily

correspond to equal differences in the underlying ability being measured. Ideally, estimates of differences in mathematical ability should not depend on whether individuals happen to take an easier or harder version of a test.

Figure 5.2.1. Comparisons of raw test scores for three test takers across two hypothetical tests



Notes: Graph adapted from Wu et al. (2017, Figure 6.2). Figure adapted using AI.

Why are these limitations of classical test theory and raw test scores particularly important for ILSAs? One of their central aims is to compare achievement across different assessment cycles. Raw scores are not however directly comparable when different versions of a test are used. A change in the percentage of questions answered correctly could hence reflect differences in test difficulty as well as genuine changes in pupils' knowledge and skills. ILSAs must also assess pupils with a very wide range of achievement. Easier questions may be more suitable for some pupils or populations, while harder questions may provide better information about others. Yet results from these different test versions then need to be placed on a common scale, requiring suitable links between the tests (e.g. shared questions). Moreover, as illustrated in Figure 5.2.1, ILSAs should also maintain the same "distance" in achievement scores between pupils independent of whether they receive an easier or harder version of the test. Although some of these issues may be addressed using methods based on CTT, IRT provides a flexible framework for creating and comparing different test versions that help facilitate comparisons across groups and over time.

5.3 Item response theory (IRT). The conceptual part.

Many textbooks and other resources provide detailed introductions to IRT (e.g. Baker, 2001; Wu et al., 2017). Rather than repeat these accounts, this chapter offers a broad overview of what IRT is and why it is widely used in international large-scale assessments. Unlike many introductions, the discussion separates the main conceptual ideas from the underlying statistical methods. The aim is to build an intuitive understanding of IRT and its potential advantages, rather than the technical details.

IRT is conceptually different from CTT and changes the way we think about educational measurement. Whereas we described CTT as being test-centric, IRT can be considered to be item-centric. In CTT, the main focus is the total score on a particular administration of a test. In IRT, the starting point is instead the relationship between each individual test question (“item”) and the underlying attribute being measured. Different versions of a test can then be constructed by selecting different combinations of items from a larger item pool, with the aim of placing pupils and items onto a single common scale.

A key task in IRT is hence to estimate the characteristics of individual test questions. In the simplest version (a one-parameter IRT / Rasch model) questions differ only in their difficulty. Now suppose a large collection of questions has already been used in previous tests or pilot studies. This will mean their level of difficulty will already be “known” (or, rather, estimated in these independent samples). One can then design different versions of the test to be used in the future that include different combinations of questions. Moreover, as the property of each question is “known” (or has been estimated) in advance, different versions of the test can be made broadly comparable in terms of their difficulty, or we will have a good idea of how challenging different versions of the test will be for different groups (i.e. higher and lower achievers).

This also allows test designers to be strategic about which questions they include. By choosing easier or harder items, it is possible to adjust the overall difficulty of a test or focus it on a particular level of ability. For example, in a school entrance examination with a minimum pass mark, it would be sensible to include many questions with difficulty close to that threshold. Similarly, a test aimed at lower-achieving pupils can be built using correspondingly easier questions while continuing to allow their achievement to be compared to higher-achieving pupils on the same measurement scale. This would be done, for instance, by asking higher and lower ability pupils some common test questions (“anchor items”), but with other questions assigned

that are more appropriate to their level²². The use of such “anchor items” within an IRT framework can also help facilitate comparisons over time (provided the anchor items continue to function sufficiently similarly across groups and years). Specifically, as the properties of these questions have already been estimated, their inclusion in different assessment years can help show whether changes in scores reflect changes in pupils’ knowledge and skills rather than differences in test difficulty.

In doing so, IRT helps address some of the limitations of CTT. In the section above we noted how question difficulty in CTT (percentage correct) can depend heavily on the particular group of pupils taking the particular administration of the test. In contrast, IRT attempts to place the difficulty of each item onto a common scale, using evidence from previous testing or calibration samples. This, in theory, thus helps to overcome the problem that item statistics such as the percentage answering correctly in CTT are “sample specific” as previously discussed. Moreover, IRT also helps researchers compare results from tests that differ in difficulty. As the difficulty of the questions is “known” (or has been estimated) from previous test administrations, scores can be interpreted in relation to how challenging each test was. In other words, this helps to overcome the problem that raw scores from CTT are “test specific” as discussed in the section above. Finally, IRT provides a useful way of linking different versions of a test (i.e. “equating”). This can help ensure that tests used in different years are broadly comparable and that any remaining differences in difficulty can be estimated and taken into account. While equating of tests is also possible under CTT, IRT’s explicit focus on items makes equating a natural aspect of this approach.

IRT is particularly useful for ILSAs because one of their main aims is to measure changes in achievement over time. These assessments regularly introduce new questions, while retaining some questions from earlier cycles. IRT helps estimate how difficult the new questions are compared with the older ones and place results from different assessments on a common scale. This supports judgements about whether changes in scores are likely to reflect changes in pupils’ knowledge and skills rather than differences in the tests themselves. ILSAs also assess diverse populations with a wide range of achievement. Different pupils (and in some ILSAs different countries) may therefore receive partly easier or harder sets of questions, with IRT providing a practical way to link these different test versions to the same underlying scale. Moreover, IRT

²² This principle also supports adaptive testing, in which pupils receive questions of different difficulty depending on how well they performed earlier in the assessment (see section 6.9).

approaches can be extended in ways that have additional benefits given the goals of ILSAs, including designing assessments that cover a broad range of content while limiting total test time for individual pupils. Further discussion of such extensions to basic IRT models follows in chapter 6.

5.4 IRT - the statistical part. One-parameter model.

We now turn to the statistical foundations of IRT, though keeping technical notation to a minimum and focusing mainly on the underlying intuition. To simplify the discussion, we consider the most basic IRT model - the one-parameter model (sometimes referred to as the Rasch model). It is also assumed (for ease of explanation) that each question is scored as either correct (1) or incorrect (0).

A key aspect of IRT is that it uses a mathematical model to describe the probability that an individual test taker answers each question correctly. This is clearly somewhat different to the CTT approach summarised in equation 5.2.1, which focuses more directly on performance across the test as a whole. In the simplest one-parameter IRT model, the probability of a correct answer depends on two main factors (a) the test taker's level of proficiency in the area being measured and (b) the difficulty of the question. Specifically, the probability that a test taker answers a question correctly can be expressed as:

$$\pi = \frac{\exp(\theta - \delta)}{1 + \exp(\theta - \delta)} \quad (\text{equation 5.4.1})$$

Where:

π = The probability the test question is answered correctly.

θ = The test taker's latent ability in the skill or attribute being measured.

δ = The difficulty of the test question.

One can also rearrange this equation and take the natural logarithm of both sides to arrive at:

$$\log\left(\frac{\pi}{1-\pi}\right) = \theta - \delta \quad (\text{equation 5.4.2})$$

This version of the equation shows that the difference between a person's latent proficiency in the skill (e.g. their ability in maths) and the difficulty of the question is equal to the logarithm of the odds of answering the question correctly. This is commonly known as a "logit" and is the scale of IRT question difficulty and test taker's proficiency (i.e. test scores).

It is often said that this model (and indeed all standard unidimensional IRT models) rely on several important assumptions. These include:

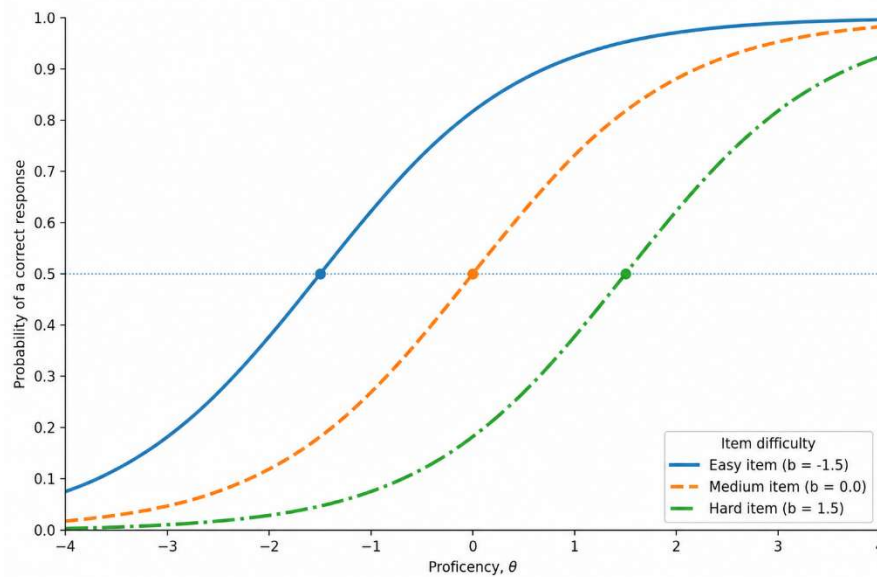
- Uni-dimensionality. The test questions mainly measure a single underlying area of proficiency²³.
- Local independence. Once a test taker's proficiency has been taken into account, their response to one question should not directly affect their response to another. For example, in a multi-part question, answering part (b) correctly should not depend heavily on having answered part (a) correctly.
- Monotonicity. as a person's proficiency increases, their probability of answering a question correctly should not decrease.

While the above are formal assumptions of many IRT models, they also reflect broader principles of good test design. For example, if a mathematics test depends too heavily on reading skills, its scores are unlikely to provide a clear measure of mathematical proficiency (regardless of whether IRT or CTT is used). Similarly, closely linked questions may provide less independent evidence about proficiency. Thus, in reality, such issues pose problems under both IRT and CTT approaches to constructing test scores.

Given IRT's emphasis on individual questions and their properties, the item difficulty parameter (δ) is worth further discussion. As it is measured in logit units, it can technically range from $-\infty$ to $+\infty$, though in reality it typically falls in the range -3 to $+3$. Its value is determined by the level of an individual's ability at which the probability of answering the item correctly is 0.5. Lower values indicate a question is easier and higher values that the question is harder. Within IRT, it is common to display such information in what is known as "item characteristic curves" (ICCs), with an example for three hypothetical questions presented in Figure 5.4.1. This plots individual proficiency in the subject along the horizontal axis with the probability of correctly answering the question on the vertical axis. Curves are then plotted for each question demonstrating the relationship between the two. In a one-parameter model, the curves never cross and simply shift to the left (easier) or right (harder) depending on the difficulty of the question.

²³ There are, however, multidimensional IRT models, where it is possible in some specifications to relax this assumption for some questions.

Figure 5.4.1. Hypothetical item characteristic curves for three items. Differences in item difficulty (equal discrimination).



Notes: Figure created using AI.

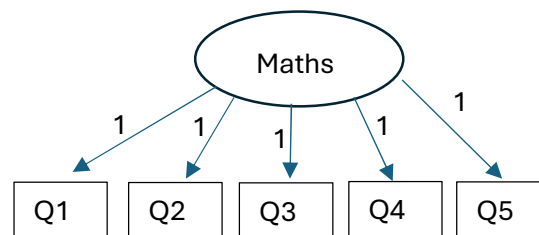
Several features of the one-parameter IRT model are worth highlighting. In particular, in a single administration of a test, one-parameter IRT “proficiency estimates” (test scores) are essentially just a logit-based transformation to the percentage of questions an individual answered correctly (i.e. CTT-based raw scores). In other words, the percent of questions a test taker answers correctly contains all the information needed to produce an IRT-based estimate of their proficiency. A similar one-to-one relationship applies to CTT and IRT estimates of question difficulty (i.e. percentage answering an item correctly and δ). This logit-based transformation helps address the potential problem with raw scores illustrated in Figure 5.2.1 in the previous section (where differences between raw scores can appear larger or smaller depending on whether the test contains mainly easy or difficult questions). Specifically, the logit transformation effectively “straightens” the relationship out, with the estimated differences in proficiency between test takers 1, 2 and 3 broadly consistent whichever test version they receive. But a clear implication is that (for test takers who receive the same set of test questions) percent correct and standard Rasch proficiency estimates will rank individuals in the same order and are usually very highly correlated with one another.

Hence it is not how IRT produces proficiency estimates per se that makes this approach attractive (Goldstein, 2015). Rather, as Wu et al. (2017, p.122) observe “IRT becomes useful when we are dealing with multiple tests where equating is needed.” In other words, its main attractions (particularly in the context of ILSAs) lie within our discussion of the broader conceptual issues discussed above. In particular, the explicit focus on the properties of individual questions makes

it possible to select items strategically to meet the goals of the assessment and the convenient linking of different versions of a test.

To conclude this subsection, we note that another way one can conceptualise a one-parameter IRT model is as a non-linear form of factor analysis (or, more broadly, as the measurement part of a structural equation model). Figure 5.4.2 illustrates this for a hypothetical maths test containing five questions. The oval labelled “maths” represents the underlying latent proficiency that we wish to estimate, while the arrows point to five binary questions (each scored as correct or incorrect). The “loadings” between mathematical proficiency and the questions are then all constrained to be equal in a one-parameter model (fixed at a value of one in Figure 5.4.2) – i.e. it is assumed that each question taps the latent maths construct equally well. The questions hence then only differ in their “intercepts” (i.e. difficulties in IRT terminology). This then demonstrates how a one-parameter IRT model can be viewed as a simple latent variable measurement model.

Figure 5.4.2. A latent variable illustration of a one-parameter item-response model



5.5 Two and three parameter IRT models.

The one-parameter model described above is the simplest form of IRT. It does, however, come with certain restrictions. In particular, every question is assumed to have the same discrimination parameter – i.e. all questions are equally sensitive to differences in the latent skill we are trying to measure. This idea is related to the earlier discussion of item discrimination under CTT in section 5.2, where discrimination was measured as the correlation between performance on an individual question and the overall test score. A one-parameter IRT model is thus conceptually similar to assuming that this correlation (between item scores and overall test scores) is the same for each test question.

A two-parameter IRT model relaxes this restriction by adding a discrimination parameter to equation 5.4.1, thereby allowing its value to vary across test questions. Equation 5.4.1 can consequently now be rewritten as:

$$\pi = \frac{\exp(\alpha(\theta - \delta))}{1 + \exp(\alpha(\theta - \delta))} \quad (\text{equation 5.5.1})$$

Where:

α = The item-discrimination parameter.

With all other terms as defined under equation 5.4.1 above.

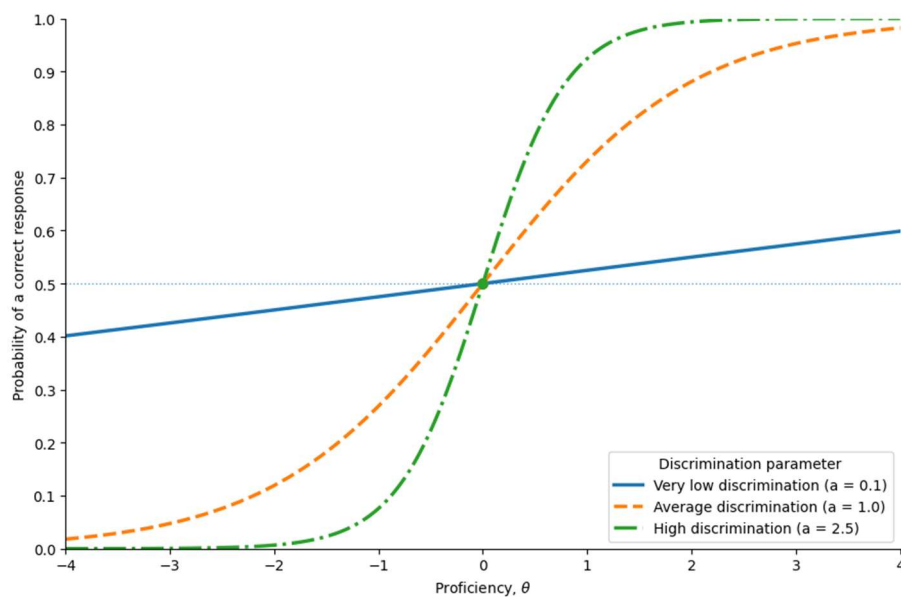
Higher values of the discrimination parameter are generally preferred to lower values. A low discrimination value suggests that a question does not distinguish well between pupils with higher and lower levels of the underlying trait we are trying to measure. It could also indicate that the question is unclear, poorly targeted or measuring something different from the other test questions. This links closely to the earlier discussion of discrimination under CTT, where a weak relationship between performance on an individual question and performance on the test overall is generally undesirable. On the other hand, questions with higher discrimination parameters tend to align more closely with the construct of interest²⁴ (or at least align more closely with the other questions included on the test).

Negative discrimination values are also possible. For example, when IRT is applied to survey responses containing a mixture of positively and negatively worded statements, some items may have negative discrimination parameters if they have not been reverse-coded. In this case, the negative sign simply reflects the direction of the relationship with the underlying construct and is not necessarily a problem. The item could instead be reverse-coded so that its discrimination parameter becomes positive (while retaining the same absolute value). Otherwise, for a positively coded question, a negative discrimination parameter will potentially indicate a possible problem with the item or its scoring.

Another way to understand this is by looking at the item characteristic curves. Under the one-parameter model (presented previously in Figure 5.4.1) the curves had the same shape and differed only in location, shifting left or right according to item difficulty. Under the two-parameter model, however, the curves can also change shape. Items with lower discrimination have flatter item characteristic curves, whereas high discrimination produces steeper curves. This is illustrated in Figure 5.5.1 below.

²⁴ Though does not by itself establish construct validity

Figure 5.5.1. Hypothetical item characteristic curves for three items with equal difficulty and different discrimination parameters.



Notes: Figure created using AI.

Notice that the curve for the item with a very low discrimination parameter is almost flat. This means that a person with very high proficiency (e.g. strong mathematical skills) has roughly the same probability of answering the item correctly as a person with very low proficiency (e.g. very weak mathematics skills). For instance, in this particular example, an individual with very high maths ability has a 60% chance of providing the correct response to the question, compared to 40% for an individual with very low maths ability. The item is therefore doing relatively little to distinguish between respondents at different maths ability levels and would not be a particularly useful mathematics test question. In fact, it could resemble the statement “my favourite food is cheese” discussed under CTT in section 5.2; responses may reflect something unrelated to the construct the test is intended to measure. Ideally, such poorly discriminating items would be identified during piloting and removed or revised before the test is used. In practice, however, such items are not always detected before they appear in the final assessment or survey instrument.

Including a separate discrimination parameter has several important implications for how we interpret and use results from an IRT model. Most importantly, our estimate of an individual’s proficiency no longer depends only on the total number of questions they answer correctly. In other words, the raw test score is no longer a sufficient statistic for the proficiency estimate as it is under the one-parameter model. The pattern of responses now matters as well. As a result, two

individuals who answer the same number of questions correctly on the same test (i.e. receive the same number of raw marks) may receive different IRT proficiency estimates.

One way to think about the two-parameter model is that a respondent's estimated ability is not determined solely by the number of questions answered correctly; the particular response pattern also matters. Items with larger discrimination parameters generally provide more information, especially for respondents whose ability is close to the item's difficulty. Consequently, two respondents with the same raw score can receive different estimated-ability scores depending on which items they answered correctly.

Returning to our example of a 50-question mathematics test discussed in section 5.2, suppose that the item "my favourite food is cheese" has a discrimination parameter close to zero. In a two-parameter model, this item would contribute very little to the estimation of mathematical proficiency. Whether a respondent answered it correctly or incorrectly would therefore have almost no effect on their estimated ability, because the item does little to distinguish between people with different levels of mathematical skill. This contrasts with the situation under a one-parameter model, where the discrimination parameter is constrained to be the same for every item. As a result, answering "my favourite food is cheese" correctly would affect the IRT test score in the same way as answering an actual mathematics question correctly (despite it clearly being unrelated to the construct the test is designed to measure).

A third parameter can also be added to the model (known as the "guessing" parameter). This in essence reflects the probability that someone with very low proficiency answers the item correctly. The clearest example is a multiple-choice item. Even if a respondent does not know the correct answer, selecting one option from four still means they have a chance of choosing it by chance. However, the guessing parameter is not determined simply by the number of response options. It may also be affected by the quality of the "distractors" (incorrect alternatives) and the ability of test takers to rule out unlikely answers. This third parameter thus actually reflects more than purely random guessing alone.

This leads to a further update to equation 5.5.1 so that it now also includes the guessing parameter:

$$\pi = \gamma + (1 - \gamma) \frac{\exp(\alpha(\theta - \delta))}{1 + \exp(\alpha(\theta - \delta))} \quad (\text{equation 5.5.2})$$

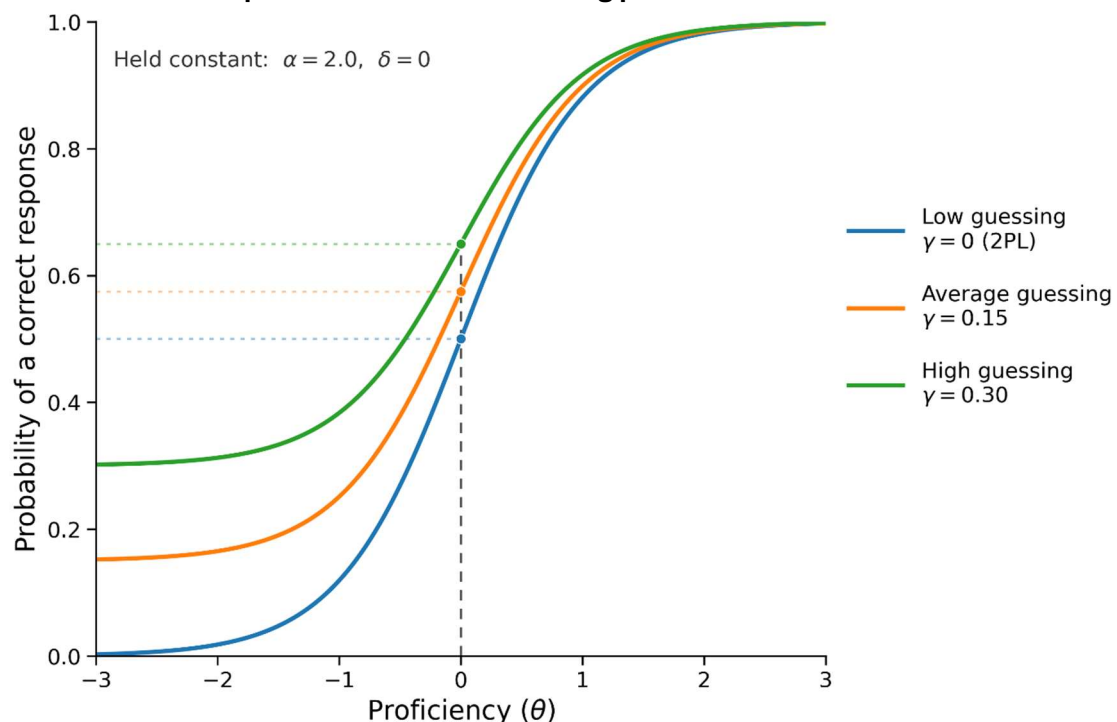
Where:

γ = The guessing parameter.

With all other terms as defined under equations 5.4.1 and 5.5.1 above²⁵.

The easiest way to understand the guessing parameter is through the item characteristic curves for each question. A hypothetical example is shown in Figure 5.5.2 below. The key feature is that the lower asymptote (i.e. where the left-hand tail of the curve levels off) sits above zero. For example, the item with a high guessing parameter has this probability sitting at 0.30. This means that even a test taker with a very low level of the latent construct being measured (e.g. very limited mathematical proficiency) still has around a 30% chance of answering the question correctly.

Figure 5.5.2. Hypothetical item characteristic curves for three items from a three-parameter model. Guessing parameter included.



Notes: Effects of different lower-asymptote parameters, with discrimination and difficulty held constant. Figure created using AI.

These different models are important as different ILSAs use different IRT models, with some having also changed which one they use over time. For example, PISA initially used a one-parameter model but moved to a “hybrid” approach from 2015 onwards which allows some item

²⁵ In this model, δ is now the point halfway between the lower asymptote (γ) and 1 (rather than necessarily the point where the overall probability equals 0.5).

discriminations to vary. TIMSS has also changed its approach over time and uses different models for different types of questions. For example, in 2019, TIMSS used a “three-parameter logistic (3PL) IRT model for multiple-choice items, the 2PL IRT model for constructed response items worth 1 score point, and the generalized partial credit model for constructed response items worth up to 2 score points” (von Davier, 2020, p. 11.4). How much the specific IRT model used matters for the key inferences drawn from ILSAs is, of course, an empirical question. Existing evidence on this matter is reviewed in the following chapter (see section 6.7).

Throughout this and the previous section, our discussion has focused on assessments formed of binary items where responses are scored as either correct or incorrect. Yet IRT models can also be extended to items with more than two ordered score categories (e.g. a two-mark test question that is scored as 0 for an incorrect response, 1 for a partially correct response and 2 for a fully correct response). As such, IRT can also be applied to questionnaire data, including a set of Likert-type items designed to measure a specific attitude, feeling or viewpoint. It is therefore useful for readers to think of IRT as a broader method for modelling responses and constructing scales from assessment or survey data. Indeed, IRT can be conceptualised as a form of categorical factor analysis or the measurement component of a broader structural equation model (an idea that we will build upon in the following chapter in section 6.4).

5.6 Differential item functioning and its links to measurement comparability

A key part of any international study is ensuring that the constructs being measured are genuinely comparable. The most obvious examples are when comparisons are being made across countries and languages, where cultural or linguistic issues may lead to different interpretations and responses to the questions being asked. Yet it also applies to other comparisons such as changes over time and differences across demographic groups (e.g. immigrants versus natives). Moreover, such issues not only apply to assessment items measuring achievement in subjects such as mathematics, but also to constructs captured through questionnaires. In TALIS, for example, researchers may wish to compare teachers’ job satisfaction across countries. For these comparisons to be meaningful, the questionnaire items used to measure job satisfaction need to function sufficiently similarly across the jurisdictions being compared (Zieger et al., 2019).

Within the literature, two broad approaches are commonly used to investigate measurement comparability. The first involves fitting a series of confirmatory factor analysis (CFA) models across the groups being compared (Millsap & Olivera-Aguilar, 2012). Increasing numbers of equality constraints are then imposed on the estimated parameters to test whether the measure

operates in the same way across groups in a process known as testing for measurement invariance. Several studies have applied this approach to data from international large-scale assessments, including Ding et al. (2023), Zieger et al. (2019) and Fang et al. (2025). Detailed overviews of the CFA approach, including the different levels of invariance that can be tested, are provided in Putnick & Bornstein (2016) and van de Schoot et al. (2012).

The second approach, which is the focus of this section, involves investigating differential item functioning (DIF). We focus on DIF because it is closely associated with the educational assessment literature and with concerns about whether test questions operate comparably across different groups. It also fits naturally with the earlier discussion of IRT and serves an important part of question development in ILSAs.

What, then, exactly is DIF? In essence, it occurs when respondents from different groups who have the same level of the construct being measured have different probabilities of answering a particular item correctly. For example, children in Country A and Country B who are equally proficient in mathematics might still perform differently on a particular question. This difference could be related to language, translation, cultural context or familiarity with the situation described. In such a case, we would say that the item may display differential item functioning. DIF analyses are commonly carried out across the items included in an assessment, particularly when new questions are being developed or existing questions are being reviewed. The aim is to understand whether each item appears to operate in a sufficiently similar way across the groups being compared. Detecting DIF does not automatically mean that an item is biased, but it does indicate that the item may require closer investigation.

5.6.1. Investigating DIF using logistic regression

Various methods have been proposed for detecting DIF. In keeping with the approach taken in this book, we focus on two methods that are either relatively intuitive or based on statistical techniques that are widely taught.

The first is logistic regression, proposed as a method for detecting DIF by Swaminathan & Rogers (1990). For ease of explanation, we focus on items scored as correct or incorrect and only two groups (e.g. two countries) being compared. However, the same general approach can be extended to comparisons involving more than two groups and to items with several ordered response categories (where ordered logistic regression is used instead). For each item being examined for possible DIF, we fit a logistic regression model in which the probability of a correct response is predicted using a measure of the individual's proficiency in the trait we are trying to

estimate (e.g. maths skills), group membership (e.g. country) and their interaction. The model can be written as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta \cdot G_{ij} + \delta \cdot \theta_{ij} + \gamma \cdot G_{ij} * \theta_{ij} \quad (\text{equation 5.6.1.1})$$

Where:

π_{ij} = The probability of a correct response to the test question.

G_{ij} = A variable for group membership (e.g. Country A compared to country B as the reference group).

θ_{ij} = The latent trait of interest (e.g. mathematics skills).

There are two parameters of interest from this model. The first is γ which provides an indication of what is known as “non-uniform DIF”. This occurs when the relationship between the trait of interest (θ_{ij}) and the probability of answering an item correctly differs across groups. In IRT terminology, this has similarities to asking whether the item discrimination parameter varies between groups (e.g. countries). Put differently, does the question distinguish more clearly between respondents with different levels of mathematical proficiency in Country A than in Country B? A sizeable and statistically significant interaction (γ) provides an indication that the item may display DIF (though rigid cut-off rules should generally be avoided), which may warrant further investigation (see section 5.6.3 below for further details).

The other parameter of interest from this model is β . This captures the association between group membership and the probability of answering the question correctly, after accounting for differences in the proficiency being measured. Provided that the γ parameter is deemed to be sufficiently small, the β parameter then provides an indication of “uniform DIF”. Again, whether such DIF is present is often decided based on the magnitude and statistical significance of β . This indicates that the question is consistently easier for one group than another for a reason unrelated to their latent proficiency. For example, a mathematics question might be more difficult for children in Country A than for equally proficient children in Country B. In an IRT framework, this is similar to asking whether the item difficulty parameter differs across groups. We note, however, that it is standard practice to only investigate possible uniform DIF for an item if it is concluded that no non-uniform DIF is present first (i.e. the interaction term – γ - is deemed to be sufficiently small).

5.6.2. Investigating DIF using IRT parameters and item characteristic curves

A broad range of IRT-based methods can be used to investigate DIF. Here we focus on two such approaches – comparisons of estimated IRT item parameters across groups and comparisons of item characteristic curves. For simplicity, the explanation again focuses on the case of two countries taking an assessment made up entirely of binary items. The broad approach does however extend to multiple groups and items with multiple score points.

One begins by fitting a “multiple-group” IRT model. This essentially means that responses from individuals within the two countries are included within the same IRT model. To place individuals from both countries onto a common scale, a set of anchor items are needed that are believed to function similarly across them (i.e. questions that are DIF free). The IRT parameters of these anchor items are then constrained to be equal across countries. The parameters of the remaining items (which may or may not display DIF) are then estimated separately for each country.

For these non-anchor items, the IRT parameter estimates are then compared across country A and country B.

If the estimated IRT parameters are similar, then there is little evidence that the item displays DIF. If the difficulty parameters differ while the discrimination parameters remain similar, this suggests uniform DIF (i.e. the item is consistently easier in one country than the other). A difference in the discrimination parameter provides an indication of non-uniform DIF - the relationship between proficiency and the probability of answering correctly differs across countries. Where both difficulty and discrimination differ, the item displays both non-uniform DIF alongside a difference in its difficulty. As under the logistic regression modelling approach, both the magnitude and the significance of the difference in the estimated IRT parameters across countries should be considered.

Table 5.6.2.1 illustrates this approach using a hypothetical 20-item test administered in two countries. For simplicity, this focuses on estimates from a multiple-group one-parameter IRT model, meaning only instances of uniform DIF are considered. Nine items are selected in advance as anchors, with their difficulty parameters constrained to be equal across countries (e.g. item 2 has a difficulty of -1.7 logits in both country A and B). The difficulty parameters for the remaining 11 items are estimated separately.

Table 5.6.2.1. A hypothetical example of an exploration of DIF using IRT on a 20-item test.

	Difficulty (b) parameter	
	Country A	Country B
Item 1	-2.00	-1.90
Item 2 (anchor)	-1.70	-1.70
Item 3 (anchor)	-1.50	-1.50
Item 4 (anchor)	-1.00	-1.00
Item 5 (anchor)	-0.80	-0.80
Item 6	-0.50	-0.40
Item 7	-0.30	0.80
Item 8	-0.20	-0.10
Item 9	-0.10	0.00
Item 10 (anchor)	0.00	0.00
Item 11 (anchor)	0.05	0.05
Item 12	0.15	0.05
Item 13	0.35	0.45
Item 14	0.55	0.45
Item 15 (anchor)	0.70	0.70
Item 16	0.90	0.85
Item 17 (anchor)	1.10	1.10
Item 18	1.50	0.00
Item 19	1.80	1.75
Item 20 (anchor)	2.00	2.00

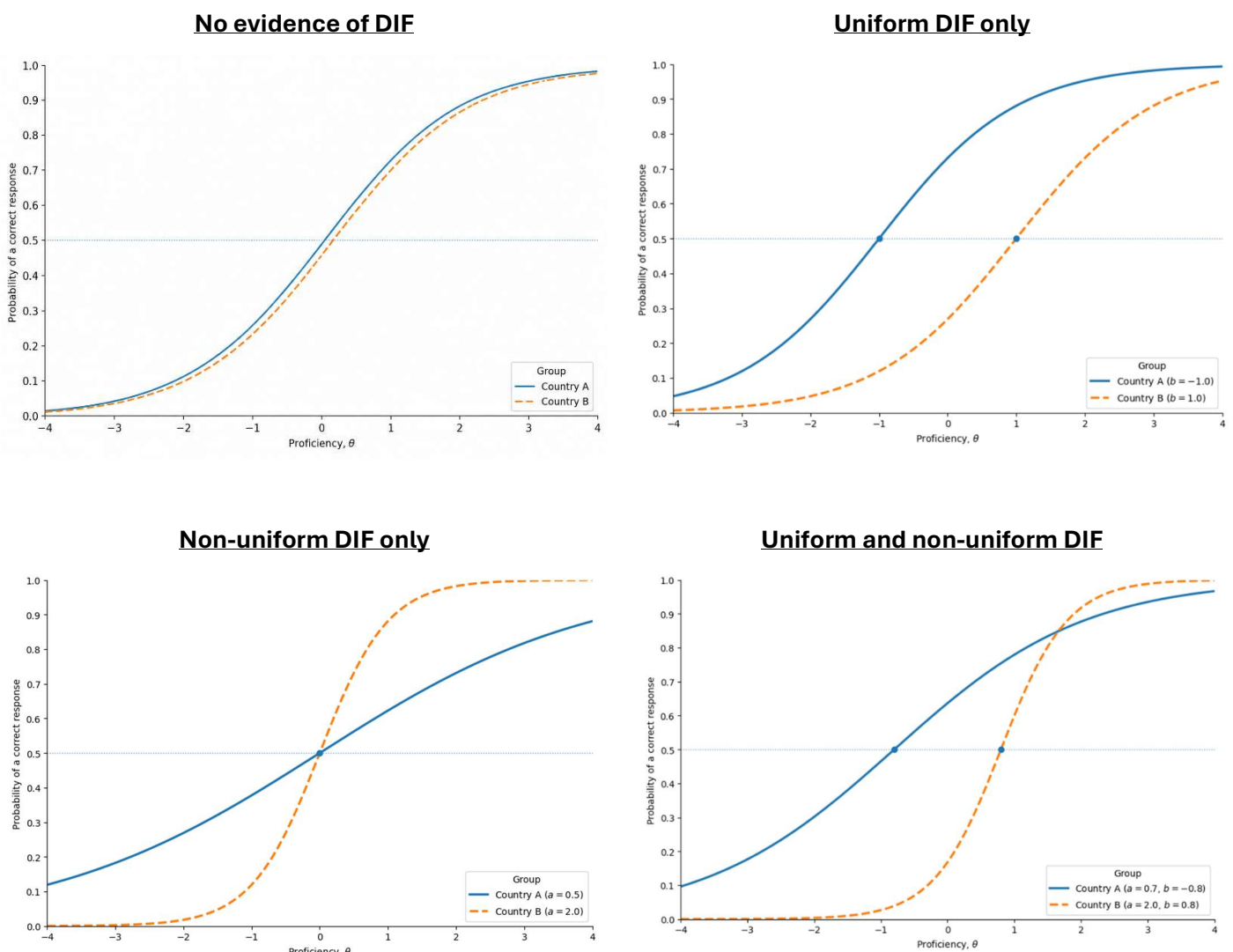
Most of the difficulty parameters on the 11 questions where these have been estimated are very similar across the two countries, providing little evidence of DIF (e.g. Item 6 has an estimated difficulty of -0.5 in Country A and -0.4 in Country B). Yet two questions clearly stand out. The first is item 7, where the difficulty is estimated to be -0.3 in Country A but 0.8 in Country B. This thus indicates that item 7 is substantially easier in Country A. On the other hand, item 18 shows the opposite pattern, with estimated difficulties of 1.5 and 0.0 respectively. This difference of 1.5 logits therefore suggests that item 18 is considerably harder in Country A, providing evidence of uniform DIF in the opposite direction.

An important point in the example above is that DIF is identified at the level of individual items. In many applications, however, the more important issue is whether these item-level differences affect the test as a whole. This is known as differential test functioning (DTF) which occurs when the combined set of items on a test systematically favours one group over another (over and above differences across groups in their latent proficiency). In the example above, the impact of item 7 and item 18 may thus partly cancel out when the overall test score is calculated.

Consequently, substantial DIF at the item level may actually have relatively little effect at the test level.

A complementary graphical approach is to inspect the item characteristic curves after fitting the multiple-group IRT model. As the shape and position of each curve are determined by the item parameters, this is closely related to comparing difficulty and discrimination estimates across countries. The difference is that the evidence is assessed visually. Curves that closely overlap suggest little obvious evidence of DIF, while clear differences may indicate that the item functions differently across groups.

Figure 5.6.2.1. Item characteristic curves illustrating different examples of DIF



Notes: Figure created using AI.

Figure 5.6.2.1 provides an example of this approach. Each panel presents the results for one item, comparing its item characteristic curve in Country A with that in Country B. In particular:

- Top left panel. The item characteristic curves are very close to one another, providing little evidence of DIF.
- Top right panel. The curve for Country A lies to the left of the curve for Country B, indicating that the item is easier in Country A. This is indicative of uniform DIF.
- Bottom left panel. The curves have different shapes, with it being flatter in Country A than B. This indicates the discrimination parameter is lower in country A, with the item thus distinguishing less clearly between individuals at different proficiency levels. This indicates non-uniform DIF.
- Bottom right panel. The curves differ in both their location and shape, indicating differences in both difficulty and discrimination. The item therefore displays non-uniform DIF alongside a difference in its overall difficulty.

5.6.3. What should be done about DIF?

After detecting DIF in an item, the first step one should take is to return to the question itself and examine it closely. Is there an obvious reason why one group might perform better than another for reasons unrelated to the latent skill being measured? Possible explanations include differences in language or translation, the way the question is displayed or its cultural relevance. For example, a mathematics question based on the sport kabaddi might be easier for children in Bangladesh than for children in the United Kingdom (e.g. due to differences in familiarity with the sport rather than mathematical proficiency).

This substantive review is critical. If DIF appears to arise from a feature of the item that is unrelated to the intended construct, it likely indicates item bias which is problematic for cross-national comparisons. Yet it may also be the case that no clear reason for DIF can be identified. The item may have thus been flagged purely by chance or a relevant aspect of the trait we are measuring that is not well captured in other parts of the test. As such, DIF may actually be of substantive interest. It is hence vital that – once DIF has been detected – one uses their subject knowledge and understanding of the groups being compared to consider why this has occurred.

After doing so, then for each DIF item, one can take the following courses of action:

- Do nothing. When an item displays DIF, one possible response is to leave it unchanged, continue to include it in the test / survey and produce the IRT proficiency estimates in the usual way. This may be reasonable where there is no clear evidence of item bias and the

item otherwise appears to function well. Indeed, automatically removing every item that shows DIF can end up just narrowing the content covered by the test (and the construct it measures). Best practice would then be to investigate how comparisons of overall test score differences between groups change with and without the DIF items.

- Remove the item. At the other extreme, the item may be removed. After reviewing the question, researchers may identify a clear and credible reason why respondents from different groups are reacting to it differently for reasons unrelated to the latent trait being measured. The item may thus be judged unsuitable for cross-group comparisons. If the problem is detected during piloting, the wording, translation or presentation may be revised before the main study is conducted. Alternatively, the item may be dropped from the final assessment. DIF may also be discovered only after the main data have been collected. In this instance, the question may be excluded from the scoring model. For example, if one item in a 50-item test shows substantial and concerning DIF, proficiency estimates could be recalculated without it. Note the question could either be removed for all groups or only for those where clear DIF is present. This, of course, then introduces the challenge of a different set of test questions being used in different countries – which has implications for face validity at the very least.
- “Free” the IRT item parameters. Rather than assuming that an item has exactly the same difficulty or discrimination parameter in every country, the IRT model used may allow selected parameters to differ for particular groups (e.g. countries or languages). In practical terms, this means recognising that an item may be more difficult or may distinguish more clearly between respondents with different proficiency levels in one group than in another. This strategy is used in ILSAs such as PISA. While most items retain common international parameters, selected questions are assigned group-specific parameters. ILSAs generally use this approach sparingly, however, as enough items must retain common parameters to maintain a shared international scale. The impact of allowing item parameters to vary across countries has been investigated by Jerrim et al. (2018) in the context of PISA. They found that allowing selected item parameters to differ across countries made little substantive difference to the resulting cross-national comparisons.

5.6.4 Advice to government officials and applied researchers on DIF

What, then, should government officials and applied researchers do in their own considerations of DIF, whether this be in the context of ILSAs or other assessments/surveys? My suggestion is that the following three-step process would be close to best practice to follow:

Step 1. Develop a priori expectations. Once an assessment or survey has been designed, and before data are collected, the items should be reviewed by subject, language and measurement experts for possible DIF. This should rely solely on theory and substantive judgement rather than statistical evidence. Experts should consider which groups might be affected by DIF, the likely direction of any difference and its possible cause. Ideally, several experts should review the items independently so that the consistency of their judgements can be assessed (e.g. measuring their inter-rater agreement), with their ratings and reasoning clearly documented.

Step 2. Conduct statistical DIF analysis. After data have been collected, statistical analyses of DIF should be carried out (including, for instance, the methods discussed above). This should include measures of both the magnitude of the DIF and an indication of statistical significance. Where practical, it should be conducted without the analyst seeing the predictions from Step 1.

Step 3. Review the combined evidence. Each item should then be reconsidered using evidence from both stages. Statistical DIF that matches an earlier concern strengthens the evidence for a particular explanation (e.g. potential item bias for cultural reasons). However, unexpected DIF should also be taken seriously and should prompt a fresh review of the item. Decisions to retain, revise, remove or free the parameters of an item should then be made. Again, initial views on such matters should ideally be formed and recorded by multiple experts independently (to explore their consistency), before a final decision is made. The evidence and reasons for each decision should then be documented transparently.

Based on the publicly available documentation, the kind of process outlined above appears to only be partially followed in the context of ILSAs. To their credit, item review procedures are extensive and often include translation checks, expert review, field trials and statistical analysis of DIF. Final modelling decisions are also commonly reported. In PISA, for instance, the technical reports clearly document where item parameters have been allowed to vary across particular countries or language groups (e.g. OECD, 2024). However, there is often less publicly available information about the substantive reasoning behind each item-level decision. In particular, it is often unclear what experts believe has caused the DIF and whether these concerns had been raised before the statistical results were examined. This means that, without such documentation, it is difficult to know how much weight has been placed on statistical evidence alone.

5.7. Test your knowledge questions

5.7.1. Task 1.

Task 1. Question 1. There is a 20-question test of pupils' arithmetic skills conducted across two countries (Country A and Country B). A one-parameter item-response theory (IRT) model is estimated separately in each country, with the resulting difficulty parameters then placed on the same scale. The difficulty (b) parameter estimated for each item is presented in Table 5.7.1.1.

Please review the questions in this table and discuss whether you believe there is any evidence of differential item functioning (DIF), and whether this suggests that there may be a problem with the comparability of the test (or particular test questions) across countries. Start your answer by providing a brief summary of what is meant by DIF.

Table 5.7.1.1. Item difficulty parameters from a hypothetical 20-item test

	Difficulty (b) parameter	
	Country A	Country B
Item 1	-2.00	-1.90
Item 2	-1.70	-1.75
Item 3	-1.50	-1.60
Item 4	-1.00	-0.95
Item 5	-0.80	-0.90
Item 6	-0.50	-0.40
Item 7	-0.30	0.80
Item 8	-0.20	-0.10
Item 9	-0.10	0.00
Item 10	0.00	0.05
Item 11	0.05	0.00
Item 12	0.15	0.05
Item 13	0.35	0.45
Item 14	0.55	0.45
Item 15	0.70	0.75
Item 16	0.90	0.85
Item 17	1.10	1.05
Item 18	1.50	0.00
Item 19	1.80	1.75
Item 20	2.00	1.95

Task 1 Question 2. You are now shown Table 5.7.1.2. This provides the wording of each question on this 20-item test. How can this information be used to help further inform the answer you provided to task 1 question 1 above? If you believe DIF is present, please discuss possible explanations for what may be causing it.

Table 5.7.1.2. Question wording on the hypothetical 20-item test

Test questions	
Item 1	What is $3 + 4$
Item 2	What is $7 + 8$
Item 3	What is $6+6$
Item 4	What is $7+9$
Item 5	What is $10+6$
Item 6	What is $11 + 15$
Item 7	What is 5×5
Item 8	What is $13 + 14$
Item 9	What is $17 + 18$
Item 10	What is $20 + 13$
Item 11	What is $21+ 16$
Item 12	What is $17 + 18$
Item 13	What is $19 + 16$
Item 14	What is $21 + 23$
Item 15	What is $43 + 27$
Item 16	What is $38 + 25$
Item 17	What is $47 + 85$
Item 18	What is $33 + 81$
Item 19	What is $85 + 57$
Item 20	What is $94 + 63$

Task 1. Question 3. Given the information provided in Tables 5.7.1.1 and Tables 5.7.1.2, what approach would you take to constructing test scores to consider any differential item functioning (DIF) you may have found? Please justify your answer.

Task 1. Question 4. If this test were to be used again, what changes do you think could be made to it to enhance its cross-national comparability.

Task 1. Question 5. What would be the implication of using a two-parameter IRT model rather than a one-parameter IRT model to construct test scores using these data? Could you name one clear advantage of using a two-parameter model here?

Task 1. Question 6. Based on the evidence above, to what extent do you believe that comparing average total summative scores (i.e. just giving pupils a score between 0 and 20 on the test) in the two countries would provide a reasonable basis for understanding differences between countries in test-takers arithmetic skills?

5.7.2. Task 2.

You have been supplied with a datafile. This includes data from 17,554 students from across 8 countries, who provided information about the disciplinary climate in their mathematics class.

They answered 10 questions in total using a two-point scale (0= Disagree; 1 = Agree). The ten questions pupils were asked are as follows:

- Question A = When the lesson began, our mathematics teacher has to wait quite a long time for us to quieten down
- Question B = We lost quite a lot of time because of students interrupting the lesson.
- Question C = There was much disruptive noise in this classroom.
- Question D = In our teacher's class, we were aware of what was allowed and what was not allowed.
- Question E = In our teacher's class, we knew why certain rules were important.
- Question F = Our teacher managed to stop disruptions quickly.
- Question G = Our teacher reacted to disruptions in such a way that the students stopped disturbing learning.
- Question H = In our teacher's class, transitions from one phase of the lesson to the other took a lot of time.
- Question I = Our teacher was immediately aware of students doing something else.
- Question J = Our teacher was aware of what is happening in the classroom, even if he or she was busy with an individual student.

Pupils' responses to these questions are to be used to create a "classroom management" scale. This captures pupil's views on the disciplinary climate within lessons.

Task 2. Question 1. To begin, a two-parameter IRT model is used to create the classroom management scale. The estimates discrimination parameters are presented in Table 5.7.2.1. Please comment on these estimated discrimination parameters and whether there is anything unusual about them.

Table 5.7.2.1. Discrimination parameters from a two-parameter IRT models used to create the classroom management scale.

Question	Discrimination parameter
A	-0.63
B	-0.66
C	-0.72
D	1.87
E	2.28
F	2.66
G	2.28
H	-0.33
I	1.73
J	1.94

Task 2. Question 2. This two-parameter IRT model is estimated a second time, but now with questions A, B, C and H reverse coded (i.e. for these questions, 0 now equals agree and 1 disagree). What do you think would happen to the estimated discrimination parameters for these questions?

Task 2. Question 3. What do you think the correlation between the classroom management scales would be under the approach used in task 2 question 1 and task 2 question 2?

Task 2. Question 4. A one-parameter IRT model is now used to create the classroom management scale instead. What does this imply in terms of the assumptions one is making regarding how well each survey question captures the “classroom management” construct?

Task 2. Question 5. Two versions of the classroom management scale are created when using a one-parameter IRT model. The first version uses the original scoring for questions A, B, C and H (0 = disagree; 1 = agree). The second version uses the reverse-coded versions (0 = agree; 1 = disagree). The correlation between these two scales is just 0.15. Can you explain why it is so low?

Task 2. Question 6. Table 5.7.2.2 provides the correlation between the four classroom management scales created thus far. Which looks to be the odd-one out? Can you think why?

Table 5.7.2.2. Correlation between the four versions of the classroom management scales created using different IRT models

	Two-parameter. Original scoring.	Two-parameter. Reversed items	One-parameter. Original scoring.	One-parameter. Reversed items.
Two-parameter. Original scoring.				
Two-parameter. Reversed items.	1.00			
One-parameter. Original scoring.	0.54	0.54		
One-parameter. Reversed items.	0.90	0.90	0.15	

Task 2. Question 7. Table 5.7.2.3 provides the correlation between the responses to each question. Does this point to anything unusual? What might explain some of the low correlations?

Table 5.7.2.3. Correlation in the responses to the ten classroom management questions

	A (reversed)	B (reversed)	C (reversed)	D	E	F	G	H (reversed)	I	J
A (reversed)										
B (reversed)	0.74									
C (reversed)	0.65	0.73								
D	0.08	0.09	0.11							
E	0.08	0.10	0.12	0.63						
F	0.17	0.18	0.20	0.39	0.46					
G	0.09	0.09	0.11	0.36	0.44	0.62				
H (reversed)	0.43	0.44	0.41	0.04	0.02	0.04	-0.02			
I	0.01	0.01	0.04	0.31	0.39	0.44	0.44	-0.06		
J	0.04	0.04	0.07	0.35	0.43	0.46	0.45	-0.04	0.62	

References

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. <https://files.eric.ed.gov/fulltext/ED458219.pdf>

Ding, Y., Yang Hansen, K., & Klapp, A. (2023). Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychology of Education*, 38, 709–732. <https://doi.org/10.1007/s10212-022-00623-y>

Ellis, J. L., & Sijtsma, K. (2024). Proof of reliability convergence to 1 at rate of Spearman–Brown formula for random test forms and irrespective of item pool dimensionality. *Psychometrika*, 89(3), 774–795. DOI: 10.1007/s11336-024-09956-7

Fang, G., Teo, T., & Chan, P. W. K. (2025). Testing for approximate measurement invariance of instructional quality in the Teaching and Learning International Survey (TALIS) 2018. *Humanities and Social Sciences Communications*, 12, Article 520. <https://doi.org/10.1057/s41599-025-04870-4>

Hambleton, R.K. & Jones, R.W. (1993). An NCME instructional module on comparison of classical Test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12: 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>

Jerrim, J., Parker, P., Choi, A., Chmielewski, A.K., Sälzer, C. & Shure, N. (2018), How Robust Are Cross-Country Comparisons of PISA Scores to the Scaling Model Used? *Educational Measurement: Issues and Practice*, 37: 28-39. <https://doi.org/10.1111/emip.12211>

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). The Guilford Press.

OECD (2024), *PISA 2022 Technical Report*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/01820d6d-en>.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental review: DR*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. <https://doi.org/10.1111/j.1745-3984.1990.tb00754.x>

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item Response Theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 11.1–11.25). Boston College, TIMSS & PIRLS International Study Center. <https://timssandpirls.bc.edu/timss2019/methods/chapter-11.html>

Wu, M., Tam, H. P., & Jen, T.-H. (2017). *Educational measurement for applied researchers: Theory into practice*. Springer Singapore. <https://doi.org/10.1007/978-981-10-3302-5>

Zieger, L., Sims, S., & Jerrim, J. (2019). Comparing teachers' job satisfaction across countries: A multiple-pairwise measurement invariance approach. *Educational Measurement: Issues and Practice*, 38(3), 75–85. <https://doi.org/10.1111/emip.12254>

6. Assessment design and plausible values

6.1 Chapter introduction.

This chapter examines the assessment designs used in ILSAs; a key topic for anyone seeking to understand how these studies generate and report measures of achievement. Several features of ILSA assessment design can be surprising to those encountering the data for the first time, particularly because these studies are not primarily designed to produce precise scores for individual pupils. The chapter thus begins by distinguishing between assessments intended to support inferences about individuals and those designed to describe groups or populations. This distinction has important implications for how tests are constructed, how many questions each pupil answers and how achievement is estimated. The concept of multiple matrix sampling in assessment design is then introduced; an approach that has long been central to ILSAs.

The chapter then provides a nontechnical overview of how ILSA achievement estimates are produced from pupils' responses to assessment questions, together with information collected within the background questionnaires. This leads us to plausible values - the closest analogue to individual test scores contained within ILSAs. We explain why plausible values are generated and how it is recommended they are used in statistical analysis. We also critically consider how much difference it makes in practice when these recommendations are not followed. Finally, the chapter turns to recent developments in assessment design, including the growing use of computer-based testing and adaptive assessment.

6.2 Trade-offs between content breadth, measurement precision and test length

For most people, the familiar purpose of a test is to produce a score that measures an individual's knowledge and skills in a particular subject. The focus is therefore on assessing individuals (e.g. determining how Child A is performing in mathematics and comparing that performance with Children B, C and D).

The goals of ILSAs are different.

They are not primarily designed to provide precise measures of the abilities of individuals. Instead, their main purpose is to facilitate comparisons at the population (and sub-population) level (e.g. how pupils in Country A perform in mathematics, on aggregate, relative to those in Countries B, C and D). This difference in purpose has important implications for assessment design.

Broadly speaking, assessment design can be thought to involve a trade-off between the following three factors:

- Breadth of content. Do we want to measure children's knowledge and skills in a single subject or across several areas? Even within one subject (e.g. mathematics) should the assessment cover as many topics and concepts as possible?
- Measurement precision. Do we want to measure achievement precisely at the individual, school or group/population level?
- Test length. How long are we willing (and able) to make the assessment?

Improving one of these aspects usually places additional pressure on the others. For example, expanding the assessment to cover a second subject will require a longer test or reduce the precision with which individual pupils' skills are measured. Similarly, improving the precision of individual scores generally requires either more test items or a narrower range of content (though this may then have implications for the validity of the assessment). Conversely, keeping the assessment reasonably short usually means accepting some compromise in either subject coverage, measurement reliability, or both.

Like all assessments, ILSAs must balance these competing priorities. In general, they are designed so that:

- Content coverage is broad. For instance, PISA assesses pupils in at least three core subjects during each cycle, while TIMSS assesses mathematics and science. Coverage within each subject is also extensive. For example, when mathematics is the focal subject in PISA, results are also reported for the "quantity", "space and shape", "change and relationships" and "uncertainty and data" sub-domains. Broad content coverage is therefore a priority in ILSA assessment design.
- Individual-level measurement precision is limited. As noted above, ILSAs are not designed to support inferences about individual test takers (or schools). Their focus is on estimates of distributions, means and relationships at the population and subpopulation levels. Individual-level measurement precision therefore has a low priority. However, achieving sufficient precision and comparability for population-level comparisons between countries and over time over time is of paramount importance.
- Test length is limited. As discussed in chapter 1 (section 1.6.3) ILSAs are low-stakes assessments. Because little is riding on the results for individual participants, there are limits to what can reasonably be asked of pupils, teachers and schools. For instance,

longer assessments may reduce test effort. Test length is therefore kept manageable. For instance, in PISA each pupil normally completes 120 minutes of cognitive testing, divided into two 60-minute sessions separated by a short break, followed by a background questionnaire.

These choices give ILSAs the best opportunity to compare population-level outcomes reliably across countries, subject areas and time. Yet this does therefore mean that ILSA test designs are somewhat unusual; their structure bears little resemblance to the tests most of us took at school. Remember, though, those assessments had a very different purpose to ILSAs; they were attempting to measure our individual knowledge and skills.

6.3 Multiple matrix sampling in assessment designs.

As noted in the previous subsection, ILSAs prioritise broad subject coverage and a manageable assessment length over producing scores that are sufficiently precise for reporting at the individual level. This has important implications for how the tests are designed. In particular, ILSAs typically draw upon a very large pool of assessment items (far more than any one individual could complete in a single test). For example, the PISA 2022 assessment contained 234 mathematics items, 197 reading items, 115 science items and 46 financial literacy items (National Center for Education Statistics, n.d.). This large item pool allows PISA to cover a very broad range of subjects, topics and skills.

Yet it is clearly impossible for individual test takers to answer so many questions, particularly in a low-stakes assessment. Consequently, what typically happens in ILSAs (although the precise design varies) is that questions are organised into separate subject-specific clusters (e.g. a cluster containing only reading items). Items are organised into subject-specific clusters. In this specific design, each item normally belonged to one cluster. Several clusters are then combined to create different test forms (i.e. assessment booklets). These forms are subsequently assigned to pupils, so each participant answers only a subset of the full item pool²⁶.

Table 6.3.1 provides a concrete example drawn from an earlier PISA cycle (2009). More recent cycles include additional complexities, such as elements of adaptive testing in PISA 2022 and

²⁶ The assignment of test booklets to pupils was traditionally random in ILSAs. In more recent cycles, however, an element of adaptive assessment has been introduced. Under this approach, test takers may be routed toward easier or more difficult questions depending on how they performed on earlier parts of the assessment. This is discussed in greater detail in section 6.9.

2025, which are overviewed at the end of this chapter (section 6.9). Nevertheless, the 2009 design provides a useful illustrative example for conveying the basic ideas.

Table 6.3.1. An example of a multiple matrix test design

Booklet	Clusters			
1	M1	R1	R3	M3
2	R1	S1	R4	R7
3	S1	R3	M2	S3
4	R3	R4	S2	R2
5	R4	M2	R5	M1
6	R5	R6	R7	R3
7	R6	M3	S3	R4
8	R2	M1	S1	R6
9	M2	S2	R6	R1
10	S2	R5	M3	S1
11	M3	R7	R2	M2
12	R7	S3	M1	S2
13	S3	R2	R1	R5

Source: Adapted from Table 2.1 of OECD (2012).

In this PISA cycle, reading was the focal subject (“major domain”), with science and mathematics being the “minor domains”. (The focal subject cycles between reading, mathematics and science each time PISA is conducted). The assessment pool thus contained 131 reading items, 35 mathematics items and 53 science items. These were organised into seven reading clusters (R1 to R7 in Table 6.3.1), three mathematics clusters (M1 to M3) and three science clusters (S1 to S3). Each cluster contained approximately 30 minutes of assessment material. The clusters were then combined to create 13 different test booklets. Each booklet included four of the 13 clusters, producing a total testing time of around two hours. Pupils were then randomly assigned to complete one of these 13 booklets.

There are some key points to note from Table 6.3.1.

First, pupils do not all answer the same test questions. Instead, each pupil completes only a subset of items drawn from the full assessment pool. This reinforces the earlier point that ILSAs are not designed to produce sufficiently reliable scores for individual reporting (different pupils answer different sets of test questions).

Second, many pupils do not answer questions in all three core PISA subjects. For example, pupils assigned to booklet 6 in PISA 2009 completed reading questions only. Those assigned to booklet 2 completed 90 minutes of reading questions, 30 minutes of science questions and no

mathematics questions. This again highlights how ILSAs are not designed to support inferences about individuals. After all, how can one judge how well a young person performs in mathematics if they answered no mathematics questions at all!?

Third, the different test booklets are distributed across large numbers of pupils within each country. For example, a country with 4,000 participants would have approximately 300 pupils completing each booklet. Moreover, as each item appears in several booklets, a substantial number of pupils respond in each country to each test question. Thus, by combining information across the full national sample, ILSAs can estimate population and subgroup outcomes with much greater precision than they can for individual test takers. This uncertainty does not simply disappear. Rather, it is incorporated into the analysis through what are known as “plausible values” (discussed in section 6.5 below).

6.4 A non-technical summary of how “test scores” (plausible values) are generated in ILSAs.

Given the complexity of the assessment design described above, it may not be immediately clear how pupils’ responses to the test questions they were assigned are converted into the achievement measures widely reported when ILSA results are released. Broadly speaking, this is done using a sophisticated IRT model that combines information from the questions each pupil answered with other available data. The full technical details are beyond the scope of this book, but they are discussed in Mislevy (1991), Wu et al. (2017), Bezirhan & von Davier (2024) and Zieger et al. (2022). As the precise details again vary across ILSAs and assessment cycles, our focus here is the general intuition of the process. Specifically, it can be broken down into five steps:

- Step 1. The process begins by estimating an initial multiple-group IRT model. Its purpose is to obtain estimates of the item parameters introduced in the previous chapter (item difficulty, discrimination and, where a three-parameter model is used, guessing). These item characteristics are estimated using responses from pupils across all the participating countries (including possibly also those who participated in ILSAs in previous cycles). Once obtained, the item-parameter estimates are treated as fixed in the later stages of the process described below.
- Step 2. Alongside the ILSA assessment, participants complete one or more background questionnaires. These collect a large amount of information about pupils’ characteristics, home environments, attitudes and educational experiences. An unusual feature of ILSAs

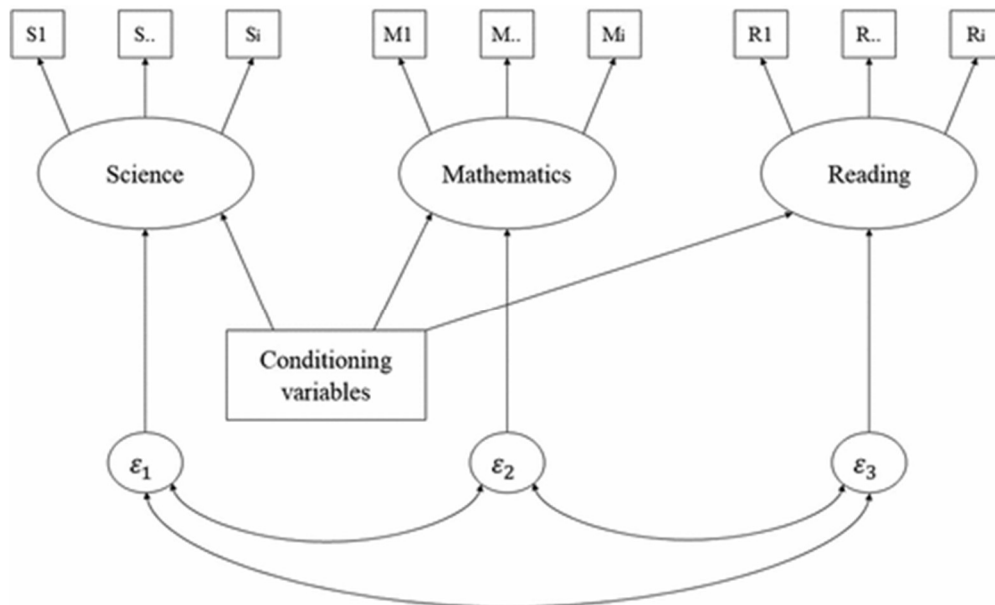
is that this information is used in the model that generates the “test scores”²⁷. However, with such a large amount of data, it would not be feasible to include all these background variables within the model. A dimensionality-reduction technique (principal component analysis) is therefore first applied to the background data. This condenses the large set of background variables into a much smaller number of components that retain most of the information contained in the original data (e.g. around 90%). These principal components (along with the item parameter estimates in step 1) are then included in the next stage of the process.

- Step 3. Pupils’ responses to the assessment items, the fixed IRT item parameters obtained in Step 1 and the principal components derived in Step 2 are then brought together within a “latent regression” model. Figure 6.4.1 provides a simplified illustration of the model used in PISA. The ovals represent the latent proficiency dimensions being measured (reading, mathematics, and science) while the squares labelled R, M, and S represent the assessment items used to measure each domain. The “conditioning variables” pointing toward the proficiency dimensions (reading, maths, science) correspond to the background information summarised through the principal components discussed in step 2 above. After allowing for the conditioning variables, each latent proficiency has a residual component (ε_1 , ε_2 and ε_3). The curved double-headed arrows show that these residual components may be correlated across domains, reflecting the fact that pupils who perform strongly in one subject often perform strongly in others. Another way this can be conceptualised is as a structural equation model in which the item parameters on the test questions (R, M and S) are treated as fixed values estimated during the first stage of the process.
- Step 4. Once the model has been estimated, each pupil is assigned a set of achievement estimates known as plausible values. The key idea is that there is considerable uncertainty about any individual pupil’s proficiency in the assessed domains (an intentional consequence intertwined with the choice of the assessment design). Therefore, rather than assigning a single test score, ILSAs provide multiple “plausible values” for each pupil in each subject. We discuss these plausible values further in the section that follows.

²⁷ Selected direct or derived covariates such as gender, assessment form and school IDs may also be included in the model.

- Step 5. A linear transformation is conducted to transform the estimated plausible values onto the ILSA reporting scale (e.g. see Bezirhan & von Davier, 2024, for further details with respect to TIMSS). This transformation is study, cycle and domain specific.

Figure 6.4.1 Simplified representation of the PISA 2012 conditioning model used to generate the PISA “plausible values”



Source: Adapted from Zieger et al. (2022, Figure 1).

Clearly, this is a complex process that is not easy to understand. Indeed, the first time this is explained to individuals outside of a quite niche psychometric community, their heads are spinning.

One of the most common questions is why information from the background questionnaires is included at all and whether this is a legitimate thing to do? After all, it means that two pupils who answer exactly the same assessment questions in exactly the same way are likely to receive different “test scores” (plausible values) if their responses to the background questionnaire differ (e.g. if they report being from a higher versus lower socio-economic background or report being a boy instead of a girl). The argument made in the literature is that including background information in the ILSA scaling model “allows a more accurate estimation of student proficiency and helps eliminate bias in group-level comparisons for those grouping variables included in the model” (Bezirhan & von Davier, 2024, p 11.12). Or, as put by Zieger et al. (2022), if background data are not included in the model, then differences between groups will be underestimated.

The logic behind this argument stems from the idea that individual’s ability in each subject is to some extent being “imputed”. In other words, pupil’s ability in maths is not only being based on their answers to maths test questions, but also how they answered questions in other subjects (e.g. reading). The principle of “congeniality” from the multiple imputation literature is thus applied (Jewsbury et al., 2024); when one wishes to explore the correlation or difference between a background variable and ILSA test scores, then this variable should be included in the “imputation” model (i.e. the model used to produce the “plausible values” – i.e. Figure 6.4.1). However, as it is impossible to include all background variables as “conditioning variables” in the model, the principal components derived in step 2 (which retain most of the variation in these data) are used instead.

The process outlined above is clearly not straightforward and, in many ways, feels counterintuitive (despite being well-grounded within the psychometrics literature). Yet reasonable arguments can also be made for adopting alternative approaches (e.g. how would ILSA results change if a simpler model were used, or if no background data were used to derive the “plausible values” at all)? Assessing these alternatives is complicated by the fact that the underlying code sitting behind models such as that depicted in Figure 6.4.1 has not been publicly released. We nevertheless return to this issue in section 6.6 of this chapter, where we review evidence on whether modifying the methods used to generate ILSA scores leads to substantively meaningful changes to the results.

6.5 What are plausible values and how are you meant to use them in your analysis?

As noted in the previous subsection, ILSAs include multiple (usually 5 or 10) “plausible values” per individual per subject, rather than a single test score. Another way to think about this is that each pupil has a distribution of possible scores they could have realistically achieved on the test, with each plausible value being a random draw from this distribution. An example of how these plausible values look in a typical ILSA dataset can be found below (PV1MATH refers to the first plausible value in mathematics; PV2MATH the second plausible value and so-forth). For context, a standard deviation equates to around 100 score points.

Table 6.5.1. An illustration of five plausible values within an ILSA for five pupils

Pupil	PV1MATH	PV2MATH	PV3MATH	PV4MATH	PV5MATH
1	281	298	164	280	319
2	245	280	244	175	216
3	486	449	457	423	427
4	356	284	323	286	237
5	381	342	439	368	307

The variation between these plausible values for each pupil reflects the uncertainty surrounding their proficiency in mathematics. This is often substantial. For instance, pupil 1 in Table 6.5.1 has a proficiency estimate of 164 in the third plausible value compared to 319 under the fifth plausible value – a difference of over 150 points (1.5 standard deviations)! This again highlights why ILSAs cannot be used to provide measures of achievement for individual pupils; there is far too much uncertainty in individual results.

Plausible values are generated for all test takers in every assessed subject, including subjects in which they answered no test questions. Returning to Table 6.3.1, for example, pupils assigned to Booklet 6 completed reading items only. Nevertheless, they are assigned plausible values for reading, mathematics and science. For these pupils, the mathematics and science plausible values are essentially imputed from their performance on the reading assessment and their responses to the background questionnaire. Understandably, for these pupils, there will be substantial uncertainty in their mathematics and science skills, reflected by large differences across the plausible values.

How, then, should these plausible values be used in one's analysis?

As noted in previous sections, plausible values can be understood as multiple imputations of an unobserved latent variable. Analysts therefore follow the well-established procedures from the multiple-imputation literature when combining results across them (Carpenter et al., 2023). In the context of ILSAs, this process can be summarised in the following six steps:

- Step 1. Calculate the statistic of interest using the first plausible value. When doing so, apply the relevant final sampling weight and replicate weights (see Chapters 2 and 3 for further details). For example, if the ILSA provides a final individual sampling weight and 80 BRR weights, one will need to produce 81 estimates using the first plausible value. The sampling variance is then calculated using the appropriate replication-weight formula (e.g. Equation 3.4.1 for BRR together with Fay's adjustment where required). This will result in an estimate of the statistic of interest (β_1) and its sampling variance (σ_1^2) based on the first plausible value.
- Step 2. Repeat Step 1 for each of the remaining plausible values in the dataset. For example, if the dataset contains five plausible values, this will produce five estimates of

the statistic of interest $(\beta_1, \beta_2, \dots, \beta_5)$ together with five corresponding estimates of the sampling variance $(\sigma_1^2, \sigma_2^2, \dots, \sigma_5^2)$.

- Step 3a. Calculate the mean of the point estimates obtained from each of the plausible values to obtain the pooled estimate (β_*) . For instance, with five plausible values:

$$\beta_* = \frac{\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5}{5}$$

This is the final estimate of the statistic of interest.

- Step 3b. Calculate the mean of the sampling variances obtained from each of the plausible values to obtain the pooled estimate (σ_*^2) . For instance, with five plausible values:

$$\sigma_*^2 = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2}{5}$$

This is your final estimate of the sampling variance.

- Step 4. Calculate the measurement-error variance (ε^2) . For each plausible value subtract the pooled estimate (β_*) from the corresponding estimate (β_{pv}) in turn. Square these differences, add them together and divide by the number of plausible values minus one (PV-1). The relevant formula is:

$$\varepsilon^2 = \frac{\sum_{pv=1}^5 (\beta_{pv} - \beta_*)^2}{PV - 1}$$

This is the final estimate of the measurement error variance.

- Step 5. Combine the sampling variance calculated in Step 3b with the measurement-error variance calculated in Step 4. The total variance is obtained using the following formula:

$$\omega^2 = \sigma_*^2 + \left(1 + \frac{1}{PV}\right) \cdot \varepsilon^2$$

Where:

ω^2 = Total error variance.

σ_*^2 = Final estimate of the sampling variance (from step 3b)

PV = Number of plausible values

ε^2 = Final estimate of the measurement error variance (from step 4)

Note that the more plausible values that are drawn, the smaller the multiplier on the measurement error variance. In particular, the term $\left(1 + \frac{1}{PV}\right)$ is included due to ILSAs only including a finite number of plausible values. As the number of plausible values gets sufficiently large, the total error variance essentially boils down to the sum of the sampling and measurement-error variances.

- Step 6. Take the square root of the total variance calculated in Step 5 to obtain the final standard error. This can then be used in subsequent analyses. For instance, to test whether differences between countries are statistically significant, this standard error can be used in the denominator for the relevant country using the two-sample t-test presented in equation 4.3.1 in chapter 4.

The mechanics of the process outlined above are easier to understand through a worked numerical example. To keep this example focused on the use of plausible values we skip steps 1 and 2, particularly as the use of replicate weights to estimate sampling variance was covered previously in Section 3.4. We therefore begin with Table 6.5.2. This contains ten estimates of the average score, ten estimates of the sampling error and ten estimates of the sampling variance. The latter (sampling variance) is simply the former (sampling error) squared²⁸. We include both here, so readers are clear we need to combine the sampling variance with the measurement error variance.

Table 6.5.2. Estimates of population mean scores, BRR sampling errors and BRR sampling variances for a single country

Plausible value	Mean	BRR Sampling error	BRR sampling variance
1	489.242	2.211	4.888
2	489.238	2.273	5.167
3	489.310	2.100	4.411
4	488.413	2.160	4.668
5	488.788	2.192	4.804
6	488.594	2.210	4.886
7	488.778	2.181	4.756
8	489.642	2.195	4.819
9	488.523	2.139	4.574
10	489.222	2.095	4.389

²⁸ Both columns have been rounded independently from unrounded results, so the displayed values may not square exactly

Starting with Step 3a, we calculate the average of the ten mean scores²⁹:

$$\beta_* = \frac{489.24 + 489.24 + 489.31 + 488.41 + 488.79 + 488.59 + 488.78 + 489.64 + 488.52 + 489.22}{10}$$

$$\beta_* = 488.98$$

Then doing the same for the sampling variance under each of the ten plausible values in step 3b:

$$\sigma_*^2 = \frac{4.89 + 5.17 + 4.41 + 4.67 + 4.80 + 4.89 + 4.76 + 4.82 + 4.57 + 4.39}{10}$$

$$\sigma_*^2 = 4.74$$

Next, in step 4 we calculate the squared differences between the estimate under each individual plausible value and the pooled estimate:

$$(489.24 - 488.98)^2 = 0.0713$$

$$(489.24 - 488.98)^2 = 0.0691$$

$$(489.31 - 488.98)^2 = 0.1124$$

$$(488.41 - 488.98)^2 = 0.3155$$

$$(488.79 - 488.98)^2 = 0.0348$$

$$(488.59 - 488.98)^2 = 0.1455$$

$$(488.78 - 488.98)^2 = 0.0388$$

$$(489.64 - 488.98)^2 = 0.4453$$

$$(488.52 - 488.98)^2 = 0.2046$$

$$(489.22 - 488.98)^2 = 0.0608$$

Then sum these differences, and divide through by the number of plausible values (10) minus one:

$$\varepsilon^2 = \frac{0.0713 + 0.0691 + 0.1124 + 0.3155 + 0.0348 + 0.1455 + 0.0388 + 0.4453 + 0.2046 + 0.0608}{10 - 1}$$

$$\varepsilon^2 = 0.1665$$

We can now combine these values together in step 5:

$$\omega^2 = 4.74 + \left(1 + \frac{1}{10}\right) * 0.1665$$

$$\omega^2 = 4.91931$$

(Note that this value was produced using unrounded results - so may not match exactly).

Then we can take the square root to get the standard error:

²⁹ Note that unround values used in calculations, so figures may not match exactly to all decimal places.

$$\omega = \sqrt{4.91931} = 2.217951$$

Fortunately, as noted in the introductory chapter (section 1.7) analysts do not need to carry out these calculations manually when working with ILSAs. The software packages listed there handle the technical work automatically. The purpose of the discussion above is instead to help readers understand how plausible values are treated “under the hood” and what the software is doing on their behalf.

It is also worth noting how small the measurement error variance is in this example (0.1665) relative to the sampling variance (4.74). Indeed, its contribution to the total variance is very modest; omitting this component would reduce the estimated standard error only from 2.22 to 2.18 (a difference of just 2%). Thus, although there is considerable uncertainty surrounding an individual pupil’s proficiency, this does not necessarily translate into equally large uncertainty for a country or group-level statistic. In particular, by combining information across many pupils, population estimates can be much more precise than individual proficiency estimates. Indeed, in this example, the contribution made by the measurement error component is trivial.

6.6 Does it really matter if we use plausible values correctly or not?

Although software packages are available to handle many of the complexities of ILSA analysis, the recommended procedures are not always followed in secondary research. For example, one review found that more than half of the studies published in leading economics journals did not apply the procedures described in the previous section for plausible values or those outlined in Chapter 3 for sampling and replicate weights (Troccoli, 2020). There may be practical reasons for this. Complex statistical models can already take considerable time to estimate. In such situations, correctly incorporating plausible values and replicate weights will greatly increase the computational burden. In PISA, for instance, analysing ten plausible values with one final sampling weight and 80 replicate weights requires the model to be estimated 810 times (just for a single country).

But does it really matter whether recommended practice regarding the plausible values is followed or not?

For those wishing to take a shortcut, the best approach is to use a single plausible value. As the OECD (2009, p. 44) notes “working with one plausible value instead of five [or more] will provide

unbiased estimate of population parameters” [author’s addition]. Using the replicate weights also produces an unbiased estimate of the sampling variance. Thus, the main consequence of using only one plausible value is that the measurement-error variance component is not incorporated into the final standard error.

How much this matters in practice depends on the specific analysis being conducted. In the numerical example above, the measurement error variance was so small that simply ignoring it would make very little difference to the final standard error. Hence using just a single plausible value would, for all intents and purposes, be sufficient. But this is not always the case.

For instance, the TIMSS 2019 technical report provides a useful breakdown of the sampling variance and measurement error (“imputation”) variance associated with average achievement scores (Foy & LaRoche, 2020). For most countries, the imputation variance for overall subject scores in mathematics and science is only around 10% the size of the sampling variance. There are, however, some notable exceptions. For example, for average Grade 8 mathematics achievement in Saudi Arabia, the imputation variance is approximately 30% of the sampling variance. Its contribution also tends to be considerably larger for subdomain scores, such as “number” and “algebra”. For instance, the measurement error (imputation) variance for the average Grade 8 number score in Cyprus is almost as large as the sampling variance (2.22 versus 2.66). In this case, failing to use the plausible values correctly would lead to a nontrivial underestimation of the final standard error. This makes sense; the matrix-sampling design will mean each pupil answers relatively few questions in a specific subdomain such as number, creating greater uncertainty around proficiency in that area. Although this uncertainty is to some extent reduced when estimating group or population-level statistics, it does not become as small as for overall mathematics or science scores.

An interesting example of how the incorrect use of plausible values can affect substantive conclusions is by Danzer and Lavy (2018), as discussed by Troccoli (2024). The original study used data from PISA 2003 and 2006 to investigate the effect of paid parental leave on PISA scores in Austria. Although it found no effect on average, it reported a sizeable positive effect for the sons of highly educated mothers.

The analysis of this subgroup was based on 482 observations and used only the first plausible value. For reading, the original study estimated a positive effect of 33 PISA points. However,

Troccoli (2024) illustrated that using the fifth plausible value instead produced an estimate close to zero. When all plausible values are combined using the recommended procedure, the estimated effect was a statistically insignificant 12 points. Thus, Danzer and Lavy's (2018) conclusion that paid parental leave had a positive effect appears to have resulted from their failure to account properly for the plausible values.

Their analysis was particularly susceptible to this problem because it focused on a relatively small subgroup. With only 482 observations, there were fewer individuals over whom measurement error could be averaged, resulting in greater variation in the estimates across plausible values. The reading results were also likely to have been especially affected because reading was a "minor domain" in PISA 2003 and 2006. Pupils therefore answered relatively few reading items, leading to particularly large levels of measurement error.

An alternative approach to using a single plausible value (which may be intuitively appealing) is to calculate the mean of the plausible values for each individual (i.e. to average across the rows in Table 6.5.1) and then use this "averaged" value in subsequent analyses. Unfortunately, this approach is likely to be even more problematic than using a single plausible value. It too fails to incorporate the measurement error variance into the standard errors. But it also artificially suppresses variation in achievement. This introduces bias into estimates of distributional population statistics such as percentiles, the percentage of pupils achieving above or below a particular threshold and the standard deviation³⁰. Thus, if one wishes to take a shortcut, it is preferable to use a single plausible value. Averaging plausible values within individuals has no clear advantages and risks introducing bias into certain point estimates.

Thus, in summary, I offer the following practical guidance:

- Follow the recommended procedures wherever possible, using one of the software packages described in section 1.7.
- Using a single plausible value will usually be sufficient when estimating country-level statistics and models using overall ILSA scores. When doing so, researchers should recognise that standard errors are likely to be slightly underestimated, though the practical impact may well be negligible.

³⁰ For a weighted mean calculated using the same observations and weights, averaging the plausible values within pupils gives exactly the same point estimate as averaging the separate plausible-value estimates. It does not, however, provide the correct total standard error.

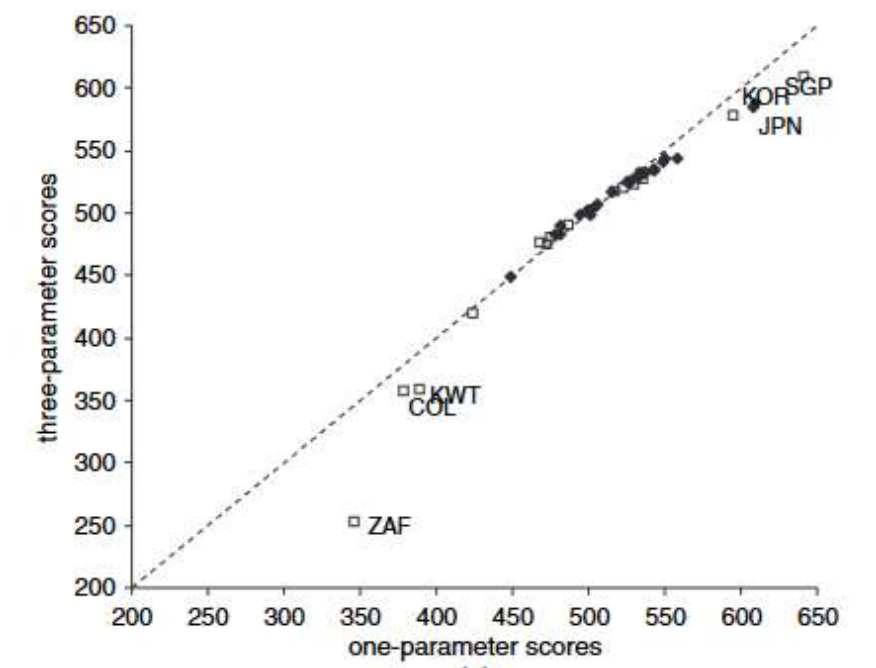
- Exercise caution when using a single plausible value to analyse relatively small subgroups and/or scores in specific subdomains.

6.7 Does using a different IRT or conditioning model make a difference to ILSA results?

Given the complexity of the processes outlined above, a reasonable question is how much particular modelling choices affect ILSA results? For example, does it matter which IRT model is used? Similarly, how sensitive are the results to changes in the background variables included in the model sitting behind the generation of the plausible values (Figure 6.4.1)?

Brown et al. (2007) investigated the importance of IRT model choice using historical data from a period when TIMSS changed from a one-parameter to a three-parameter model. Their results for countries' median scores are presented in Figure 6.7.1. Estimates from the one-parameter model are shown on the horizontal axis, while those from the three-parameter model appear on the vertical axis³¹.

Figure 6.7.1. The country-level association between country maths medians in TIMSS estimated using one-parameter and three-parameter IRT models



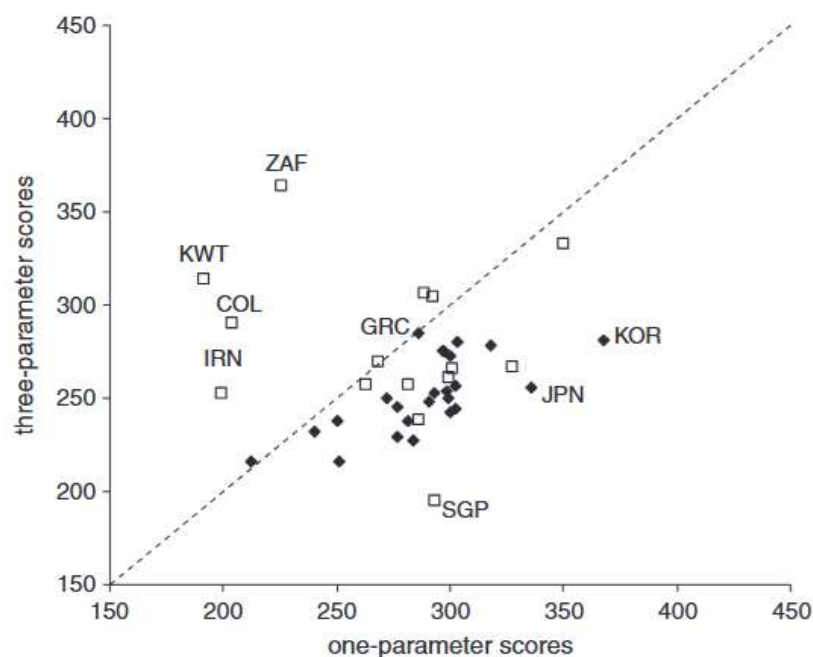
Source: Brown et al. (2007). Figure 4a.

³¹ Note that, although these are presented in terms of differing number of IRT parameters, Brown et al. (2007) also noted the three-parameter versions included a set of conditioning variables (while presumably the one-parameter versions did not).

The two sets of results are strongly related, with a Pearson correlation of approximately 0.97. For the vast majority of countries, the choice of model therefore had little effect on estimated median achievement. There were, however, several nontrivial differences, concentrated mainly amongst countries at the upper and lower ends of the achievement distribution. The clearest example is South Africa (ZAF) for which the estimated median differed by around 100 points (approximately one standard deviation).

Brown et al. (2007) also examined how the choice of IRT model affects estimates of the spread of achievement, measured as the difference between the 95th and 5th percentiles. Their key findings are reproduced in Figure 6.7.2. In contrast to the results for median achievement, the estimated spreads differ substantially between the one-parameter and three-parameter models. For the mathematics results shown, the correlation between the two sets of estimates is effectively zero (at 0.03), though the corresponding correlation for science was higher (0.67). This near-zero correlation in maths is driven almost entirely by the results for non-OECD countries.

Figure 6.7.2. The country-level association between country spread (95th – 5th percentile) in TIMSS maths scores estimated using one-parameter and three-parameter IRT models



Source: Brown et al. (2007). Figure 6a.

Jerrim et al. (2018) conducted a similar investigation using PISA data. They examined several technical choices made within the production of ILSA scores, including the use of one-parameter versus two-parameter IRT models, changes and differences in item-parameters (including country-specific parameters) and whether questions that pupils did not reach were treated as

incorrect responses or as missing data. In contrast to Brown et al. (2007), they found that these decisions had little substantive effect on comparisons of PISA performance across countries. This conclusion held for both average scores and various distributional statistics.

Beyond the choice of IRT model, another aspect of ILSA score generation that has attracted academic debate is the use of background variables in the “conditioning model” (i.e. the model depicted in Figure 6.4.1). Zieger et al. (2022) examined this issue in detail using PISA data. They found for the major domain (the focal subject in a particular PISA cycle) results were largely unaffected by the inclusion or exclusion of different sets of background variables. This makes sense; as pupils generally answer a comparatively large number of questions in the focal subject, the background variables and other aspects of the model have less “work” to do (in the imputation of pupil’s proficiency). However, some statistics in the minor domains (most notably reading) were more sensitive to the background variables included. Specifically, they found that measures of achievement inequality in the minor domains were particularly sensitive to the specification of the conditioning model.

Benton (2019) also investigated the use of background variables, focusing particularly on the application of principal components analysis to reduce the number of variables included in the conditioning model (i.e. step 2 of the process outlined in Section 6.4). He demonstrated how the inclusion of these principal components could potentially lead to bias in estimates of PISA achievement. The OECD (2020) subsequently examined this issue using PISA 2015 data for Northern Ireland, considering how different methods of specifying the conditioning model affected the results. Table 6.7.1 summarises their findings for three approaches (one for the approach initially used in PISA 2015 and two alternatives)³².

Interestingly, the different approaches led to some changes in Northern Ireland’s estimated mean scores. Under one specification, the mean mathematics score was around five points higher than the official PISA 2015 estimate, while the corresponding reading score was approximately seven points higher. In contrast, the estimated mean science score changed very little (likely related to science being the major domain in PISA 2015, with the background variables thus playing a less influential role).

³² We do not discuss the specific differences between these approaches in detail here to avoid becoming overly technical. Interested readers can however consult Annex H of OECD (2020) - pages 3–7 - for a fuller explanation.

Table 6.7.1. Average PISA 2015 scores for Northern Ireland. Alternatives under different ways of specifying the conditioning model.

	Reported	Alternative 1	Alternative 2
Maths	492.8	493.7	497.8
Reading	497.0	504.0	503.9
Science	500.1	500.6	499.1

Source: PISA 2018 technical report OECD (2020, Annex H, Table H2).

The effects were even more pronounced for estimates of the spread of achievement. Under alternative approaches to specifying the conditioning model, the estimated standard deviation increased substantially in all three subjects. In reading, it rose from approximately 84 points to around 102 points. Once again, the smallest change occurred in science, which was the major domain in PISA 2015. These findings are consistent with Zieger et al. (2022), who similarly concluded that distributional measures (most notably inequality in achievement) can be particularly sensitive to the specification of the conditioning model. The OECD (2020) primarily put this difference in results down to estimating the two alternative models on data only from only Northern Ireland (rather than the whole UK), arguing that the inclusion of school IDs as direct regressors was minimal. Interested readers should turn to OECD (2020) for further discussion.

Table 6.7.2. Standard deviations of PISA 2015 scores for Northern Ireland. Alternatives under different ways of specifying the conditioning model to produce the proficiency estimates.

	Reported	Alternative 1	Alternative 2
Maths	77.5	98.4	96.1
Reading	83.8	101.7	102.5
Science	89.8	100.4	99.9

Source: PISA 2018 technical report OECD (2020, Annex H, Table H2).

More recently, Rutkowski and Rutkowski (2025) examined how alternative specifications of the model used to generate ILSA scores affect estimates of achievement distributions for populations and subgroups. Their analysis suggests that the results for the major domain are generally robust across different model specifications, though note that less reliable assessments (e.g. sub-domains) could be more sensitive. However, their study focuses on the major domain within a PISA cycle within a single country (United States), in which each pupil answers a relatively large number of questions.

Thus, to summarise, we believe the literature on the robustness of ILSAs to alternative model specifications currently suggests the following:

- Country mean scores are generally robust to the model changes that have been studied (i.e. similar results have generally been produced regardless of the model used). This is particularly true when the number of test questions each pupil answers in a given subject are reasonably large (e.g. PISA major domains).
- Evidence regarding the sensitivity of cross-country comparisons of distributional statistics (particularly measures of achievement spread) is mixed.
- Distributional statistics in PISA's minor domains (and sub-domains more generally) are likely to be the most vulnerable to change under different model specifications.

6.8 A note on trend items and linking error

As well as facilitating comparisons across countries, another key function of ILSAs is to measure changes in educational achievement over time. This is achieved by including a mixture of newly developed questions and questions retained from previous assessment cycles. The latter are known as “trend items”. These trend items are not released into the public domain, with access to them tightly restricted. This is intended to prevent pupils from becoming familiar with specific test items in advance of sitting the assessment. The trend questions included in the pool of assessment items are distributed across the question clusters described in section 6.3. Since the same questions are administered to comparable samples of pupils in different cycles, they provide a common reference point for linking the achievement scales across years (recall the discussion of “anchor items” in section 5.3). Responses to these trend questions are used within the ILSA scaling process to place results from different cycles on a common metric (i.e. to link the tests). This in turn means that observed changes in scores should be due to changes in achievement rather than simply differences in the difficulty or composition of the assessment.

A process known as concurrent calibration is then used to link achievement scales across ILSA cycles. Although the technical details are beyond the scope of this book, the intuition underpinning it is reasonably straightforward. All responses to the test questions from the most recent assessment are pooled with responses to the test questions from the previous cycle (though only for countries that participated in the ILSA at both time points). The trend items – included in both assessments - provide the bridge that connects the two cycles. IRT item parameters are estimated through this joint calibration and then used to estimate pupils' proficiency distributions in both survey years, following a process similar to that outlined in

section 6.4. Finally, a linear transformation is made which aligns the re-estimated distribution for the previous cycle with its originally reported distribution. The same transformation is then applied to the current-cycle results, placing them on the same scale. While the specifics of this process may vary across ILSAs, similar broad ideas and intuition apply. For further details on this process see Foy et al. (2020) and von Davier et al. (2019).

One technical issue when measuring trends over time (including comparing scores across assessment cycles) is known as linking error. This reflects the uncertainty in placing results from different cycles onto the same scale. Part of the issue is that the link between cycles is based on only a finite number of trend items (i.e. any two tests we wish to equate only have a limited set of common “anchor” questions). Moreover, the measurement properties of these items may also vary between cycles (e.g. specific values of the IRT item parameters).

The technical details of how linking error is estimated are again beyond the scope of this book (interested readers are directed to Martin et al., 2012, for an accessible introduction). Moreover, as with many aspects of ILSAs, the precise details vary across studies. But the basic idea (providing a simplified account of the approach used in TIMSS and PIRLS studies) is as follows.

Say you have two ILSA data from two time points (2023 and 2027). Achievement in the earlier (2023) cycle is estimated twice using the same question response data. The first estimate uses the original (2023) set of IRT item parameters. The second uses an updated set of IRT parameters estimated from data that also include the current (2027) assessment cycle. This produces two estimates of the achievement distribution for the previous cycle, including two estimates of the country mean. The standard error of the difference between these two estimates is used to quantify the linking error.

The discussion above is clearly becoming somewhat technical. Let us therefore turn to what this means in practice when analysing trends over time. In particular, there is an argument that linking error should be incorporated into the standard error when comparing changes in average ILSA scores across assessment cycles. When doing so, analysts should use estimates of the linking error reported in the relevant technical documentation (e.g. see Fishbein et al., 2024, Exhibit 12.20-12.24 for TIMSS 2023 and National Center for Education Statistics, n.d, for PISA)³³.

³³ TIMSS/PIRLS report country-specific linking-error estimates. PISA reports, for each domain and pair of assessment cycles, a common link-error estimate that is applied across participating countries. This

Now return to the two-sample t-test for independent samples discussed in section 4.3 of Chapter 4. When comparing a country's results across assessment cycles, the denominator of the test statistic should be adjusted to incorporate the uncertainty arising from linking the two achievement scales over time. The relevant formula is therefore updated to include a term for the link error ($LE_{1,2}$):

$$T - statistic = \frac{\beta_1 - \beta_2}{\sqrt{SE_1^2 + SE_2^2 + LE_{1,2}^2}}$$

Where:

β_1 = Mean score (or other appropriate statistic) in the country at time point 1.

β_2 = Mean score (or other appropriate statistic) in the country at time point 2.

SE_1 = The standard error of the estimate at time point 1 (estimated applying the replicate weights and plausible values as appropriate).

SE_2 = The standard error of the estimate at time point 2 (estimated applying the replicate weights and plausible values as appropriate).

$LE_{1,2}$ = The value of the link error for the country for the comparison between time point 1 and time point 2.

An illustration of how to apply this formula is available from the National Center for Education Statistics. (n.d.). This compares the change in the mean score for the United States between the PISA 2018 and 2022 cycles. In this instance, the average maths scores in the USA in 2018 and 2022 were 478 and 465, with standard errors of 3.24 and 4.01. The value of the link error for this comparison is 2.24. Plugging these values into the formula above gives a final t-statistic of:

$$T - statistic = \frac{478 - 465}{\sqrt{3.24^2 + 4.01^2 + 2.24^2}} = 2.313$$

If the link error were ignored, the value of the t-statistic would of course be larger (2.52) and the associated p-value smaller.

Interestingly, the international organisations responsible for delivering ILSAs differ in how strict they are in including the linking error in their comparison of average scores over time. The OECD

difference is due to differences in the specific details of how the link error is derived in the respective studies.

typically does include the linking error when comparing trends in average PISA test scores over time, with further discussion of the issue available from OECD (2023, Annex A7). On the other hand *"TIMSS currently does not include estimates for the trend scale linking error in the standard error of the difference for comparing results between assessment cycles. TIMSS 2023 uses an approach consistent with previous cycles to simplify calculations"* (Fishbein et al., 2024, p. 13.8).

6.9 The introduction of adaptive test designs

More recent ILSAs have introduced several new features into their assessment designs. Many of these have been made possible by the transition from paper-based to computer-based testing (recall section 1.6.2 for some of the challenges of this transition). This shift has enabled the use of new and more interactive question formats, but it has also created opportunities to incorporate elements of adaptive testing. Under an adaptive design, at least some pupils are routed toward easier or more difficult sets of questions depending on how well they performed during earlier stages of the assessment. The aim is to present pupils with items that are better matched to their level of skill, while still supporting reliable group and population comparisons.

Consider the second cycle of the Survey of Adult Skills, part of the OECD's Programme for the International Assessment of Adult Competencies (PIAAC). Its literacy and numeracy assessments used a hybrid multistage adaptive design intended to match test difficulty more closely to each participant's knowledge and skills. In stage 1, all respondents first completed the same 16-item locator test, comprising of eight literacy and eight numeracy items. Among those proceeding to the main literacy or numeracy assessment, most followed an adaptive route. In Stage 2, they were more likely to receive either lower-difficulty or higher-difficulty questions according to their performance on the first stage of the test. In Stage 3, they were then routed to lower-difficulty, average-difficulty or higher-difficulty material based on their performance in the previous two stages³⁴. Further details are provided in OECD (2025). Other ILSAs have also introduced adaptive elements, although these do not all operate in the same way. For instance, PISA 2022 used individual multistage adaptive testing in its computer-based mathematics and reading assessments. On the other hand, PIRLS 2021 and the digital component of TIMSS 2023 used group-adaptive designs, in which the mix of easier and more difficult assessment booklets was adjusted at the country level (rather than routing individual pupils).

³⁴ The assessment also included a linear component completed by a randomly selected 25% of participants. For these individuals, performance on the initial 16 questions did not determine their subsequent routing. Instead, they were assigned clusters of questions of an average level of difficulty, which also covered a broader range of item difficulties than those used in the adaptive component.

What are the advantages of adaptive testing over standard "linear" approaches?

First, adaptive testing can make an assessment more efficient by increasing the information obtained at a given test length. In general, items provide the most information when their difficulty is well matched to the participant's skills and there remains substantial uncertainty about whether they provide the correct response³⁵. Repeatedly presenting a highly skilled individual with questions that are far too easy adds relatively little information, as does presenting less-skilled individuals with questions that are far too hard for them. Adaptive testing addresses this problem by directing participants towards appropriately challenging test questions. This in turn makes their responses to the test questions more informative without necessarily increasing the time they spend taking the assessment.

Second, adaptive testing can improve measurement precision across a broad range of achievement levels, particularly at the upper and lower ends of the distribution. In conventional fixed assessments, the available items are often concentrated around the difficulty level of the average respondent. This means the test becomes most informative near the centre of the achievement distribution; it is less precise for the highest and lowest achievers. Adaptive testing attempts to address this issue by assigning questions that are better matched to each participant's ability (or to broader lower-achieving, average-achieving and higher-achieving groups) and thus provide more information across the achievement distribution.

Third, adaptive testing may help sustain participants' motivation in low-stakes assessment settings. Higher-performing individuals may become bored when repeatedly presented with questions that are too easy, while lower-performing individuals may become discouraged or disengaged when questions are consistently too difficult. Since limited test effort is a recognised concern in ILSAs (recall the discussion of this issue in section 1.6.3) presenting participants with questions that are better matched to their proficiency could help reduce boredom and frustration. Admittedly, however, the empirical evidence that adaptive testing improves motivation is mixed (see Frey et al., 2024, for meta-analytic evidence on this issue).

The main implication for readers is that the discussion presented in section 6.3 focused primarily on some of the basics of ILSA assessment design. More recent assessments have introduced additional layers of complexity. Although the broad issues discussed throughout this chapter

³⁵ For a dichotomously scored (right/wrong) item under a simple IRT model, this occurs when the probability of a correct response is around 50%.

continue to apply, these newer designs may create important additional nuances where further research is needed.

6.10 Test your knowledge questions

6.10.1 Task 1

The table below presents data for 10 pupils who have completed a mathematics test. Five “plausible values” have been generated for each pupil in the dataset.

ID	PV1MATH	PV2MATH	PV3MATH	PV4MATH	PV5MATH
1	2	1	3	1	1
2	2	4	2	3	5
3	3	1	1	1	1
4	4	5	4	5	3
5	4	5	3	3	5
6	5	7	5	7	7
7	7	5	5	6	6
8	8	9	10	10	9
9	9	7	6	7	7
10	9	10	8	7	7

Task 1. Question 1. Please use the data in the table to calculate the average mathematics score. Make sure that, when doing so, you handle the plausible values appropriately.

Task 1. Question 2. Please now use the data in the table to calculate the standard deviation of mathematics scores. Make sure that, when doing so, you handle the plausible values appropriately.

Task 1. Question 3. The *incorrect* way to use plausible values is to average them for each individual in the dataset and then calculate the statistic of interest using these averaged values. Using the data presented in the table above, please calculate the mean and standard deviation using this *incorrect* approach. Then please comment on whether/how this incorrect approach leads to biased estimates.

6.10.2 Task 2.

Pupils in a country have completed a reading test, with their responses to the test questions converted into a set of 10 plausible values. The mean score and its standard error estimated using each plausible value in turn can be found in the table below.

	Mean	Standard error
PV1READ	494.5338	2.398543
PV2READ	494.2325	2.26398
PV3READ	494.1211	2.26654
PV4READ	494.4079	2.271027
PV5READ	494.4568	2.299061
PV6READ	495.5099	2.358802
PV7READ	494.2411	2.446005
PV8READ	494.8552	2.334797
PV9READ	493.8338	2.263278
PV10READ	493.8542	2.224371

Task 2. Question 1. Please use the data reported in the table above to create an estimate of average reading test scores in the country.

Task 2. Question 2. Please use the data reported in the table above to create an estimate of the sampling error for the estimate of the mean reading score.

Task 2. Question 3. Please use the data reported in the table above to create an estimate of the measurement error for the estimate of the mean reading score.

Task 2. Question 4. Please use the data reported in the table above to create an estimate of the final standard error (incorporating both sampling error and measurement error) for the estimate of the mean reading score.

Task 2. Question 5. If you calculated the standard error as just the average values of the ten standard errors reported in the table, what answer would this give you? What does this estimate of the standard error miss out and how important is this?

Task 2. Question 6. If you were to produce the estimate using only five plausible values rather than 10, what would likely happen to your estimate of the mean and the standard error? Explain your answer and give an approximate estimate of the range in which you think the mean and standard error might fall.

6.10.3 Task 3.

The table below includes estimates of the standard deviation of achievement in a population, along with the estimated sampling variance. These estimates are provided for each of the five plausible values in the dataset. (E.g. The standard deviation of the first plausible value is 110 with the sampling variance of this estimate being 2).

Plausible value	Standard deviation	Sampling variance
1	110	2
2	90	2
3	100	3
4	110	4
5	90	2

Task 3. Question 1. Using the data provided in the table, please provide your best estimate of the population standard deviation.

Task 3. Question 2. Using the data provided in the table, please provide your best estimate of the sampling variance of this population standard deviation estimate

Task 3. Question 3. Using the data provided in the table, please provide your best estimate of the measurement error variance of this estimate

Task 3. Question 4. Using your answers to the previous questions, please provide a 95% confidence interval for your estimate of the population standard deviation.

Task 3. Question 5. Can you think of a simple way to make this confidence interval narrower without collecting additional data? Why might this be particularly effective in this instance?

Task 3. Question 6. What do you think the consequences would be if a researcher used only a single plausible value in this instance?

References

- Benton, T. (2019). The effect of using principal components to create plausible values. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: 83rd annual meeting of the Psychometric Society, New York, NY 2018* (pp. 293–306). Springer. https://doi.org/10.1007/978-3-030-01310-3_26
- Bezirhan, U., & von Davier, M. (2024). TIMSS achievement scaling methodology: Item response theory and population models. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures)* (pp. 11.1-11.15). Boston College, TIMSS & PIRLS International Study Center. <https://doi.org/10.6017/lse.tpisc.timss.rs8288>
- Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(3), 623–646. <https://doi.org/10.1111/j.1467-985X.2006.00439.x>
- Carpenter, J. R., Kenward, M. G., Bartlett, J. W., Morris, T. P., Quartagno, M., & Wood, A. M. (2023). *Multiple imputation and its application* (2nd ed.). Wiley. <https://doi.org/10.1002/9781119756118>

Fishbein, B., Siegel, P., Atasever, U., Cooney, D., & Gonzalez, G. (2024). Standard error estimation in TIMSS. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report: Methods and procedures* (pp. 13.1–13.10). Boston College, TIMSS & PIRLS International Study Center. <https://doi.org/10.6017/lse.tpisc.timss.rs6950>

Foy, P., & LaRoche, S. (2020). Estimating standard errors in the TIMSS 2019 results. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 14.1–14.60). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch14-estimating-standard-errors.pdf

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 12.1–12.146). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch12-scaling-implementation.pdf

Frey, A., Liu, T., Fink, A., & König, C. J. (2024). Meta-analysis of the effects of computerized adaptive testing on the motivation and emotion of examinees. *European Journal of Psychological Assessment*, 40(5), 427–443. <https://doi.org/10.1027/1015-5759/a000821>

Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How robust are cross-country comparisons of PISA scores to the scaling model used? *Educational Measurement: Issues and Practice*, 37(4), 28–39. <https://doi.org/10.1111/emip.12211>

Jewsbury, P. A., Jia, Y., & Gonzalez, E. J. (2024). Considerations for the use of plausible values in large-scale assessments. *Large-scale Assessments in Education*, 12, Article 24. <https://doi.org/10.1186/s40536-024-00213-y>

Martin, M. O., Mullis, I. V. S., Foy, P., Brossman, B., & Stanco, G. M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 5, 35–47. https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_5/IERI_Monograph_Volume_05_Chapter_2.pdf

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196.

National Center for Education Statistics. (n.d.). *Methodology and technical notes*. U.S. Department of Education, Institute of Education Sciences. <https://nces.ed.gov/surveys/pisa/pisa2022/technical-notes/index.asp>

Organisation for Economic Co-operation and Development (2009). *PISA data analysis manual: SPSS*: OECD Paris.

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 technical report*. OECD Publishing. <https://doi.org/10.1787/9789264167872-en>

Organisation for Economic Co-operation and Development. (2020). *PISA 2018 technical report*. OECD Publishing. <https://www.oecd.org/pisa/data/pisa2018technicalreport>

Organisation for Economic Co-operation and Development. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing.

https://www.oecd.org/en/publications/2023/12/pisa-2022-results-volume-i_76772a36/full-report/comparing-mathematics-reading-and-science-performance-across-pisa-assessments_1c417d6f.html

Organisation for Economic Co-operation and Development. (2025). *Survey of Adult Skills 2023 technical report*. OECD Publishing. <https://doi.org/10.1787/80d9f692-en>

Rutkowski, L., & Rutkowski, D. (2025). The basics of conditioning models: Stability of marginal and conditional achievement to model specification. *Large-scale Assessments in Education*, 13, Article 34. <https://doi.org/10.1186/s40536-025-00270-x>

Troccoli, C. (2020). Comment on “Paid parental leave and children’s schooling outcomes”. SSRN. <https://doi.org/10.2139/ssrn.3744650>

Troccoli C. (2024). Does paid parental leave affect children's schooling outcomes? Replicating Danzer and Lavy (2018). *Journal of Applied Econometrics*, 39(2), 356–362. <https://doi.org/10.1002/jae.3021>

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, 26(4), 466–488. <https://doi.org/10.1080/0969594X.2019.1586642>

Wu, M., Tam, H. P., & Jen, T.-H. (2017). *Educational measurement for applied researchers: Theory into practice*. Springer Singapore. <https://doi.org/10.1007/978-981-10-3302-5>

Zieger, L. R., Jerrim, J., Anders, J., & Shure, N. (2022). Conditioning: how background variables can influence PISA scores. *Assessment in Education: Principles, Policy & Practice*, 29(6), 632–652. <https://doi.org/10.1080/0969594X.2022.2118665>

7. A primer on causal inference using large-scale international assessment data

7.1 Chapter introduction

The main value of ILSAs lies in their use as descriptive tools that help policymakers and researchers understand broad educational patterns. Such findings are often most useful as “question starters” that draw attention to patterns that warrant further investigation. There is, however, growing interest in using ILSAs to answer what are inherently causal questions. This is not where these studies necessarily offer their greatest value. Other data sources (e.g. administrative records, randomised experiments and longitudinal cohort studies) are often better suited to this task.

The aim of this chapter is to examine how ILSAs have been used to address questions that are inherently causal. As elsewhere in the book, the emphasis is on developing intuition rather than providing a comprehensive technical discussion. More detailed guides specifically focused on causal inference methods can be found in texts such as *Mostly Harmless Econometrics* (Angrist & Pischke, 2009). The distinctive contribution here is to consider how commonly used causal methods operate when applied to international data (ILSAs in particular).

The chapter begins by discussing a fundamental limitation of most ILSA datasets; they tend to be cross-sectional and thus usually contain no direct measure of pupils’ prior achievement. We then consider how longitudinal extensions of ILSAs help to move researchers somewhat closer to credible causal inferences via allowing stronger confounder control. This includes an explicit discussion of links between ILSA data and administrative records. The chapter then introduces pupil fixed-effects models, discussing both their strengths and limitations, particularly with respect to their application to ILSAs.

The discussion then moves on to country-level analyses. When doing so, we first consider the use of cross-country scatterplots and explaining why causal conclusions cannot be naively drawn from the associations they display. We then move on to discussing country-level differences-in-differences, which exploit changes in policies or institutions over time. The chapter concludes with a brief introduction to synthetic control methods, which construct a weighted comparison group for a country experiencing a particular reform or intervention.

By the end of the chapter, readers should understand the main ways in which ILSAs have been used to try and move away from correlations and towards cause and effect. In doing so, readers

should be able to recognise the important assumptions, limitations and caveats that this involves.

7.2 The fundamental challenge of ILSA's cross-sectional data

The vast majority of ILSA microdata are cross-sectional at the individual level. This means that information is collected from each pupil at a single point in time. Consequently, many secondary analyses of ILSAs published in the academic literature rely on some form of cross-sectional regression model. These models typically examine how a particular factor or exposure is associated with an outcome after accounting for a set of other observed background characteristics. For example, researchers might investigate whether pupils who attend private schools achieve higher ILSA test scores than those attending public schools, while controlling for differences in family background and other relevant characteristics. When the outcome is continuous (as is usually the case with ILSA achievement scores) the analysis is commonly based on a linear regression model of the form:

$$Y_{ij} = \alpha + \beta \cdot T_{ij} + \gamma \cdot X_{ij} + \varepsilon_{ij} \quad (7.2.1)$$

Where:

Y_{ij} = The outcome of interest (e.g. ILSA test scores).

T_{ij} = The covariate of interest (e.g. whether the child attends a private school).

X_{ij} = A vector of control variables included in the model.

ε_{ij} = Random error term (with standard errors estimated using the relevant replicate weights and plausible values discussed earlier in the book).

i = Pupil.

j = School.

The parameter of interest in such models is β . For a continuous covariate of interest, it represents the expected change in the outcome associated with a one-unit increase in this covariate, conditional on the other control variables (X_{ij}) included in the model. For a binary covariate (e.g. public versus private school), β represents the adjusted difference in the expected outcome between the two categories.

What assumptions are required to interpret the estimated coefficient (β) as causal? Perhaps the most important is the “selection on observables” assumption (i.e. “conditional exchangeability” or the absence of unmeasured confounding). This means that the control variables (X_{ij}) must

account for the pre-existing factors that jointly influence the covariate of interest (T_{ij}) and the outcome (Y_{ij}). Suppose, for example, that we wish to estimate the effect of private-school attendance on pupils' ILSA scores. The model would need to account adequately for all relevant differences between pupils before they entered into private versus public schools. This would likely include (but not limited to) family resources, prior achievement, parental education and school preferences³⁶. All variables included in the model should also be well-measured. For instance, poorly measured confounders may leave residual confounding, meaning that the estimated β coefficient could continue to reflect pre-existing differences rather than capturing the causal effect.

Unfortunately, these assumptions cannot be established conclusively from the observed data. In particular, analysts cannot prove that the controls (X_{ij}) are sufficient to account for all relevant confounding. Their plausibility must instead be justified by researchers using their substantive knowledge, an explicit account of the assumed causal relationships and sensitivity analyses. This is challenging when using ILSA data. Although the background questionnaires collect a wide range of contextual information, time and space are limited and many potentially important confounders are unlikely to be collected. Consequently, selection on observables is usually difficult to defend in conventional cross-sectional regressions using ILSAs.

This does not mean that attempts to use ILSA data for such analyses should be abandoned. Rather, the resulting estimates should be interpreted with appropriate caution, accompanied by clear caveats and sensitivity analyses examining how conclusions change under alternative assumptions and model specifications.

A common challenge in studies using this approach is deciding which variables should be included in the set of controls (X_{ij}). This decision should be guided by the research question at hand and an understanding of the relationships amongst the variables. Broadly speaking, the following principles apply:

- Control for “pre-treatment” confounders. These are variables that help account for noncausal differences in the covariate of interest which are also associated with the

³⁶ Other assumptions required for a causal interpretation would be that there is sufficient overlap between comparable pupils in the public and private school sectors. Likewise, there should be no important measurement error in the variables and the model should be appropriately specified.

outcome. Although the specifics depend on the precise research question, common examples in ILSA include measures of demographic characteristics (e.g. gender, immigrant status, family background)³⁷.

- Do not normally control for consequences of the covariate of interest. These are variables that lie on the causal pathway between the covariate of interest and the outcome. In particular, such variables often capture the mechanism through which the effect of the covariate of interest may operate. Controlling for them can thus remove part of the effect the analysis is trying to estimate. These are sometimes known as “bad controls” (Cinelli et al., 2024).

Although this sounds straightforward, deciding which variables to include in practice often requires considerable judgment. Analysts must weigh two competing risks; omitted-variable bias on the one hand (i.e. where important confounders related to both the covariate and the outcome are excluded) and “overcontrolling” (e.g. including variables on the causal path between the covariate of interest and the outcome) on the other.

Jerrim and Moss (2019) provide a useful illustration of this challenge. Using PISA, they examined how frequently 15-year-olds read fiction and how this was associated with their performance on the PISA reading assessment. Their initial regression models controlled for numerous background characteristics, including socioeconomic status, immigrant status, gender, preschool attendance and grade repetition. Even after accounting for these factors, they found a strong positive association between fiction reading and reading achievement. Yet an important concern remained. Pupils who are higher achievers may be more likely to read fiction frequently. Yet PISA’s cross-sectional design means that this potential confounding factor was difficult to control. Omitted-variable bias was therefore a significant concern, making the selection-on-observables assumption difficult to defend.

In response to this concern, Jerrim and Moss (2019) estimated an additional specification that controlled for pupils’ contemporaneous performance on the PISA mathematics assessment. They argued that mathematics achievement could act as a proxy for broader academic ability and thereby account for some of the confounding arising from higher-achieving pupils also being more likely to read fiction. At the same time, they recognised that this control could attenuate the

³⁷ I note, however, that pre-treatment status alone does not establish that a variable is an appropriate control. In particular, one should avoid conditioning on any pre-treatment variables that could open noncausal pathways.

estimated association if fiction reading itself influenced mathematics performance. Their modelling choice therefore involved a trade-off between reducing omitted-variable bias and potentially introducing overcontrol. Adding mathematics scores reduced the estimated association between fiction reading and PISA reading performance by more than half (from around 17 to 7 points per one-category increase in reading frequency) but the remaining relationship was still positive and statistically significant. The authors nevertheless continued to emphasise that important uncontrolled confounders might remain, meaning their estimates should be interpreted as conditional associations rather than necessarily causal effects.

This example illustrates that, when cross-sectional ILSA data are used to examine the relationship between a factor of interest and an outcome (e.g. test performance), model specification is often as much an “art” as it is a “science”. Analysts should carefully justify which controls are included based on their research question and assumed causal relationships amongst the variables. It is also advisable to show how the estimated association changes as particular controls are added or removed. This is especially important for variables that might reasonably be viewed as “bad controls” (e.g. such as PISA mathematics scores in the case of Jerrim and Moss, 2019, discussed above).

What else can analysts do when working with cross-sectional ILSA data? One potentially valuable approach (which remains uncommon in studies using ILSA data) is to conduct formal sensitivity analyses examining how robust the results are to unmeasured confounding. For instance, Cinelli and Hazlett (2020) have developed a framework that quantifies how strongly an omitted confounder needs to be associated with the covariate and outcome to overturn a research conclusion. This approach also allows analysts to benchmark the assumed strength of unobserved confounding against the explanatory strength of observed covariates. Helpfully, user-friendly software packages in R and STATA are available to implement this approach (Cinelli et al., 2024). Although the technical details are beyond the scope of this book, we encourage readers to consider such methods when assessing the robustness of cross-sectional analyses using ILSAs.

Thus, in summary, our advice to applied researchers using cross-sectional ILSA microdata is as follows:

- Clearly define the question and the effect being estimated. Devote sufficient space in the methodology section to explaining the preferred model specification and the reasoning underpinning the choice of controls.
- Estimate a transparent set of alternatives and theoretically defensible model specifications. Report how the estimate changes when certain controls are included or excluded (particularly potential “bad controls”).
- Use formal methods such as the Cinelli and Hazlett (2020) framework to assess how strongly unmeasured confounding would need to affect the analysis to alter the substantive conclusions. Where necessary, apply such sensitivity analyses even if the software packages do not easily allow users to account for the ILSA complex survey design.
- Discuss the selection-on-observables assumption explicitly, including plausible sources of unmeasured confounding and uncertainty about temporal ordering.
- Be transparent about possible measurement error in the covariate of interest, outcome and control variables. Explain how imperfectly measured confounders could leave residual confounding. Acknowledge that the direction and magnitude of the resulting bias may be difficult to determine.

7.3 Longitudinal international data (including administrative links)

Given that a major limitation of ILSAs is their cross-sectional nature, several research teams have sought to extend these data by adding a longitudinal or panel dimension. This has usually been done in one of three ways:

- Links to administrative or register records. Some countries hold administrative data on young people’s educational background or outcomes. These usually cover most young people in the education system (or at least in the state school sector). While the specific information available will vary across countries, such records often include details of school grades, attendance records, additional demographic information and possibly post-school destinations (e.g. enrolment in higher education). In some countries, these administrative records have been linked to ILSA data. Examples include England (e.g. Jerrim, 2023), Sweden (Borger et al., 2024), Italy (Borgonovi et al., 2026) and Finland (Pulkkinen & Rautopuro, 2022). These linkages are not limited to pupils – e.g. in England TALIS has also been linked to administrative records held on teachers (Sims & Jerrim, 2020). Given the broad population coverage of administrative records, a major advantage

of this approach is that data are usually available for most of the original ILSA participants (as well as potentially those who were originally sampled but did not respond).

- Individual country follow-ups. A small number of countries have used data originally collected through ILSAs as a baseline for subsequent longitudinal follow-up studies. This has effectively turned the original assessment into the first wave of a panel dataset. In Australia this has been done regularly for PISA via its Longitudinal Surveys of Australian Youth, with further examples including Canada's Youth in Transition Survey, Switzerland's TREE study and the Czech PISA-L study. These data allow researchers to examine how assessment and questionnaire data collected during the ILSA predict later educational, labour-market and social outcomes. The main disadvantage, as with any longitudinal survey, is attrition and nonresponse. For example, in the LSAY, about half of ILSA participants have dropped out by wave 4 (Australian Government Department of Education, 2016, Figure 28).
- Coordinated multi-country panel studies. A recent development has been the emergence of coordinated multi-country panel extensions of existing ILSAs. The clearest example is TIMSS 2023 Longitudinal (or TIMSS-L) overviewed previously in section 1.4.3. Nine of the fourth-grade participants were reassessed in Grade 5, with three also participating in the grade 8 to grade 9 longitudinal component. These studies provide genuine panel data, with comparable assessment and questionnaire information collected from the same pupils one academic year apart. Similar data collections are also planned for PIRLS 2026, while the TALIS Video Study (discussed in section 1.4.6) also had a longitudinal design (though with only a short time horizon and being very specifically focused on the teaching of quadratic equations).

Why are these data useful? Returning to the discussion in the previous section, longitudinal ILSA data make it possible to estimate a value-added model specification (Todd & Wolpin, 2003). The key advantage is that analysts can control for pupils' prior achievement, thereby accounting for some of the pre-existing differences in knowledge and skills before the later outcome is measured. This can make estimates more credible than those based on cross-sectional data alone, although it does not eliminate all threats to causal inference. Concretely, the model presented in Equation 7.2.1 becomes:

$$Y_{ij} = \alpha + \beta.T_{ij} + \gamma.X_{ij} + \delta.A_{ij,0} + \varepsilon_{ij} \quad (7.3.1)$$

Where:

$A_{ij,0}$ = A measure of prior achievement.

All other variables are defined as discussed under equation 7.2.1 in the previous section.

Note that, in Equation 7.3.1, ILSA scores may serve either as the prior achievement measure ($A_{ij,0}$) or as the outcome (Y_{ij}) when linked administrative records are available (depending on what additional information those records contain). In individual-country longitudinal follow-ups, ILSA scores usually function as the baseline achievement measure ($A_{ij,0}$) with later educational or labour-market outcomes collected in subsequent waves as the dependent variable. In coordinated multi-country panel studies (e.g. TIMSS-L) ILSA scores can enter as both the pre and post-achievement measures ($A_{ij,0}$ and Y_{ij}).

This arguably makes the key selection-on-observables assumption more plausible. The hope is that controlling for prior achievement absorbs much of the pre-existing variation between pupils that might otherwise confound the relationship of interest. Prior achievement is a particularly powerful control because it is strongly associated with many later educational outcomes (e.g. future test scores) and may also influence the covariate of interest. It does not however eliminate all threats to causal inference, particularly if prior achievement is measured with error or has itself been affected by the covariate of interest. Longitudinal or linked administrative data may also contain additional background information not always available in public-use ILSA datasets (e.g. ethnicity, special educational needs) which may also be included in the vector of controls (X_{ij}) to further support the selection on observables assumption.

Consider again the example in which private secondary school attendance is the covariate of interest (T_{ij}) and ILSA performance is the outcome (Y_{ij}). Young people's knowledge and skills before entering secondary school may influence whether they obtain a place at a private school, particularly where admission involves an entrance test (or other forms of academic selection). In cross-sectional ILSA data, prior achievement is therefore a potentially important unobserved confounder (i.e. it may be related both to private-school attendance and to later outcomes such as end-of-secondary test scores). Having a "pre-treatment" (i.e. prior to entering private school) measure of prior achievement can help address this issue. Yet its effectiveness depends on how well prior achievement is measured, the strength of its relationship with the outcome and whether there is sufficient overlap between pupils attending different school types. In particular, a prior-achievement measure with low levels of reliability may only provide a false sense of security.

At the same time, it is important to recognise that the usefulness of a prior-achievement control depends on the specific research objective. Including such a control partly changes the question being answered. In particular, it shifts the focus towards whether the covariate of interest is associated with the *progress* young people make between the baseline achievement measure ($A_{ij,0}$) and the later outcome (Y_{ij}).

For instance, consider again the example of fiction reading and children's reading achievement studied by Jerrim and Moss (2019) discussed in the previous section. In seven countries, they were able to control for a proxy measure of prior reading achievement (though its quality was questionable – being based on pupils' reports of earlier school marks). Doing so changed the interpretation of their estimates; it now captured the association between fiction reading and PISA reading performance conditional on pupils' reported marks in their previous school report. Any influence that earlier fiction reading may already have had on children's prior achievement would therefore be partly controlled away.

Of course, longitudinal data on ILSA participants offer advantages beyond making the selection-on-observables assumption more plausible³⁸. One advantage is that the outcomes recorded in administrative data may be higher stakes, such as national examination grades. Such measures are likely to be less susceptible to problems such as low-test effort than ILSA scores (recall section 1.6.3 for further discussion). Another is that longitudinal data make it possible to address new research questions. For example, follow-up studies linked to PISA have allowed researchers to examine how achievement and background measures collected at age 15 are associated with later outcomes, including university entry patterns and future earnings (e.g. Piacentini & Pacileo, 2019; Meehan et al., 2023).

What advice would I therefore give to researchers who have access to some form of longitudinal ILSA data?

For many research questions, a useful strategy is to begin with the publicly available international data (covering a large number of countries) to examine the association between the covariate and

³⁸ We also note that panel studies such as TIMSS-L can be used to estimate individual fixed-effects (or first difference models), exploiting the variation that occurs within-pupils over time. Such high-dimensional fixed-effects models are discussed in the following section, though we focus on the way this broad approach has more commonly been applied using ILSAs (exploiting variation in pupils' performance across subjects rather than variation in their performance over time).

outcome of interest using a sensible set of common controls. The advantage of this approach is broad international coverage and considerable statistical power, though the estimates may be particularly susceptible to omitted-variable bias. Researchers can then complement this with country-specific replications where longitudinal data are available. These can be used to show how estimates change once prior achievement is included (e.g. when one estimates a value-added specification) along with any additional relevant covariates available in the linked or follow-up data. Such analyses using longitudinal ILSA data will typically have less statistical power and lower levels of external validity, but with the selection-on-observables assumption becoming more plausible. Combining evidence from the full international dataset with more detailed longitudinal analyses from a single country may thereby provide researchers with the best of both worlds.

7.4 Pupil (high-dimensional) fixed effects models

Another analytical approach that has become increasingly popular in academic studies using ILSA data is the use of pupil fixed-effects models. These belong to a broader class of high-dimensional fixed-effects models that are now widely used in the economics literature (Guimarães & Portugal, 2010).

The intuition behind these models is as follows. In some ILSAs (e.g. PISA, TIMSS), achievement is measured in more than one subject (e.g. mathematics and science). Information on the covariate of interest may also be collected separately by subject (e.g. instruction time devoted to maths and science). Researchers can therefore examine whether within-pupil differences in the covariate across subjects are associated with within-pupil differences in achievement across subjects. The key advantage is that this approach controls for all differences between pupils that are constant across subjects (e.g. family background, ethnicity, general levels of academic ability and school attended). It does so by effectively including a separate dummy variable for each pupil in the dataset. The appeal of pupil fixed-effects models is therefore that they implicitly account for a wide range of otherwise difficult-to-measure confounders.

Formally, such models are specified as:

$$Y_{sij} = \beta.T_{sij} + \gamma.X_{sij} + u_{ij} + \theta_s + \varepsilon_{sij} \quad (7.4.1)$$

Where:

Y_{sij} = The outcome of interest in subject s for pupil i who attends school j.

T_{sij} = The covariate of interest in subject s for pupil i who attends school j .

X_{sij} = A vector of subject-varying controls.

u_{ij} = Pupil fixed effects.

θ_s = A subject fixed effect (accounting for differences in average levels of the outcome across subjects).

ε_{sij} = Random error term.

s = Subject s .

i = Pupil i .

j = School j .

This model can be estimated separately for each country or using pooled data from multiple countries simultaneously. When a pooled international sample is used, country-level influences on the outcome that are common across subjects are absorbed by the pupil fixed effect (as each pupil belongs to only one country).

The parameter of interest in this model is β . This captures how within-pupil differences in the covariate of interest across subjects are associated with within-pupil differences in ILSA test scores across those same subjects. To interpret this estimate causally, a “selection-on-observables” (conditional independence) assumption is still required³⁹. The hope is that including pupil fixed effects makes this assumption more plausible. This is because the fixed effects control for all potential confounders that are constant across subjects for the same pupil. However, the fixed effects do not account for subject-varying confounding (i.e. factors that explain why the covariate of interest varies across subjects for the same pupil and are also related to differences in achievement across subjects). Such factors can be included in the vector of subject-varying controls (X_{sij}) – though obviously only when they are measured in the ILSA (usually in the background questionnaire). The main threat is therefore unobserved subject-varying confounding (including any subject-varying variables that are included as controls but which are poorly measured).

³⁹ Or, more precisely, conditional on the pupil fixed effects, subject effects and any subject-varying controls, the residual within-pupil variation should be unrelated to the residual determinants of achievement.

A common source of confusion when implementing pupil (or other high-dimensional) fixed-effects models for the first time is how the dataset should be structured. In particular, it may not be immediately obvious how it is possible to include a dummy variable for each pupil without eliminating all the variation in the data. To illustrate the basic idea, Table 7.4.1 presents a simple hypothetical example. For brevity, it includes data for just three pupils, uses only one plausible value and ignores additional complications such as sampling weights and replicate weights. The key point is that the dataset is arranged in what is sometimes known as “long format”. In the context of pupil fixed-effects models using ILSA data, this means that each pupil appears in multiple rows (one for each subject). Each row contains the outcome of interest (e.g. the first plausible value in Table 7.4.1), the covariate of interest (e.g. time spent studying in minutes per week) and any subject-varying control variables included in the model. This structure makes clear why including pupil fixed effects is possible; although each pupil has their own dummy variable there are still multiple observations for that pupil across subjects. Variation therefore remains in the data after the pupil fixed effects are included.

Table 7.4.1. A hypothetical illustration of how ILSA data might look before estimating a pupil fixed-effect model.

Pupil	Subject	First plausible value	Time spent studying (minutes per week)
1	Maths	510	240
1	Reading	490	150
1	Science	500	200
2	Maths	623	150
2	Reading	647	170
2	Science	602	210
3	Maths	421	200
3	Reading	453	180
3	Science	399	240

Notes: Only three pupils and the first plausible value illustrated for convenience.

From the discussion so far, pupil fixed-effects models appear attractive. They seem to offer a convenient way of addressing many of the omitted-variable concerns discussed in section 7.2. However, pupil fixed-effects models also have important limitations.

One key example is measurement error. The literature has noted that high-dimensional fixed-effects models can be especially sensitive to measurement error (Angrist & Pischke, 2009; Jennings et al., 2024). The intuition is that fixed effects remove much of the between-unit variation in the data, meaning that the estimate of β is identified only from within-pupil differences across

subjects. If the true within-pupil variation in the covariate is relatively small, random measurement error may account for a larger share of the remaining variation. In other words, the ratio of “noise” to “signal” increases, which can magnify bias due to measurement error. Moreover, this bias does not always attenuate the estimate of β towards zero. For instance, recent work has shown that, in high-dimensional fixed-effects models, measurement error can sometimes increase the risk of false-positive findings (Jennings et al., 2024). Relatedly, because fixed effects remove much of the broader variation in the data, even modest unobserved subject-varying confounding may be sufficient to bias the estimate of β .

This point may be particularly important in the context of the complex ILSA test designs discussed in chapter 6. Recall that ILSA assessments are designed primarily to support reliable population-level estimates across a broad range of content, rather than precise reporting for individual pupils. Yet as each pupil answers only a subset of the available test items, individual proficiency estimates in a given subject have substantial uncertainty. In some cases, pupils may not answer any items in a particular subject area, as illustrated by the rotated PISA booklet design in Table 6.3.1. In such cases, plausible values for the untested subject are generated using pupils’ responses in other subjects, information from the background questionnaire and random imputation error. As Jerrim et al. (2017) note, this raises important concerns for pupil fixed-effects models which rely on within-pupil differences across subjects. If a pupil has answered questions in only one subject, it is unclear how much genuine between-subject achievement variation remains in their plausible values (rather than variation induced by the imputation procedure).

A further issue is how much useful variation remains after the fixed effects have been included. In TIMSS, for example, a pupil fixed-effects model will rely on only two observations per pupil (one for maths and one for science). With pupil fixed effects included, the estimate of β is therefore identified from within-pupil differences between these two subjects. This is a relatively narrow source of variation and is asking quite a lot of the data. As emphasised by Mummolo and Peterson (2018), researchers using fixed-effects models should carefully describe the variation that remains after fixed effects have been applied and consider whether the implied comparisons are substantively plausible. Estimates may also be sensitive to pupils with unusually large within-pupil differences between subjects, with it also potentially unclear how much the covariate of interest can realistically vary within pupils across subjects. For instance, how much does instruction time plausibly vary for the same pupil between mathematics and science?

Despite these limitations, the ability to control for subject-invariant sources of confounding has led to pupil fixed-effects models being used in a growing number of studies using ILSA data. In particular, it has been used to examine how factors such as teacher self-efficacy (Jerrim et al., 2023), between-class ability grouping (Jerrim et al., 2025), instructional time (Lavy, 2015), homework (Jerrim et al., 2020), teaching approaches (Schwerdt & Wuppermann, 2011) and teacher professional development (Kirsten et al., 2023) are related to pupils' achievement. When doing so, authors vary in how strongly they note the strengths and (particularly) the limitations of models including pupil fixed effects. They also differ in the extent to which they present the resulting estimates as causal, rather than as associations that control for more factors than conventional cross-sectional models.

We again conclude with practical advice for researchers considering using pupil fixed-effects models when analysing ILSA data:

- Begin by estimating cross-sectional regressions separately by subject. This should include a recognition that these models may be vulnerable to omitted-variable bias and violations of the conditional independence assumption.
- Clearly state which subject-invariant unobserved confounders the pupil fixed-effects are intended to control. This should include concrete examples (e.g. pupil ethnicity).
- Recognise the limitations of pupil fixed-effects models as well as their strengths. This should include explicit discussion of measurement error in ILSA test scores, the subject-varying covariate of interest and any subject-varying controls.
- Be clear about the source of variation used by the pupil fixed-effects model. This should include how much of the variation the pupil fixed effects explain and how much variation remains after their inclusion. Tools such as the `fedistr` and `sumhdfe` STATA packages (Correia et al, n.d.; Ahrens, 2023; Breuer, M., & deHaan, 2024) can be useful for this purpose.
- Interpret the covariate of interest appropriately. This should include a recognition that fixed effects remove between-pupil variation, changing what should be regarded as a “typical” within-pupil, across-subject difference.
- If the cross-sectional and pupil fixed-effects estimates differ substantively discuss plausible reasons for this difference. In particular, subject-invariant unobserved confounding (addressed by the pupil fixed-effects model) is often only one out of several plausible explanations. Others include, for instance, the effect of measurement error and/or the limited within-pupil variation used to identify the fixed-effects estimate.

7.5. Cross-country scatterplots and country-level correlations

We now turn from analyses of individual-level microdata to the use of country-level data in attempts to draw causal conclusions. This sub-section begins by examining the common practice of presenting cross-country correlations and scatterplots - and considers what such descriptive relationships can (and cannot) tell us about cause and effect.

We have already encountered such examples in this book. Recall, for instance, Figure 1.5.1 which plotted the so-called “Great Gatsby Curve.” This widely cited graph highlights how countries with higher levels of income inequality tend to also have lower levels of intergenerational social mobility. Scatterplots of this kind (and their associated cross-country correlations) are common in international comparative research. For instance, they frequently appear in OECD reports accompanying the release of PISA results (e.g. OECD, 2013, Figure III.4.5), while readers who have seen Andreas Schleicher present findings from ILSAs will be familiar with his extensive use of such figures (and his often bold claims about what such relationships show). Likewise, the influential book *The Spirit Level* (Wilkinson & Pickett, 2009) drew heavily on such cross-country scatterplots and correlations to support its central argument (that income inequality has a negative impact on a wide array of life outcomes).

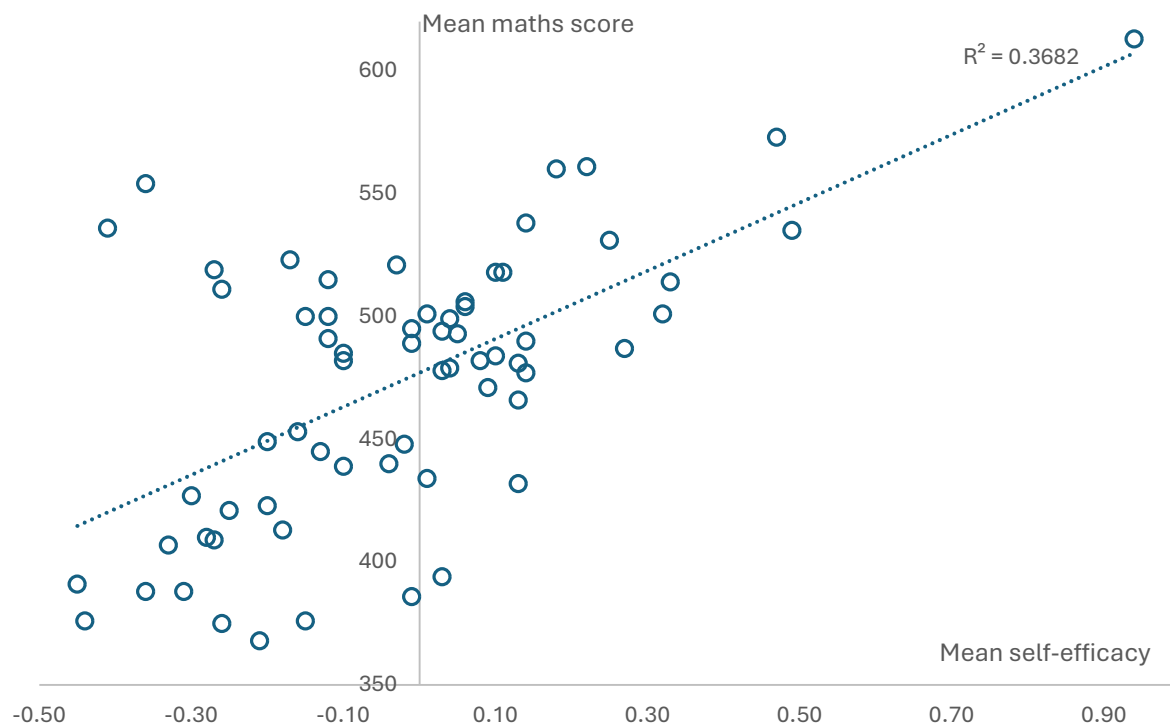
Correlation does not, of course, establish causation. Countries will differ in many ways beyond the two characteristics displayed in a simple scatterplot, which may explain part or all of the apparent relationship. Cross-national analyses also often rely on a relatively small number of countries, meaning that the estimated correlation can be sensitive to a handful of outliers. Researchers and policymakers should therefore be extremely cautious when interpreting such plots. While they can be useful for describing broad cross-national patterns and visually displaying results, they rarely provide convincing evidence that one characteristic of the country is causing the other.

I shall now illustrate this point with what is a (hopefully) fun example.

Let’s begin with Figure 7.5.1, taken from the release of one set of PISA results. The horizontal axis shows each country’s average level of mathematics self-efficacy based on pupils’ responses to questions in the PISA background questionnaire. The vertical axis presents each country’s average PISA mathematics score. A fairly strong positive relationship is immediately apparent - countries where pupils report higher average levels of mathematics self-efficacy also tend to

achieve higher average mathematics scores. The figure proudly reports an R^2 value of 0.37, suggesting that country-level differences in self-efficacy account for 37% of the cross-national variation in mathematics scores. This corresponds to a Pearson correlation of approximately 0.61. Its certainly an eye-catching association.

Figure 7.5.1. The cross-national correlation between average levels of mathematics self-efficacy and average PISA mathematics scores



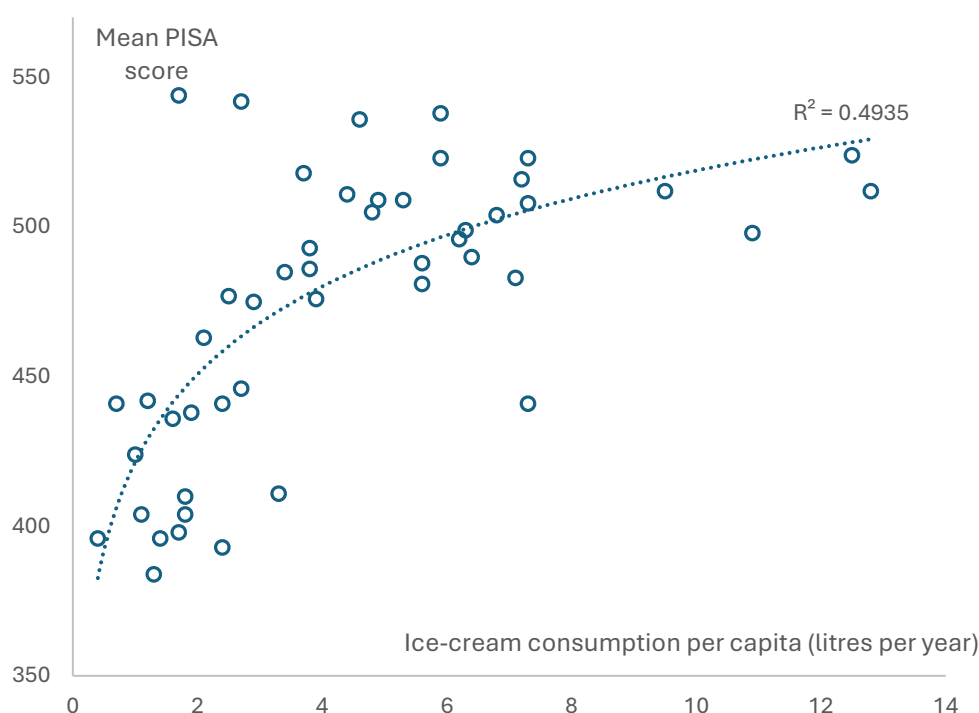
Source: OECD (2013, Figure III.4.5).

Now, taken at face value, it seems plausible that maths self-efficacy may well be linked to mathematics skills. But let's extend this to a different country-level factor. Fish.....

Schmiedel et al. (2017), writing in an epidemiology journal, similarly explored country-level correlates of PISA performance. But they had a rather different explanatory variable in mind - national fish consumption. After controlling for countries' levels of economic development, they reported that fish consumption remained a "significant predictor" of PISA scores. They argue that "this effect is likely due to fish...[being] important for the omega-3 supply to mothers and the early neurodevelopment of their children" [author addition]. The implication, at least implicitly, is that countries may achieve higher PISA scores partly because their populations eat more fish. It is certainly an intriguing idea - and perhaps an appealing one for seafood enthusiasts.

Now, to me, this feels a bit of a stretch..... but I guess maybe there could be something to it....Or perhaps it could be a different food? How about ICE-CREAM!?

Figure 7.5.2. The cross-national correlation between ice-cream consumption and average PISA reading scores



Source: Recreated/adapted from *The Economist* (2016). Plotted values digitised from The Economists public graphic.

And that is precisely what *The Economist* explored in a rather wonderful article. Its graph, reproduced in Figure 7.5.2, plots each country's annual per-capita ice-cream consumption on the horizontal axis against its average PISA reading score on the vertical axis. And - would you believe it—the figure reveals another striking positive relationship. The reported R^2 is 0.49. Thus, statistically, how much ice-cream a country eats can account for almost half of the cross-country variation in PISA reading scores!

Interestingly, when I told my children that they needed to eat more fish to do better at school, they remained decidedly unconvinced. Yet the moment I explained that the evidence apparently pointed instead toward eating more ice cream, their scepticism vanished remarkably quickly!

In all seriousness, the lesson from these graphs and examples should now be clear. Researchers and policymakers must treat causal interpretations of cross-country ILSA results with considerable caution. There are likely to be hundreds, if not thousands, of ways in which

countries differ that could help explain why some perform better in ILSAs than others. In practice, separating these competing explanations - and identifying the effect of any single factor- is extraordinarily difficult.

7.6 Country-level differences-in-differences

We now turn to another way in which country-level data (or potentially regional data within countries) can be used to examine questions attempting to capture cause and effect - country-level difference-in-differences. The broad intuition behind this approach is as follows.

Suppose countries can be divided into two broad groups. In Group A, a particular policy, intervention or structural feature of the education system is introduced within a particular period of time. One example is the tracking of pupils into different secondary schools on the basis of academic ability or potential (which has been studied using difference-in-differences methods with ILSA data as we shall discuss below). In Group B, the same policy, intervention or structural feature is absent, with these countries then acting as a comparison group. The logic of difference-in-differences is to compare how outcomes change in Group A with how they change in Group B, before and after the policy or structural feature takes hold.

A hypothetical example is illustrated in Figure 7.6.1. To keep the explanation simple, we assume data is only available at two time points. Countries are then divided into two groups, as described above. In Group A, a particular policy, intervention or structural feature occurs between time point 1 and 2. In Group B, the same policy, intervention or structural feature does not occur. Our interest is in how this change affects inequality in educational achievement, as plotted on the vertical axis.

The solid lines in Figure 7.6.1 show the observed levels of inequality in educational achievement for the two groups of countries at the two time points. At time point 1, inequality is already higher in Group A than in Group B (0.30 versus 0.15). By time point 2, this gap has widened to 0.60 in Group A versus 0.25 in Group B. The dashed black line represents the assumed counterfactual for Group A (i.e. what we believe would have happened had pupils not been tracked into different schools on the basis of academic ability). Importantly, this assumed counterfactual line runs parallel to the observed trend for Group B. In other words, we assume that, in the absence of the policy or structural feature, the change in inequality for Group A between time points 1 and 2 would have been the same as the change that actually occurred in Group B. In this simple example, the estimated causal effect is the difference between the observed and counterfactual values for Group A at time point 2 (0.20 standard deviations).

Of course, with real-world data, we cannot observe this counterfactual trend for countries in Group A. The difference-in-differences estimate thus needs to be estimated as follows.

First, one takes the difference in the outcome of interest between countries in group A and group B at Time 1:

$$0.3 - 0.15 = 0.15$$

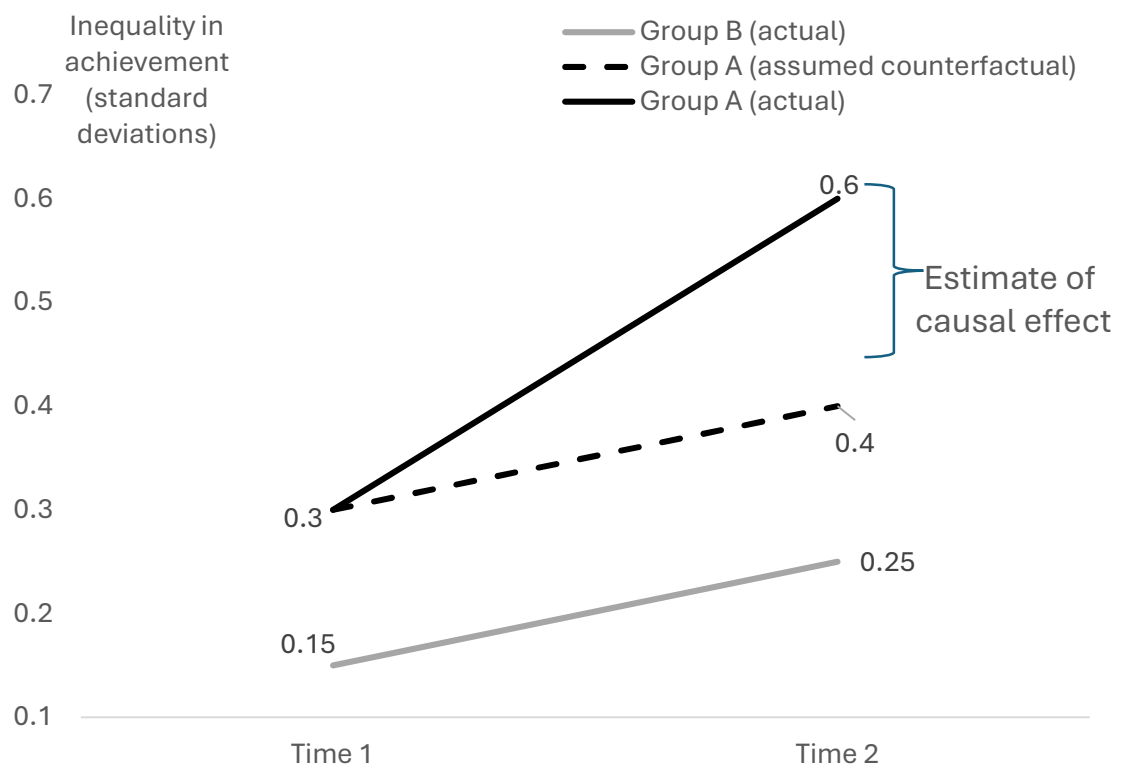
The analogous difference between countries that form group A and group B is then calculated at time point 2:

$$0.6 - 0.25 = 0.35$$

Finally, one then calculates the difference between these values (hence the name “difference-in-differences”):

$$0.35 - 0.15 = 0.2$$

Figure 7.6.1 Hypothetical example of country-level difference in differences to estimate the effect of school tracking on inequality in pupil achievement



Working through the example above makes clear that several assumptions are required if the estimates are to be interpreted as capturing cause and effect. Some of the most important are:

- Parallel (common) trends. In the absence of the policy, intervention or structural feature, the outcome of interest would have followed the same trend in the two groups of countries. This assumption cannot be tested directly as the counterfactual trend for the treated group is unobserved. However, as a proxy, researchers often examine trends in the outcome before the policy, intervention or structural feature came into force. If these pre-trends differ substantially across the two groups, it is often considered less plausible that the common-trends assumption holds. On the other hand, while similar pre-trends may be reassuring, this does not provide definitive evidence that the “common trends” assumption holds.
- Stable composition of the treatment and comparison groups. As ILSAs generally provide repeated cross-sectional samples rather than observations on the same pupils over time, changes in migration, enrolment, exclusions, nonresponse or the demographic composition of the sampled populations may affect the observed outcomes. Of particular concern are compositional changes caused by the policy or institutional feature being studied.
- No spillover effects. The policy, intervention or structural feature in Group A should not affect outcomes in Group B. In the example above, changes to educational inequality in countries where tracking occurs should not influence educational inequality in countries where tracking does not occur. In country-level difference-in-differences analyses, this assumption seems likely to hold.
- No simultaneous confounding shocks. There should not be another policy, intervention or structural change occurring at the same time as the policy, intervention or structural change of interest that differentially affects the two groups. For example, if the introduction of school tracking coincides with other education reforms that affect countries in Group A but not Group B, the difference-in-differences estimate cannot separate out the two effects.
- No anticipation effects. The policy, intervention or structural feature should affect the outcome only after it has occurred. This assumption may be undermined in the context of tracking if, for example, parents in Group A anticipate future academic selection and

respond by investing more time or money in their child's education before tracking takes place.

Despite these somewhat strong assumptions, several studies have nevertheless applied difference-in-differences to ILSAs. Many of these have focused on the issue of school-tracking, as per our example above. These have often drawn data from across multiple ILSAs in order to have data before and after the tracking has taken place.

For instance, in their seminal paper, Hanushek and Wößmann (2006) combined primary and secondary-school evidence from PISA, TIMSS and PIRLS to examine the impact of early tracking on pupil achievement. They report that early tracking increases educational inequality, with some limited evidence of lower mean scores. Subsequent re-analyses, replications and extensions using related difference-in-differences approaches include Jakubowski (2010), Ammermüller (2013), Lavrijsen and Nicaise (2015), Contini and Cugnata (2020) and Strello et al. (2021). Ruhose and Schwerdt (2016) similarly use international large-scale assessment data to study the effect of early tracking on migrant-native achievement gaps, finding no significant overall effect. Hermann and Kopasz (2021) extend this literature to gender gaps in test scores, reporting that early selection increases the gap in favour of girls.

Beyond the literature on school tracking, difference-in-differences approaches have also been used with ILSAs to investigate the impact of a range of other factors. These include school closures during the COVID-19 pandemic (Binkiewicz & Pokropek, 2025, 2026), whether countries require young people to follow a specialised curriculum in upper secondary school (Câmara Leme et al., 2020) and educational reform in Poland (Jakubowski et al., 2011). Ammermüller (2013) also used a country-level difference-in-differences approach to examine school-system features such as instructional time, private schooling and school autonomy.

Despite the number of studies that have used this approach, there are several important challenges in applying difference-in-differences methods to ILSAs. These include:

- Small number of “treated” countries. In studies that apply country-level difference-in-differences, the “treatment” (policy change, intervention or structural feature) occurs at the country-level. Countries therefore provide the key source of variation, rather than individual pupils. This issue is exacerbated by the fact that many applications of this approach combine information from ILSAs over time (with some countries participating in some ILSAs but not in others).

- Comparability across ILSAs. As noted elsewhere in this book, ILSAs differ in several important ways including their target populations, response rates, exclusion rules, assessment frameworks and test designs. Different ILSAs are thus not usually designed to be directly compared with one another. Country-level difference-in-differences studies, however, often require researchers to use them in this way (see Contini & Cugnata, 2020, for a discussion of related issues).
- Inability to compare pre-treatment trends. Data are often not available at enough points to examine whether outcomes were already evolving similarly before the policy, intervention or structural feature occurred. This is particularly important in studies of school tracking using ILSA data, where researchers may have only one pre-tracking and one post-tracking assessment point (as per our hypothetical example in Figure 7.6.1). This leaves researchers with limited empirical evidence to support the key common-trends assumption.
- Other key assumptions discussed above may not hold. Education systems can change in many ways over time or as pupils progress from primary to secondary education. It is hence plausible that other reforms, shocks or structural changes occur at the same time as the policy or feature being studied and differentially affect the “treatment” and “control” countries. For example, countries that track pupils into different secondary schools may also share other institutional features or reforms that do not occur in non-tracking countries. Similarly, as noted above, anticipation effects may be important in the context of tracking if parents, pupils or schools respond before formal selection takes place.

My advice to applied researchers is hence that estimates from country-level difference-in-differences should be interpreted cautiously. In particular, it is important that the limitations and caveats noted above are appropriately discussed.

7.7 Synthetic controls

One limitation of country-level difference-in-differences designs is that they are most straightforward to apply when a similar policy, intervention, shock or structural feature affects several countries at roughly the same time. In practice, however, many policy changes or shocks occur in only a single country (or only a single region within a country). In such cases, there is essentially only one “treated” unit. This makes causal analysis challenging.

In such situations, a natural strategy is to compare trends in the focal country with trends in a set of comparator countries. For instance, suppose New Zealand introduced a major education policy at a particular point in time. One might then compare changes in educational outcomes in New Zealand before and after the policy with changes over the same period in a set of peer countries. Key candidates might include Australia, Canada, the United Kingdom and the United States (i.e. other predominantly English-speaking OECD countries with many cultural, political and economic similarities). In doing so, these countries are being used as potential counterfactuals; what might have happened to educational outcomes in New Zealand had the policy not been introduced. The difficulty is that no single country is likely to provide the ideal comparison.

This is where synthetic control methods become useful. Rather than relying on a single comparator country, this approach constructs a weighted combination of several untreated countries to better reproduce the treated country's pre-policy trajectory. In essence, it creates a weighted average of other countries to form a "synthetic New Zealand". For example, synthetic New Zealand might be composed of 50% Australia, 35% the United Kingdom, 15% the United States and nothing from Canada (zero weight), with these weights estimated from the data available (discussed further below). Before New Zealand introduced its major education policy, the trend in the outcome of interest should be very similar in "real" and "synthetic" New Zealand. Any divergence between the two after the policy is introduced is then interpreted as the estimated effect of the policy.

The intuition above highlights that a key issue in synthetic control methods is how weights are assigned to each country in the donor pool. The technical details are beyond the scope of this book. However, the basic idea is straightforward.

First, we begin with a "donor pool" of potential comparators. This is the set of untreated countries from which the synthetic control will be constructed, with a weight attached to each country. The donor pool could be broad, for example including all countries with the necessary data available. Alternatively, it may be restricted for clear and justifiable reasons. In the example of New Zealand above one might, for instance, restrict the donor pool to members of the OECD. This would make it more likely that "synthetic New Zealand" is constructed from countries with a broadly similar level of economic development.

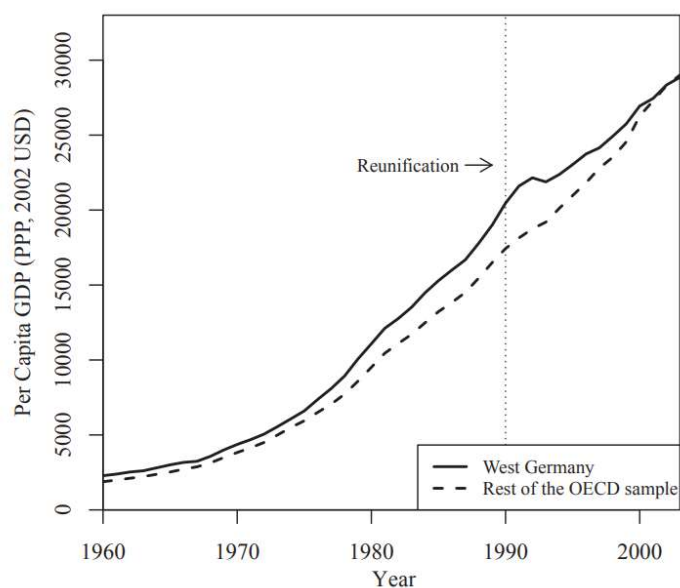
We also need information about each country before the focal country (e.g. New Zealand) introduced its major education policy. The most important information is usually the outcome of interest itself, such as a time series of educational achievement scores before the policy was

introduced. Other country-level predictors may also be included (e.g. GDP per capita, pupil-teacher ratios, educational expenditure). Synthetic control methods then use an algorithm to choose weights for the countries in the donor pool, so that the resulting synthetic control resembles the focal country as closely as possible in the pre-policy period. In standard implementations, the method also determines how much emphasis to place on different predictors, often by selecting the combination that best fits the pre-policy outcome trajectory. In the classic version of the method, donor-country weights are not allowed to be negative (though they can and often are zero for several countries) with the sum of the weights across countries equal to one.

A classic country-level example of this approach, albeit outside education, comes from Abadie et al (2015). This examined the impact of German reunification following the fall of the Berlin Wall. In particular, the authors examined how reunification affected per capita GDP in former West Germany. Figure 7.7.1, drawn from Abadie et al. (2015), compares trends in West German GDP with the rest of the OECD sample over time. The figure shows that West Germany and the OECD comparison group were already following different trajectories before reunification. It would also be possible to compare West Germany with individual peer countries, such as Austria, France or the United Kingdom. However, each of these countries represents a different possible counterfactual and may itself have been on a different trajectory (in the period prior to German reunification). Such comparisons could therefore produce different estimates, highlighting the appeal of constructing a synthetic control rather than relying on a single comparator.

In implementing their synthetic control approach, Abadie et al. (2015) used data for West Germany and 16 potential comparator countries in the donor pool. Their pre-reunification predictors included GDP per capita, the investment rate, trade openness, schooling, inflation and industry share. As is often the case in synthetic control applications, the lagged outcome (GDP per capita) received the largest predictor weight, reflecting its importance in reproducing West Germany's pre-reunification trajectory. The authors then created a "synthetic" West Germany as a weighted average of the 16 donor countries. In practice, Austria was the largest contributor (accounting for 42% of synthetic West Germany), followed by the United States (22%), Japan (16%), Switzerland (11%) and the Netherlands (9%). The remaining 11 countries were assigned a zero weight.

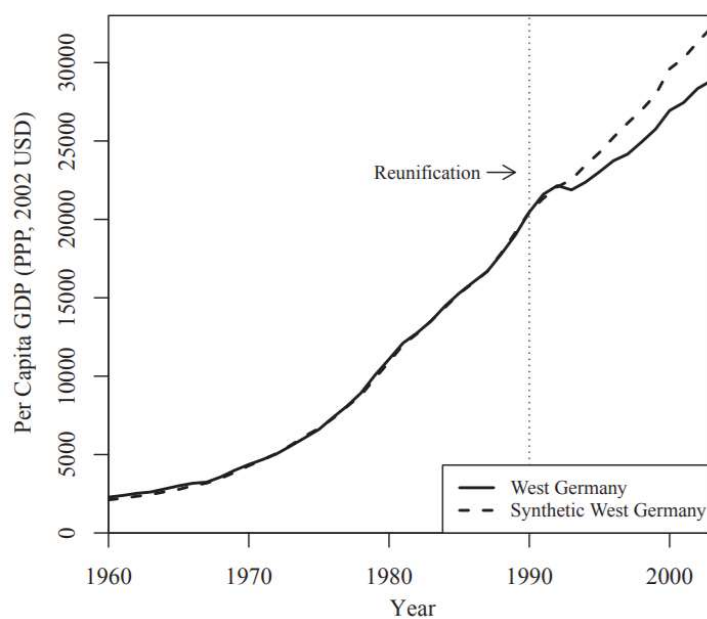
Figure 7.7.1. GDP per capita in West Germany over time compared to the rest of the OECD



Source: Abadie et al. (2015, Figure 1).

Their headline results are reproduced in Figure 7.7.2. Real and synthetic West Germany followed very similar trends in GDP per capita between 1960 and 1990, prior to reunification. After reunification, however, the two trends diverged, with actual West Germany falling behind synthetic West Germany. This counterfactual comparison suggests that GDP per capita in West Germany would have been higher had reunification not occurred.

Figure 7.7.2. GDP per capita in West Germany over time compared to “synthetic” West Germany



Source: Abadie et al. (2015, Figure 2).

It is easy to see the appeal of synthetic control methods for ILSAs and education research more generally. More often than not, education policies and reforms are specific to a particular country or a region within a country. Researchers are therefore often faced with a situation similar to the West German reunification example above (i.e. a single “treated” unit and a set of potential comparators). This is precisely the type of setting for which synthetic control methods are well suited. Despite this, relatively few studies have applied synthetic control methods to ILSA data in practice.

There are, however, some notable exceptions. For instance, Beneito and Vicente-Chirivella (2022) apply synthetic control methods to regional data from Spain, focusing on the effects of school mobile-phone bans. They report reductions in bullying in the treated regions, as well as positive effects on mathematics and science PISA scores in Galicia. Also using PISA data from Spanish regions, Anghel et al. (2015) study the impact of introducing and publishing standardised external test results in Madrid. They find a positive effect on reading performance, but no strong evidence of an effect in mathematics. Soh et al. (2021) use TIMSS data and synthetic control methods to study Malaysia’s switch to English as the medium of instruction for mathematics and science. In their mathematics analysis, synthetic Malaysia is constructed mainly from Tunisia (with a weight of 40%) and Chinese Taipei (35%). They find that the policy had a negative impact on pupils’ TIMSS performance. Finally, Belot and Vandenberghe (2009) apply synthetic control methods to PISA data to study the reintroduction of grade retention in the French-speaking Community of Belgium. They find evidence that grade retention filtered out weaker pupils, but no evidence that it improved their educational outcomes.

Nevertheless, as ILSA data include more time points, users may start to make greater use of synthetic control methods. Our advice to those applying this approach to country or regional level ILSA data is as follows:

- Remember that the approach is, at its core, simply a weighted average across untreated countries or regions.
- Be clear about what the counterfactual actually represents. In the German reunification example, synthetic West Germany combines Austria, the United States, Japan, Switzerland and the Netherlands. The “control condition” is therefore a mixture of what was happening in those countries. Applied to education trends, this can be difficult to interpret. Researchers should at least have a good understanding of what was happening in the countries assigned the largest weights.

- Recognise that the number of time points available in most ILSAs remains limited. TIMSS and PISA have the longest histories but still have data from less than ten cycles at the time of writing. This gives researchers limited information with which to model pre-policy trends.
- Remember that post-intervention time points are also limited and are not collected annually. This makes it difficult to distinguish the effect of the focal policy from later changes that may also affect outcomes.
- Focus on policies or shocks that are large enough to plausibly show up in national or regional achievement trends. Countries often make many education-policy changes over time, so the intervention being studied needs to be substantial enough that its effect can be distinguished from other contemporaneous developments.
- Check the quality of the pre-intervention fit. Synthetic control methods are most convincing when the synthetic unit closely tracks the treated country before the intervention. If the pre-policy fit is poor, the post-policy comparison is much less persuasive.
- Report robustness checks. These might include changing the donor pool, removing heavily weighted countries one at a time, using alternative predictors or testing whether the results depend on one particular comparison country.
- Use placebo or permutation tests where possible. For example, researchers can apply the same method to countries that did not experience the policy change and examine whether similarly large “effects” appear elsewhere.

7.8 Test your knowledge questions

7.8.1. Task 1.

You are interested in the association between time spent on studying mathematics outside of school and mathematics test scores across three countries (A, B and C). This information is available within the PISA dataset, when pupils are age 15. The following OLS regression model is estimated to investigate this relationship:

$$T_{ij} = \alpha + \beta \cdot H_{ij} + \gamma \cdot X_{ij} + \mu_j + \varepsilon_i$$

Where:

T_{ij} = Pupil test scores standardised within countries to mean zero and standard deviation one.

H_{ij} = Number of hours spent studying mathematics outside of school.

X_{ij} = A vector of background controls.

μ_j = School fixed effects (a dummy variable for every school, except for the reference school).

ε_i = Random error term.

Task 1. Question 1. Here is a list of variables available to you, which you might include in the vector X_{ij} . Please state for each variable whether you would include it as a control in X_{ij} .

Remember that your aim is to estimate the effect of time spent studying mathematics outside of school on pupil's test scores.

- Gender
- Ethnicity
- Interest in mathematics
- Socio-economic status
- Test anxiety in mathematics

Task 1. Question 2. How would you interpret the β coefficient from this model?

Task 1. Question 3. The model includes a control for a school fixed effect (μ_j). What do you think are the advantages of including this as a control variable?

Task 1. Question 4. What do you think are the key assumptions you would need to make to interpret the β parameter as capturing the causal effect of hours spent studying outside of school on mathematics achievement.

Task 1. Question 5. In country B, all the children in the country sit a national mathematics examination at age 14, with this information held in administrative records. Please state how linking the PISA data to administrative records in country B might help to

Task 1. Question 6. The table below presents the results after estimating the model for the three countries. A policymaker in country A says the following: *“The evidence shows that the effect of studying outside of school is stronger here than in any other country. We are therefore going to mandate in law that all students are going to be required to study outside of school for at least 7 hours every week. That will raise the performance of our students by more than one standard deviation”*. Please critically evaluate the statement the policymaker has made.

	Estimate	Standard error
Country A	0.15	0.03
Country B	0.04	0.03
Country C	0.10	0.02

7.8.2. Task 2.

Your goal here is the same as in task 1 – to estimate the effect of an additional hours spent studying outside of school on student achievement across three countries (A, B and C). However, you now have information on time spent studying outside of school in three subjects – reading, mathematics and science. You also have student test scores in the same three subjects, though this test used a multiple matrix design, with half of students receiving questions in two or more subjects.

Task 2. Question 1. Please explain how you might try to exploit the fact you now have access to data from across multiple subjects to estimate the causal effect of time spent studying outside of school on student achievement. What model may you now try to estimate?

Task 2. Question 2. Please state what you believe is the main advantage of this approach, as compared to the OLS approach used in task 1.

Task 2. Question 3. Please state at least two limitations – or potential challenges – with this approach.

Task 2. Question 4. What are the implications of the multiple matrix sample design for using this approach? What sensitivity analyses might you run in response?

Task 2. Question 5. What are the assumption(s) are you making to interpret the estimates from these models as capturing cause and effect?

Task 2. Question 6. What additional data would you like access to get closer to a causal effect when using this approach?

7.8.3. Task 3.

The data in the table below presents average test scores for ten countries, with students taking an annual test between ages 7 and 14.

In the first five countries (A to E), children are divided into separate secondary schools based on their performance on an academic selection test. In the last five countries (F to J), a lottery is used to assign students to secondary school places.

Country	Age							
	7	8	9	10	11	12	13	14
A	450	460	470	480	490	500	510	520
B	465	480	495	510	525	540	555	570
C	480	485	490	495	500	505	510	515
D	400	420	440	460	480	500	520	540
E	415	430	445	460	475	490	505	520
F	525	530	535	540	545	550	555	560
G	550	552	554	556	558	560	562	564
H	530	533	536	539	542	545	548	551
I	530	535	540	545	550	555	560	565
J	560	562	564	566	568	570	572	574

Task 3. Question 1. Using just the data at ages 10 and 11, please produce a simple difference-in-difference estimate of the effect of these different ways of allocating students to secondary schools on their academic achievement.

Task 3. Question 2. Can you think of two simple ways that this approach might be extrapolated to utilise all the data presented in the table.

Task 3. Question 3. Please give what you believe to be the two most important limitations of this difference-in-difference estimate with respect to capturing a causal effect. Please explain your answer.

References

- Abadie, A., Diamond, A. & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2), 495–510. <https://doi.org/10.1111/ajps.12116>
- Ahrens, L. (2023). *FEDISTR: Stata module for plotting distributions after absorbing fixed effects* (Statistical Software Components No. S459249). Boston College Department of Economics.
- Ammermueller, A. (2013). Institutional features of schooling systems and educational inequality: Cross-country evidence from PIRLS and PISA. *German Economic Review*, 14(2), 190–213. <https://doi.org/10.1111/j.1468-0475.2012.00565.x>
- Anghel, B., Cabrales, A., Sainz, J. & Sanz, I. (2015). Publicizing the results of standardized external tests: Does it have an effect on school outcomes? *IZA Journal of European Labor Studies*, 4, Article 7. <https://doi.org/10.1186/s40174-014-0029-3>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. <https://doi.org/10.2307/j.ctvc4j72>
- Australian Government Department of Education. (2016). *A review of the Longitudinal Surveys of Australian Youth*. https://www.lsay.edu.au/_data/assets/pdf_file/0029/181973/Review_of_Lsay_2844.pdf
- Belot, M., & Vandenberghe, V. (2009). *Grade retention and educational attainment: Exploiting the 2001 reform by the French-Speaking Community of Belgium and synthetic control methods* (Discussion Paper No. 2009-22). Université catholique de Louvain, Institut de Recherches

Beneito, P., & Vicente-Chirivella, Ó. (2022). Banning mobile phones in schools: Evidence from regional-level policies in Spain. *Applied Economic Analysis*, 30(90), 153–175. <https://doi.org/10.1108/AEA-05-2021-0112>

Binkiewicz, M., & Pokropek, A. (2025). *Estimating causal effects of extended school closures on non-cognitive factors: Evidence from TIMSS and PISA* (Working Paper No. 2025-16). Faculty of Economic Sciences, University of Warsaw.

Binkiewicz, M., & Pokropek, A. (2026). *Beyond weeks closed: Reassessing pandemic educational disruption with TIMSS 2003–2023* (Working Paper No. 2026-21). Faculty of Economic Sciences, University of Warsaw.

Borger, L., Johansson, S., & Strietholt, R. (2024). How representative is the Swedish PISA sample? A comparison of PISA and register data. *Educational Assessment, Evaluation and Accountability*, 36, 365–383. <https://doi.org/10.1007/s11092-024-09438-5>

Borgonovi, F., Ferrara, A., Grindel, C., & Palmerio, L. (2026). Longitudinal insights into academic resilience in PISA in Italy: The role of prior achievement and upper-secondary school factors. *Large-scale Assessments in Education*, 14, Article 25. <https://doi.org/10.1186/s40536-026-00298-7>

Breuer, M., & deHaan, E. (2024). Using and interpreting fixed effects models. *Journal of Accounting Research*, 62, 1183–1226. <https://doi.org/10.1111/1475-679X.12559>

Câmara Leme, A., Escardíbul, J. O., Nunes, L. C., Balcão Reis, A., & Seabra, C. (2020). The effect of a specialized versus a general upper-secondary school curriculum on students' performance and inequality: A difference-in-differences cross-country analysis. *Applied Economics*, 52(39), 4317–4331. <https://doi.org/10.1080/00036846.2020.1734183>

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 39–67. <https://doi.org/10.1111/rssb.12348>

Cinelli, C., Ferwerda, J., & Hazlett, C. (2024). Sensemakr: Sensitivity analysis tools for OLS in R and Stata. *Observational Studies*, 10(2): 93–127.

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, 53(3), 1071–1104. <https://doi.org/10.1177/00491241221099552>

Contini, D., & Cugnata, F. (2020). Does early tracking affect learning inequalities? Revisiting difference-in-differences modeling strategies with international assessments. *Large-scale Assessments in Education*, 8, Article 14. <https://doi.org/10.1186/s40536-020-00094-x>

Correia, S., deHaan, E., & de Kok, T. (n.d.). *sumhdfe: Summary and diagnostic information for evaluating within-fixed-effect variation*. GitHub. Retrieved July 24, 2026, from <https://github.com/ed-dehaan/sumhdfe>

Guimarães, P., & Portugal, P. (2010). A Simple Feasible Procedure to fit Models with High-dimensional Fixed Effects. *The Stata Journal: Promoting Communications on Statistics and Stata*, 10(4), 628–649. <https://doi.org/10.1177/1536867X1101000406>

Jakubowski, M. (2010). Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data. In: Dronkers, J. (eds) *Quality and Inequality of Education*. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3993-4_3

Jakubowski, M., Patrinos, H. A., Porta, E. E., & Wisniewski, J. (2011). *The impact of the 1999 education reform in Poland* (OECD Education Working Papers No. 49). OECD Publishing. <https://doi.org/10.1787/5kmbjgkm1m9x-en>

Jennings, J., Kim, J. M., Lee, J., & Taylor, D. (2024). Measurement error, fixed effects, and false positives in accounting research. *Review of Accounting Studies*, 29(2), 959-995.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutiérrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review*, 61, 51–58.

Jerrim, J., & Moss, G. (2019). The link between fiction and teenagers' reading skills: International evidence from the OECD PISA study. *British Educational Research Journal*, 45(1): 181–200. <https://doi.org/10.1002/berj.3498>

Jerrim, J., Lopez-Agudo, L. A., & Marcenaro-Gutierrez, O. D. (2020). The association between homework and primary school children's academic achievement: International evidence from PIRLS and TIMSS. *European Journal of Education*, 55(2), 248–260. <https://doi.org/10.1111/ejed.12374>

Jerrim, J. (2023). Test anxiety: Is it associated with performance in high-stakes examinations? *Oxford Review of Education*, 49(3), 321–341. <https://doi.org/10.1080/03054985.2022.2079616>

Jerrim, J., Sims, S., & Oliver, M. (2023). Teacher self-efficacy and pupil achievement: much ado about nothing? International evidence from TIMSS. *Teachers and Teaching*, 29(2), 220–240. <https://doi.org/10.1080/13540602.2022.2159365>

Jerrim, J., Lopez-Agudo, L. A., & Marcenaro-Gutierrez, O. D. (2025). Does within-school between-class ability grouping harm the educational outcomes of socio-economically disadvantaged children? International evidence. *European Sociological Review*, 41(5), 722–736. <https://doi.org/10.1093/esr/jcae050>

Kirsten, N., Lindvall, J., Ryve, A., & Gustafsson, J.-E. (2023). How effective is the professional development in which teachers typically participate? Quasi-experimental analyses of effects on student achievement based on TIMSS 2003–2019. *Teaching and Teacher Education*, 132, Article 104242. <https://doi.org/10.1016/j.tate.2023.104242>

Lavrijsen, J., & Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal*, 14(3-4), 206-221.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, 125(588), F397–F424. <https://doi.org/10.1111/ecoj.12233>

Meehan, L., Pacheco, G., & Schober, T. (2023). Basic reading and mathematics skills and the labour market outcomes of young people: Evidence from PISA and linked administrative data. *Economic Record*, 99(327), 473–491. <https://doi.org/10.1111/1475-4932.12755>

Mummolo, J., & Peterson, E. (2018). Improving the Interpretation of Fixed Effects Regression Results. *Political Science Research and Methods*, 6(4), 829–835. doi:10.1017/psrm.2017.44

OECD (2013), *PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs*, OECD Publishing, Paris.

Piacentini, M., & Pacileo, B. (2019). *How are PISA results related to adult life outcomes?* (PISA in Focus No. 102). OECD Publishing. <https://doi.org/10.1787/7b60595e-en>

Pulkkinen, J., & Rautopuro, J. (2022). The correspondence between PISA performance and school achievement in Finland. *International Journal of Educational Research*, 114, Article 102000. <https://doi.org/10.1016/j.ijer.2022.102000>

Ruhose, J., & Schwerdt, G. (2016). Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries. *Economics of Education Review*, 52, 134–154. <https://doi.org/10.1016/j.econedurev.2016.02.004>

Schmiedel, V., Vogt, H., Walach, H. (2018) Are pupils' 'Programme for International Student Assessment (PISA)' scores associated with a nation's fish consumption? *Scandinavian Journal of Public Health* 46(7):675-679. doi:10.1177/1403494817717834

Schwerdt, G., & Wuppermann, A. C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review*, 30(2), 365–379. <https://doi.org/10.1016/j.econedurev.2010.11.005>

Soh, Y. C., Del Carpio, X. V., & Wang, L. C. (2021). *The impact of language of instruction in schools on student achievement: Evidence from Malaysia using the synthetic control method* (Policy Research Working Paper No. 9517). World Bank. <https://documents1.worldbank.org/curated/en/786821611171903155/pdf/The-Impact-of-Language-of-Instruction-in-Schools-on-Student-Achievement-Evidence-from-Malaysia-Using-the-Synthetic-Control-Method.pdf>

Sims, S., & Jerrim, J. (2020). *TALIS 2018: Teacher working conditions, turnover and attrition* (Statistical working paper). Department for Education. https://assets.publishing.service.gov.uk/media/5f6484c28fa8f5107025c17a/TALIS_201_teacher_working_conditions_turnover_and_attrition.pdf

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: Difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability*, 33, 139–167. <https://doi.org/10.1007/s11092-020-09346-4>

The Economist (2016). Ice cream and IQ. 1st April 2016.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485), F3–F33. <https://doi.org/10.1111/1468-0297.00097>

Wilkinson, R., & Pickett, K. (2009). *The spirit level: Why more equal societies almost always do better*. Allen Lane.

8. Data visualisation and interpretation from ILSAs.

8.1 Chapter introduction

ILSAs are large and complex data collections, often involving many countries and (in some analyses) several time points. As a result, even relatively straightforward analyses can generate a substantial number of estimates. It is thus often not possible nor desirable to present every single result. One must instead present key findings in ways that make the central messages immediately clear. When done well, the presentation of ILSA results can be extremely powerful. A carefully designed table or figure can capture attention and make complex findings accessible to a broad audience. Yet it is also easy to get this wrong, or to not pay this matter the due care and attention it deserves. This can lead one to present too many estimates, cluttered graphics and results that are hard to interpret.

The aim of this chapter is to encourage readers to think carefully about how results from ILSAs can be presented in ways that are both transparent and impactful. None of what is discussed is particularly complicated. But applying the ideas does require care and attention. In doing so, we note that this chapter is not intended to be a complete guide to data visualisation – such resources are available elsewhere (e.g. Kirk, 2024). Instead, the distinctive contribution of this chapter is in its explicit consideration of the presentation of international data, particularly from ILSAs.

The chapter begins by examining the widespread use of international “league tables” and explaining why they are often a poor way to present ILSA results (and so must be interpreted with care). It then provides a series of simple, step-by-step examples showing how initially unattractive and difficult-to-read tables can be transformed into cleaner, crisper visualisations. These examples first focus on descriptive statistics before turning to estimates from regression models.

The chapter then applies the same approach to several common graphs used to present results from ILSAs, including scatterplots, line charts and bar charts. In each case, I begin with an exaggerated example of how the graph should not be constructed, before gradually improving it to demonstrate how design choices can make the main message easier to identify and understand. The discussion is not tied to any particular software package. Everything covered can be implemented using widely available tools such as STATA, R, Python or even Excel.

Thus, by the end of the chapter, readers should have developed a practical understanding of the key issues to consider when presenting results from ILSAs and other international comparative data. In particular, they should have developed a better understanding of how to communicate international comparative results clearly and effectively without oversimplifying, exaggerating or misleading their audience.

8.2. International league tables. Strengths and weaknesses.

One way in which ILSA results are reported - at least in the media - is through “league tables” and country rankings. An example, taken from the BBC News website, is presented in Figure 8.2.1. In this, the United Kingdom is the focal country, with the main information presented focusing on its ordinal position relative to other education systems. Such rankings are commonly used in the media and cited by politicians when ILSAs are used for “benchmarking” purposes. For instance, in Ireland, the Minister for Education at the time of the publication of the PISA 2022 results noted how *“we have managed to move from twenty-first to eleventh place in the listings for mathematics”* (Department of Education, 2023). The Irish government website then proceeded to present Ireland in a ranking, highlighting the number of rank positions it had risen since the 2018 cycle.

In general, presenting and interpreting ILSA results in terms of country rankings is not advisable. There are at least three major problems with doing so:

- Small differences in achievement can lead to large differences in rank position. It is often the case that several countries are separated by only a small number of ILSA points. For instance, in PISA 2022 mathematics, only five points (around 0.05 standard deviations) separated a cluster of countries between 12th and 20th position. Such small differences in scores may not represent genuine differences in young people’s mathematics skills across countries; rather this could just be due to sampling variation or relatively minor differences in non-sampling errors (e.g. exclusions and response rates) as discussed in chapter 2. This may then also lead to apparent swings in rank positions over time even when there has been no genuine change in the outcome under investigation. For example, a country ranked around 12th in one cycle could plausibly fall several positions in the next because of sampling variation or modest changes in non-sampling error.

Figure 8.2.1. Example of how the BBC displayed results from the OECD PISA study
How the UK compares in reading over time

	2009	2012	2015	2018
1	Shanghai	Shanghai	Singapore	China
2	South Korea	Hong Kong	Hong Kong	Singapore
3	Finland	Singapore	Canada	Macau
4	Hong Kong	Japan	Finland	Hong Kong
5	Singapore	South Korea	Ireland	Estonia
6	Canada	Finland	Estonia	Canada
7	New Zealand	Ireland	South Korea	Finland
8	Japan	Taiwan	Japan	Ireland
9	Australia	Canada	Norway	South Korea
10	Netherlands	Poland	New Zealand	Poland
11	Belgium	Estonia	Germany	Sweden
12	Norway	Liechtenstein	Macau	New Zealand
13	Estonia	New Zealand	Poland	US
14	Switzerland	Australia	Slovenia	UK
15	Poland	Netherlands	Netherlands	Japan
16	Iceland	Belgium	Australia	Australia
17	US	Switzerland	Sweden	Taiwan
18	Liechtenstein	Macau	Denmark	Denmark
19	Sweden	Vietnam	France	Norway
20	Germany	Germany	Belgium	Germany
21	Ireland	France	Portugal	Slovenia
22	France	Norway	UK	Belgium
23	Taiwan	UK	Taiwan	France
24	Denmark	US	US	Portugal
25	UK	Denmark	Spain	Czech Republic
26	Hungary	Czech Republic	Russia	Netherlands
27	Portugal	Italy	China	Austria
28	Macau	Austria	Switzerland	Switzerland
29	Italy	Latvia	Latvia	Croatia
30	Latvia	Hungary	Czech Republic	Latvia

Note: China score calculated using results from four provinces only - Beijing, Shanghai, Jiangsu and Zhejiang

Source: OECD

BBC

Source: <https://www.bbc.co.uk/news/education-50563833>

- Countries change their participation in ILSAs over time. Second, not all countries participate in every ILSA cycle. Participation in a given study can change over time. Take Egypt's participation in the TIMSS Grade 8 study. This country took part in 2003 and 2007, did not take part in 2011, participated again in 2015 and 2019, and then did not participate in 2023. Such changes can affect the rank position of other countries. Suppose, for example, that nothing else changed between cycles (the same other countries participated and their average scores remained exactly the same). Countries that performed below Egypt would still move up or down by one rank position simply because Egypt entered or left the comparison. This illustrates that rankings depend not only on performance, but also on the set of education systems participating in a particular cycle.
- Different ILSAs include a different set of countries. Third, different numbers of countries participate in different ILSAs. PISA has generally included very broad participation among OECD members, as well as many partner countries. Other ILSAs (e.g. TIMSS, PIRLS) include only the subset of countries that choose to take part. In addition, more low-

income and middle-income countries have joined several ILSAs in recent cycles. Consequently, a country's rank position can differ across ILSAs because the set of participants differs.

This point can be illustrated with a simple hypothetical example. Table 8.2.1 presents data from an ILSA at two time points. At time point 1, ten countries participated, while at time point 2 only eight countries participated. For each country, the table reports their rank position and their average score. Assume that the standard deviation of the score scale presented is around 100 points and that the standard error around each country mean is approximately 2.5 points.

Table 8.2.1. Hypothetical ranking of countries across two ILSA cycles.

Time 1			Time 2		
Country	Rank	Score	Country	Rank	Score
A	1	550.0	A	1	580.0
B	2	545.0	B	2	570.0
C	3	540.0	C	3	565.0
D	4	534.1	J	4	560.8
E	5	534.0	E	5	560.2
F	6	533.8	G	6	559.9
G	7	533.4	D	7	559.4
H	8	533.2	I	8	489.9
I	9	532.9			
J	10	532.7			

Consider first Country J, highlighted in green. At time point 1, it sits at the bottom of the ranking (10th position). However, its average score is only very slightly below that of Country D which is ranked 4th (532.7 compared to 534.1). This is a difference of less than two points (i.e. less than 0.02 standard deviations) and is well within the range of sampling error. Yet it produces a large difference in their rank position.

Yet, by time point 2, Country J has risen in the ranking to 4th place. But while its average score has increased, so too have the scores of most other countries. It thus remains very close to Country D ranked 7th (560.8 compared with 559.4). The apparent rise of country J from 10th to 4th is likely to therefore reflect a combination of fewer participating countries, sampling variation and minor differences in non-sampling error across cycles, rather than a clear improvement in performance relative to other nations.

Country I provides a different example. At time point 1 it is ranked 9th, though with less than a two-point difference in average scores from Country J (10th) and Country D (4th). At time point 2, Country I then appears to have risen from 9th to 8th position. If one looks only at the rankings, this might seem like good news (or at least not bad news). In reality, however, its average score

has fallen by more than 40 points. Moreover, the average score of every other participating country has increased over the same period. This substantial deterioration in Country I over time is therefore masked by the simple use of rankings, again illustrating why rank positions can be misleading when interpreted in isolation.

The advice for government officials, applied researchers and journalists is therefore straightforward. Do not rely on rankings for benchmarking. They may be easy to understand, communicate and to turn into headlines. Yet they can also give a misleading impression of differences between countries and of changes in performance over time. Benchmarking should instead focus primarily on score estimates while also indicating their uncertainty due to sampling and measurement errors (e.g. via confidence intervals). Potential non-sampling errors (e.g. exclusions, nonresponse, coverage limitations and comparability issues) should also be noted.

8.3. A walk-through example of how to create beautiful tables.

Presenting results in tables is very common in international comparative studies. One of their main advantages is efficiency; a well-designed table can communicate a large amount of information within a relatively small space. The trade-off is that tables can be difficult to read if they are poorly organised or contain too much detail. In such cases, the main message may be hard to decipher. Tables should therefore be treated as an important part of the data visualisation toolkit for ILSAs, rather than simply as a place to report numbers. Effective tables require careful decisions about what information to include, what can be left out (or put in an appendix) and how the results should be arranged. When designed with the reader in mind, tables can actually communicate certain results more effectively than a graph.

When presenting ILSA results in the form of a table, we offer the following advice to government officials and applied researchers:

- Make it self-contained. Readers should be able to understand the broad findings without any further information.
- Focus on essentials. Include only what is needed to communicate the main message. Additional results and technical details can be placed in an appendix.
- Remember less is more. Consider whether results for every country are necessary or whether a smaller, carefully selected set would communicate the point more effectively.
- Choose a logical order. Sort rows, for instance, in ascending or descending order by one of the columns. Only use alphabetical order when it is genuinely the best option.

- Informative title. Keep the title to the table as short (but as informative) as possible.
- Keep the design clean. Excessive borders and grid lines can make a table appear crowded and distract attention from the results that matter most.
- Use full country names. Abbreviations such as ISL or MNE may be unclear to many readers.
- Use colour selectively. Conditional formatting can highlight important patterns and comparisons.
- Add clear table notes. Explain definitions, symbols, interpretation and statistical significance so readers are not left uncertain about how to interpret the results.
- Provide useful benchmarks. Add an appropriate international average or comparison point at the bottom of the table to place results for individual countries into context.
- Show uncertainty where possible. Providing standard errors, confidence intervals and significance markers helps articulate uncertainty due to sampling variation and measurement error in ILSA test scores. However, they also add a lot of extra information to the table (and so should be included with care).
- Avoid erroneous precision. Do you really need to report results to 17 decimal places? Think carefully about the level of precision that offers an honest overview of the results without adding extra numbers to the table for little gain.
- Seek AI feedback. Ask an AI tool to review a draft table and suggest improvements to its structure and readability.
- Test with non-specialists. Show the final table to a non-specialist and ask them to describe its main findings back to you.

To illustrate these points in practice, we draw on an example from Zieger et al. (2022). Returning to Chapter 6 (section 6.7) recall that their study examined how sensitive estimates of inequality in PISA reading scores are to the inclusion of different conditioning variables in the scaling model. The authors compared results across countries using eight different model specifications. This necessarily produced a large volume of information, which they wanted to present clearly on a single page. It therefore serves as a useful example of the challenge involved in presenting many related country-level estimates at once, while still making the key findings easy for readers to understand.

An initial presentation of their results is provided in Table 8.3.1. In particular, it illustrates eight different estimates of inequality in reading scores (model specifications M0 to M7). Greater values indicate more educational inequality.

Table 8.3.1. An illustration of a poorly presented ILSA results table.Table. Results

	M0	M1	M2	M3	M4	M5	M6	M7
ALB	217	258	264	257	266	259	264	268
ARG	200	233	237	236	237	236	239	240
AUS	223	240	239	240	251	256	249	258
AUT	209	208	234	236	235	231	242	240
BEL	229	209	187	223	216	226	226	232
BRA	184	216	224	216	224	217	223	223
BUL	273	309	310	310	311	311	310	310
CAN	205	215	243	189	231	200	229	226
CHE	204	235	240	231	243	234	239	241
TPI	212	182	177	139	180	176	182	186
CHL	174	128	127	125	139	149	148	150
COR	149	182	188	181	192	180	190	190
HRV	193	219	240	216	228	215	224	223
CZE	200	218	226	220	222	217	224	221
DEU	210	198	216	203	216	216	225	230
DNK	191	190	176	131	196	193	204	209
ESP	199	226	231	227	233	227	233	234
EST	179	172	192	138	200	164	201	202
FIN	213	239	259	235	254	236	251	249
FRA	240	247	260	258	253	255	261	274
GRC	218	250	257	252	255	251	262	259
HK	192	201	202	202	206	201	210	211
HUN	208	191	205	203	191	200	204	214
IND	155	188	198	188	198	189	198	198
IRL	193	175	193	186	189	189	213	201
ISL	216	242	253	241	250	243	249	249
ISR	239	216	300	247	265	227	298	271
ITA	214	245	252	245	253	245	252	251
JOR	195	230	234	228	235	230	234	234
JPN	223	270	277	273	253	263	268	231
KAZ	156	184	195	185	193	186	195	192
LAT	182	210	214	208	217	210	217	218
LTH	197	222	228	223	228	222	229	230
LUX	240	268	276	269	277	269	277	277
MCU	181	208	208	210	211	210	212	212
MAL	184	213	215	217	215	216	217	217
MEX	161	191	195	192	196	193	197	197
MON	203	237	236	237	237	238	239	239
NLD	215	237	245	240	241	239	245	242
NOR	220	204	145	165	187	231	192	244
NZL	238	266	270	265	271	266	270	272
PRE	194	226	227	229	229	229	231	232
POL	196	240	212	209	212	213	212	212
PRT	214	213	206	188	224	205	200	219
QAT	251	286	284	284	286	285	285	286
ROM	202	234	237	232	238	234	237	239
RUS	201	271	275	280	274	272	279	273
SGP	232	220	206	213	218	240	215	230
SVK	240	262	258	256	263	261	254	272
SVN	213	231	264	242	258	204	265	259
SWE	231	230	244	234	245	235	258	243
THL	173	196	203	201	204	199	207	201
TUN	190	227	229	226	232	227	229	232
TUR	190	216	234	216	224	218	223	221
UK	213	240	240	236	240	241	239	239
UAE	214	168	164	185	171	172	192	202
URG	204	239	238	238	248	240	240	243
USA	207	210	192	174	187	168	183	183
VNM	161	183	189	183	187	180	188	186

This is, of course, an exaggerated example. But it makes the problem clear. All the results are technically there, yet the table is unattractive and the main story hard to identify. It hence does little to support the aims of the study or communicate the findings effectively. In particular, it is not easy to see how estimates of educational inequality change as different sets of conditioning variables are introduced, or whether some countries are particularly affected.

What could be done to improve the table and help it meet the authors' aims more effectively? Below I suggest a series of possible steps, starting with the easiest "wins" first:

- Step 1. Improve the title. The original table is simply titled "Table. Results". This gives the reader no useful indication of what it contains. Changing this to something such as "Estimated P90–P10 gaps in PISA 2012 reading scores, by country and conditioning-model specification" provides a much clearer introduction to the results being presented.
- Step 2. Make the column headings bold. In the original table, the column headings use the same font as the data below them. This makes it difficult to distinguish the labels from the results. Using bold text for the headings helps readers identify the structure of the table more quickly.
- Step 3. Use full country names. Abbreviations make the table harder to read. Although they save space, some codes may not be immediately recognisable (e.g. TPI, HRV, SVN). Using full country names makes the table clearer, quicker to read and less ambiguous.
- Step 4. Sort the countries meaningfully. Alphabetical ordering is not especially helpful in this example because the main aim is not to locate the result for a particular country. Sorting the table by one of the columns instead (e.g. M0) organises the results in a way that is relevant to the research question. Countries near the top will then tend to have lower levels of inequality in reading scores under specification M0.
- Step 5. Remove the grid lines. The large number of lines makes the table look cluttered and does not add useful information. Removing them creates more white space, gives the table a cleaner appearance and makes the results easier to read.
- Step 6. Remove some countries. The changes made so far improve the appearance of the table without removing any information. However, the table still contains results for 59 countries, which is a substantial amount to present on one page. As the study focuses on comparisons across OECD countries, the partner countries could be moved to an appendix. This involves sacrificing some potentially useful information and so is a judgement call. In this instance, I believe the gains in clarity are likely to outweigh the losses.

- Step 7. Add conditional formatting. The main research question concerns how comparisons of educational inequality change as the set of conditioning variables changes. Conditional formatting can provide a quick visual summary of this. Within each column, colours could run from green to red to represent lower and higher levels of inequality. Reading across a row would then show whether a country's relative position changes across model specifications. This also makes it easier to identify countries whose results shift substantially depending on the conditioning variables included.
- Step 8. Add summary measures. An OECD average can provide a useful reference point for each model specification, helping readers see whether a country is above or below the average level of inequality in reading scores. It may also be useful to add a row showing the correlation between each specification and M0 (the basic model with no conditioning variables) and another row for the correlation with M7 (the model including the most extensive set of conditioning variables). These summary measures provide a direct indication of how sensitive country comparisons are to the conditioning model used to generate the PISA scores.
- Step 9. Add table notes. The table contains technical information that may not be immediately clear. Readers may need to know which measure of educational inequality is being used, what the shading represents and whether the shading should be read horizontally or vertically. The addition of table notes can thus improve understanding and reduce the risk of misinterpretation.

The final version is presented in Table 8.3.2 and is much more visually appealing. It is now easy to identify countries with relatively low levels of inequality in reading scores (e.g. Chile) and those with relatively high levels (e.g. Luxembourg) regardless of the model specification used. The conditional formatting also draws attention to countries such as Norway, where inequality appears low under some specifications but closer to the OECD average under others. More generally, the table gives an immediate sense of how stable the results are across the different models. The cross-country correlations are mostly between around 0.7 and 0.8, suggesting that the broad pattern remains reasonably similar, although some non-trivial differences remain. The conditional formatting reinforces this point by acting as a simple heat map, allowing changes in countries' relative positions to be identified quickly.

Table 8.3.2. Estimated P90–P10 gaps in PISA 2012 reading scores, by country and conditioning-model specification

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	161	191	195	192	196	193	197	197
Chile	174	128	127	125	139	149	148	150
Estonia	179	172	192	138	200	164	201	202
Turkey	190	216	234	216	224	218	223	221
Denmark	191	190	176	131	196	193	204	209
Ireland	193	175	193	186	189	189	213	201
Poland	196	240	212	209	212	213	212	212
Spain	199	226	231	227	233	227	233	234
Czech Republic	200	218	226	220	222	217	224	221
Switzerland	204	235	240	231	243	234	239	241
Canada	205	215	243	189	231	200	229	226
United States	207	210	192	174	187	168	183	183
Hungary	208	191	205	203	191	200	204	214
Austria	209	208	234	236	235	231	242	240
Germany	210	198	216	203	216	216	225	230
Finland	213	239	259	235	254	236	251	249
Slovenia	213	231	264	242	258	204	265	259
United Kingdom	213	240	240	236	240	241	239	239
Italy	214	245	252	245	253	245	252	251
Portugal	214	213	206	188	224	205	200	219
Netherlands	215	237	245	240	241	239	245	242
Iceland	216	242	253	241	250	243	249	249
Greece	218	250	257	252	255	251	262	259
Norway	220	204	145	165	187	231	192	244
Australia	223	240	239	240	251	256	249	258
Japan	223	270	277	273	253	263	268	231
Belgium	229	209	187	223	216	226	226	232
Sweden	231	230	244	234	245	235	258	243
New Zealand	238	266	270	265	271	266	270	272
Israel	239	216	300	247	265	227	298	271
France	240	247	260	258	253	255	261	274
Luxembourg	240	268	276	269	277	269	277	277
Slovakia	240	262	258	256	263	261	254	272
OECD mean	211	222	229	218	229	223	233	234
OECD median	213	226	239	231	235	227	239	239
Correlation with M0	1.00	0.71	0.63	0.71	0.73	0.77	0.74	0.83
Correlation with M7	0.83	0.80	0.79	0.82	0.91	0.88	0.89	1.00

Notes: Figures illustrate how the difference between the 90th and 10th percentile of PISA reading scores change across model specifications. Colours produced separately within columns; green shading indicates less inequality in reading scores relative to other countries in that column and red cells greater inequality. M0 to M7 refer to the different model specifications used to create the estimates of educational inequality (see Zieger, 2022, for further details). OECD mean and median are unweighted summaries of the 33 countries shown. South Korea is omitted because complete estimates across M0–M7 were unavailable owing to computational difficulties.

Hopefully, the example provided above has illustrated the central point. Table 8.3.1 was little more than a dump of information, whereas Table 8.3.2 turns a subset of the results into a visualisation that speaks directly to the research question. The changes suggested took only around ten minutes to implement in Excel but made a substantial difference to the table's clarity and usefulness. Readers can now identify the main patterns much more quickly, compare countries more easily and see how sensitive the results are to different model specifications. Small improvements to the title, ordering, formatting and use of colour can therefore make a major difference to the messages that readers take away.

8.4. Reporting results from regression analyses

Regression analysis is one of the central tools used by applied researchers. Consequently, many secondary analyses of ILSAs involve some form of regression model. As discussed in sections 7.2, 7.3 and 7.4, these models are often used to explore associations between background questionnaire variables and assessment scores (or another outcome), once a selection of background factors have been controlled. Regression can also be used to investigate whether differences between countries remain after other factors have been taken into account (e.g. whether countries continue to differ in the socio-economic gap in achievement after accounting for the migration background of the participants).

Regression analysis also creates an important challenge for data presentation. A single model may produce many parameter estimates along with associated inferential statistics (e.g. standard errors, test statistics, confidence intervals). When results are reported for several countries (e.g. across multiple time points or using different model specifications) the volume of information can quickly become unmanageable. It is rarely possible, or useful, to present every result on a single page. Researchers therefore need to decide which estimates are most relevant to the research question (and which communicate the results in a transparent yet accessible way).

In general, when presenting results from regression analyses from ILSAs in the form of a table, I offer the following advice:

- Compare countries directly. When the research question concerns cross-country comparisons, present estimates in a single table wherever possible so differences are easy to identify. Try not to split results across multiple tables.

- Report essential estimates only. Think carefully about which parameters readers need. Estimates for control variables are often unnecessary unless they are substantively important to the analysis.
- Include sample information. Consider reporting country sample sizes in the table notes or appendix.
- Use clear variable labels. Ensure that variable names are easy to understand and make the reference category for each categorical variable explicit.
- Select countries carefully. Present only the countries needed to communicate the main findings, placing results for additional countries in an appendix where appropriate.
- Consider pooled estimates. A pooled estimate can provide the main result in the main body of the paper, while country-specific estimates are presented separately in an appendix.
- Balance countries and models. There is often a trade-off between the number of countries included and the number of model specifications displayed clearly. Whether you choose to present more country-specific estimates or results from a greater number of specifications will depend on your specific research question.
- Use interpretable metrics. Avoid reporting logistic regression results only as log-odds. Consider odds ratios, marginal effects or approximate risk ratios where these are more meaningful⁴⁰.
- Consider standardisation. Standardising continuous variables can help readers compare effect sizes across measures by expressing estimates in standard deviation units rather than original scales (see section 4.8 for further discussion regarding standardisation using data from ILSAs).
- Limit the number of categories. For categorical variables, consider whether results for every group are needed or whether selected comparisons would communicate the main message more clearly.
- Apply consistent formatting. Follow the general guidance presented in section 8.3 above, including on informative titles, clear headings, table notes, white space and the removal of unnecessary grid lines.
- Explain model changes. Make it clear which variables are added or removed between specifications so readers can understand why estimates change across columns.

⁴⁰ It is useful to know that, when a binary outcome measure is “rare” (e.g. less than 20% experience the outcome) the odds ratio approximates the risk ratio (Davies et al., 1998). This is a useful property to know, given risk ratios are often easier to interpret and communicate.

- Keep samples consistent. Indicate whether the analytical sample changes across models, as differences in estimates may otherwise be incorrectly attributed to added control variables.

To illustrate why the above matters, we provide a hypothetical example. Suppose the aim is to estimate socio-economic differences in teenagers' mathematics self-efficacy and compare the size of this gap across three countries. The estimates are based on an ordinary least squares regression model. This is estimated once per country, capturing the socio-economic gap after controlling for gender, ethnicity, language spoken at home and special educational needs. An initial version of the results is shown in Table 8.4.1. This is deliberately exaggerated to demonstrate what poor presentation can look like and to make the improvements easier to identify. Estimates refer to differences in mathematics self-efficacy compared to the reference group.

Table 8.4.1. Initial presentation of estimates of the socio-economic gap in teenagers' mathematics self-efficacy. An example of poor presentation.

	Par 1	err 1	Par 2	err 2	Par 3	err 3
_QSES_2	0.054637	0.122312	0.032453	0.0328627	0.103312	0.0458928
_QSES_3	0.095018	0.145493	0.674923	0.0524794	0.203242	0.0248929
_QSES_4	0.198193	0.189944	0.738974	0.0222936	0.312394	0.0442782
_QSES_5	1.101837	0.112323	1.232892	0.0390093	0.439001	0.0258939
INDSEX	0.489544	0.44492	0.759464	0.45458	0.849421	0.4840222
_ETH_B	-0.56938	0.283212	-0.63902	0.299491	-0.59211	0.2680312
_ETH_A	1.034323	0.592422	1.147447	0.572949	1.192212	0.593231
_ETH_O	0.263292	1.042323	0.244246	1.006634	0.212232	1.059602
_OLANG	0.064032	0.483232	0.082879	0.439634	0.093234	0.4032321
_SPECEDNEED	-1.39531	0.542211	-1.48925	0.579634	-1.38492	0.483212

As in the previous section, there are several ways in which the presentation of these results can be improved. Specific suggestions are:

- Remove unnecessary coefficients. The research question concerns the socioeconomic gap within each country, so the estimated coefficients for the control variables are largely irrelevant to the main discussion. Removing them substantially reduces the size of the table and helps readers focus on the results that matter. The full set of coefficients can be reported in an appendix if needed. The table notes should still identify which variables were included as controls.

- Clarify the socioeconomic groups. Label each socioeconomic group clearly and indicate which category is being used as the reference group. The table notes should also explain how the socio-economic groups were constructed.
- Standardise the estimates. Express the coefficients in standard deviation units of the original self-efficacy scale. This makes the size of the differences between socioeconomic groups easier to interpret and compare across countries. The table notes should state how the outcome was standardised (e.g. whether it has been standardised within or across countries).
- Improve the column headings. Replace vague labels with full country names wherever possible. The headings should be as informative as space allows and should make clear what each column contains. In this instance, one should try to make clear the socio-economic differences are now reported in standard deviation differences (after the standardisation has taken place in the point above).
- Report the sample sizes. Add the sample size for each country in the final row of the table. Alternatively, include these figures in the table notes if the table is becoming cramped.
- Simplify the formatting. Reduce the number of grid lines and use bold text for the column headings. A cleaner layout makes it easier to distinguish the structure of the table and prevents unnecessary visual clutter from competing with the results.
- Reduce numerical precision. Report fewer decimal places to reduce clutter in the table and to avoid erroneous precision. Two decimal places are likely to be sufficient here, although three could also be justified. Reporting more than three decimal places would add clutter without providing meaningful information.
- Remove less useful groups. Consider whether the results for the second and fourth socioeconomic quintiles need to appear in the main text. In this example, they add little to the main comparison, so they can be removed to make the table more focused. The complete results can still be provided in an appendix for transparency.
- Indicate statistical significance. Decide which comparisons are most relevant to the research question. One option is to show whether each coefficient differs significantly from zero. However, if the main interest is whether the socioeconomic gap differs across countries, it may be more useful to test whether the estimates for Countries B and C differ significantly from the estimate for Country A. (This is the approach used in our updated table below).
- Add table notes. Include enough information in the notes to the table so that it can be understood on its own. This should include details about the outcome measure, the

construction of the socioeconomic groups, control variables, the standardisation procedure and the meaning of any significance symbols.

After making these changes, a revised version of the table is presented in Table 8.4.2. The revised table is now much easier to read, with the key estimates now standing out much more clearly. Specifically, by presenting fewer estimates, it directs the reader's attention towards those that are most relevant to the research question posed (and the main message of the analysis). It is also easier to compare findings across countries and to interpret the size of the estimates in terms of standard deviation differences. The notes beneath the table are still fairly detailed and some of this material could perhaps be shortened. Even so, they provide useful information about the model, variables and interpretation of the results. However, overall, I would argue that this version of the table is now an effective way to present the results.

Table 8.4.2. Improved presentation of estimates of the socio-economic gap in teenagers' mathematics self-efficacy. An example of improved presentation.

	Country A		Country B		Country C	
	SD difference	Standard error	SD difference	Standard error	SD difference	Standard error
Socio-economic group (reference = most disadvantaged 20%)						
Third Quintile	0.05	0.07	0.31*	0.02	0.11	0.02
Top quintile (most advantaged 20%)	0.55	0.06	0.56	0.02	0.24*	0.02
Sample size	5,384		4,953		5,215	

Notes: Socio-economic groups based on within-country quintiles of the ILSA socio-economic status scale. Estimates based on an OLS regression model controlling for gender, ethnicity, language spoken at home and special educational needs. Maths self-efficacy has been standardised within countries, with estimates therefore reported in terms of standard deviation differences from the reference group. "SD difference" column illustrates the estimated difference in maths self-efficacy compared to the most disadvantaged 20% of pupils as the reference group. Results for the second and fourth quintile not presented for brevity. * indicates where the estimates are significantly different from country A at the 5% level.

8.5. Cross-country scatterplots.

As readers may have noticed, cross-country scatter plots have appeared several times throughout this book. They are one of the most useful ways of presenting results from cross-national comparative research, particularly when countries need to be compared across two dimensions at the same time. For example, a scatterplot might compare average reading and mathematics scores, the size of socioeconomic inequalities in achievement and life satisfaction or results from two different assessment cycles. These graphs can be especially powerful when they reveal a common pattern across countries. A plot comparing girls' average reading scores with boys' scores would, for instance, illustrate how there is a similar pattern to the gender gap

across a wide range of countries. Such scatterplots are also useful when no overall global pattern is visible, immediately providing striking visual evidence that this is the case. A particular advantage of these graphs is that they can continue to display a substantial amount of information while remaining more visually engaging than a large table. This makes them particularly effective for communicating the main message of an analysis.

If one decides to present results from an ILSA in this form, we offer the following points to consider:

- Prioritise clear communication. Remember that graphs are primarily communication devices. Their main advantage over tables is that they can make patterns and comparisons easier and quicker to grasp. Before finalising a plot, ask yourself whether it communicates the main message of the analysis quickly and clearly.
- Keep the design clean. Remove unnecessary gridlines, borders and other decorative features. Aim for a simple layout with enough white space for the data to stand out. Every element on the graph should serve a clear purpose.
- Choose markers carefully. Two-letter or three-letter country codes are often more informative than standard circular markers because they allow readers to identify each country directly⁴¹. Full country names can make the plot cluttered, especially when many countries are included. Explain the abbreviations elsewhere in the document (e.g. as a read across to full country names in another table) and refer readers to this information in the figure notes.
- Use honest axis scales. Choose axis ranges that make the data easy to read without exaggerating the size of cross-country differences. Avoid scales that are unnecessarily narrow. When the same measure appears on both axes, use the same scale wherever possible.
- Label axes clearly. Axis labels should be brief but informative. Keeping labels horizontal makes them easier to read.
- Select countries carefully. It may be reasonable to show only a subset of countries when this is relevant to the research question. For example, the plot might focus on OECD countries or exclude an extreme outlier that would make the rest difficult to see. Any exclusions from the plot should be clearly justified and documented, with the full results (including these countries) reported in an appendix.

⁴¹ A list of such codes is available from <https://www.iban.com/country-codes>

- Highlight key countries. Particular countries can be emphasised using different shapes, colours or font styles. This may be useful when benchmarking one country against others or identifying countries with statistically significant differences. Any highlighting should be explained clearly in the figure notes.
- Write useful figure notes. All figures should be accompanied by clear and concise notes. These should explain important definitions, abbreviations, sample restrictions, statistical details and any other information readers need to interpret the graph correctly.
- Add a reference line. A line can make the main comparison easier to interpret. When the same measure appears on both axes, a 45-degree line shows where values would be equal. A regression line can instead summarise the overall cross-country relationship. The choice should be guided by the research question.
- Report the correlation. Regardless of which reference line is used, including the cross-country correlation in the figure notes almost always provides useful auxiliary information.
- Consider interactive versions. Static graphs are usually necessary for articles, reports and books, but interactive versions can provide additional detail. They may allow readers to zoom in, identify countries or inspect exact values. These can be useful as supplementary material alongside the main static figure.
- Use AI for feedback. Once a first draft has been produced, AI tools can be used to identify possible improvements in layout, labelling and readability. They can also suggest code for implementing changes.
- Test with non-specialists. Show the finished plot to a friend or colleague who is not closely involved in the analysis. Ask them to explain what they think it shows. If they cannot identify the main message quickly without guidance, the plot is likely to need further work.

We again illustrate these points with an example. In particular, two closely related studies examined whether mothers' or fathers' education is more strongly associated with pupils' scores in ILSAs. Marks (2008) and Jerrim and Micklewright (2011) both asked whether one parent's education appeared to matter more than the others and whether the same pattern could be found across countries⁴². For example, is mothers' education consistently more strongly related to pupil achievement across education systems?

⁴² Although Marks (2008) and Jerrim and Micklewright (2011) addressed related questions, they did not analyse the same exact dataset. Specifically, Marks used PISA 2000, whereas Jerrim and Micklewright used PISA 2003 and mainly focused their cross-country discussion on mathematics.

Yet, despite asking similar research questions and conducting their analyses using similar data and methods, they presented their findings in markedly different ways. In particular, Marks (2008) reported the country-level estimates in a table, while Jerrim and Micklewright favoured a cross-country scatterplot. This provides a useful opportunity to think about how the presentation impacts the reader's understanding. A table allows the precise estimates for each country to be inspected, whereas a scatterplot may communicate the overall pattern and highlight unusual countries more quickly.

Table 8.5.1. Table of results illustrating the association between maternal and paternal education with PISA reading test scores.

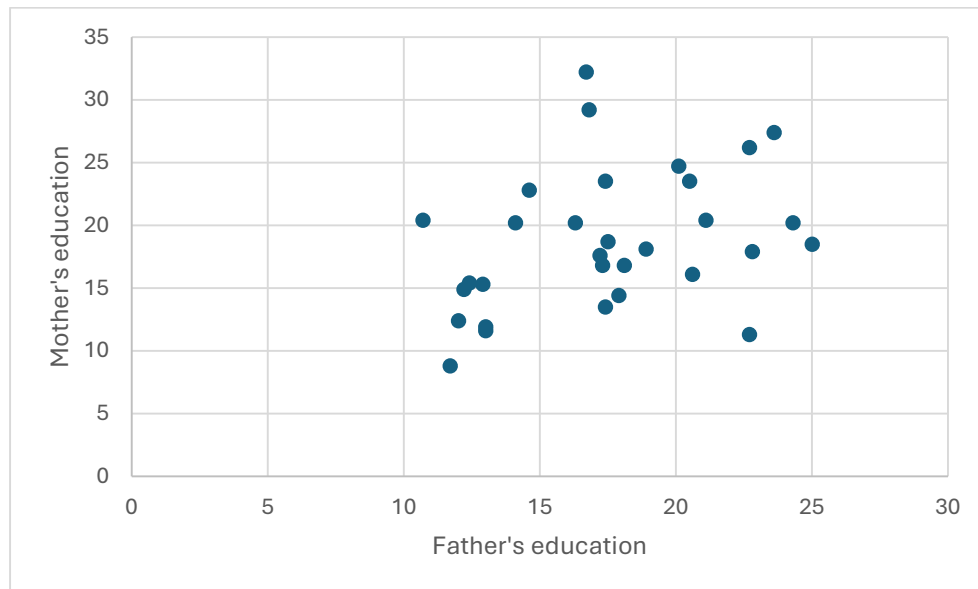
Country	Intercept	Father's education	Mother's education	R Sq.
Australia	533	22.7	11.3	0.09
Austria	511	17.3	16.8	0.10
Belgium	517	23.6	27.4	0.19
Brazil	396	14.6	22.8	0.15
Canada	536	12.9	15.3	0.06
Czech Republic	497	25	18.5	0.16
Denmark	503	20.1	24.7	0.17
Finland	550	13	11.6	0.06
France	511	17.4	23.5	0.16
Germany	502	16.8	29.2	0.16
Greece	474	14.1	20.2	0.10
Hungary	482	22.7	26.2	0.24
Iceland	511	10.7	20.4	0.09
Ireland	529	12.2	14.9	0.06
Italy	489	17.4	13.5	0.09
Korea	524	11.7	8.8	0.07
Latvia	463	12.4	15.4	0.06
Luxembourg	450	24.3	20.2	0.17
Mexico	425	20.5	23.5	0.22
Netherlands	539	18.1	16.8	0.13
New Zealand	541	20.6	16.1	0.09
Norway	511	17.9	14.4	0.07
Poland	485	17.2	17.6	0.11
Portugal	474	18.9	18.1	0.12
Russia	464	12	12.4	0.05
Spain	496	16.3	20.2	0.15
Sweden	520	13	11.9	0.05
Switzerland	499	16.7	32.2	0.20
United Kingdom	533	21.1	20.4	0.14
United States	513	22.8	17.9	0.12
Average		17.5	18.7	0.12

Source: Results drawn from Marks (2008).

Table 8.5.1 reproduces selected results from Marks (2008) in a format similar to that used in the original paper. As an aside, readers may wish to revisit the advice on table design in the previous two sub-sections and consider how the presentation could be improved further. Following the approach used throughout this chapter, we then create an initial cross-country scatter plot as an alternative way of presenting the same findings (Figure 8.5.1). The coefficient for fathers'

education is shown on the horizontal axis, while the coefficient for mothers' education is shown on the vertical axis. As in earlier examples, this first version of the graph is deliberately poor and provides a starting point for discussing how its design and presentation might be improved.

Figure 8.5.1. Initial scatterplot of the association between maternal and paternal education with PISA reading test scores.



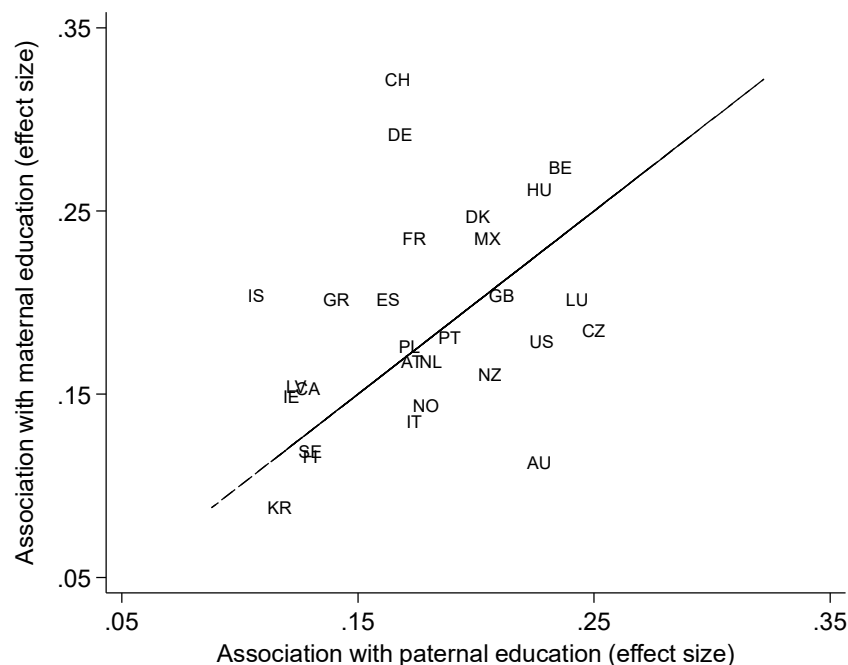
Following the guidance offered above, we note several ways in which the presentation of this graph can be improved. This includes:

- Remove visual clutter. Removing gridlines and borders gives the graph a cleaner appearance and helps readers focus on the comparisons that matter.
- Focus on OECD countries. Excluding Brazil and Russia creates a more comparable sample without changing the main pattern shown in the graph.
- Standardise the coefficients. Expressing estimates using an international standard deviation of 100 makes their size easier to interpret than raw PISA points.
- Clarify the axis titles. More informative labels make it immediately clear what each axis represents and how the plotted values should be interpreted.
- Use country-code markers. Country codes allow readers to identify individual countries quickly without creating the clutter that full country names would introduce.
- Add an equality line. The 45-degree line shows whether associations are stronger for mothers or fathers and reveals countries scattered on both sides.
- Cross-country correlation. The notes to the figure now report the cross-country correlation of 0.34.
- Include informative notes. Clear figure notes explain the sample, measures, standardisation and design choices needed for readers to interpret the graph correctly.

- Highlight significant differences. Countries with statistically significant differences could be emphasised, although this was not possible because the original paper by Marks (2008) did not report the relevant figures.

The revised results are presented in Figure 8.5.2. This, arguably, provides a more effective summary than either Table 8.5.1 or Figure 8.5.1. Some detail from the original table has inevitably been lost, but the central message is now much easier to identify from a quick scan of the graph. In particular, one can easily see that there is no consistent tendency for maternal education to be more strongly associated with children's reading skills than paternal education (or vice versa). Instead, countries are scattered on both sides of the 45-degree line. The graph also makes unusual cases much easier to spot. For instance, Switzerland (CH) and Germany (DE) stand out because the association between reading skills and maternal education is noticeably stronger than the corresponding association with paternal education. Identifying these features, and ensuring that readers notice them, is much more difficult in the original table and almost impossible in the poorly presented version of the plot in Figure 8.5.1. In doing so, Figure 8.5.2 hopefully illustrates the value of cross-country scatterplots as a communication device.

Figure 8.5.2. Revised scatterplot of the country-level association between maternal and paternal education and PISA reading test scores.



Notes: Each label represents one country. Estimates are the country-specific reading-score coefficients reported by Marks (2008) divided by 100 (the approximate standard deviation of the PISA reading scale). Parental education was standardised within country. Brazil and Russia are omitted, leaving 28 countries. The 45-degree line is where the effect size for mothers' and fathers' education is equal. Countries can be identified by two-letter country codes. The cross-country correlation is 0.34.

8.6. bar charts

Bar charts are widely used and will be familiar to most readers. In ILSAs they provide a useful alternative to presenting results in a “league table”. A bar chart can still show the rank order of countries (though, as discussed in Section 8.2, country rankings are often not particularly helpful). Yet, more importantly, bar charts can make the size of differences between countries much easier to see. They can also display uncertainty (e.g. by adding confidence intervals around each estimate) as discussed in Section 4.3.

However, much like our discussion of cross-country scatterplots, bar charts should primarily be treated as a communication device. They generally use more space than tables and therefore present less information within the space. This means they must be designed to take full advantage of what they do well; drawing attention to the main comparisons, making the magnitude of differences easy to understand and helping readers remember the key findings. A good bar chart should thus simplify interpretation of the results without oversimplifying the underlying evidence.

To achieve this in the context of presenting results from ILSAs, the following advice is offered:

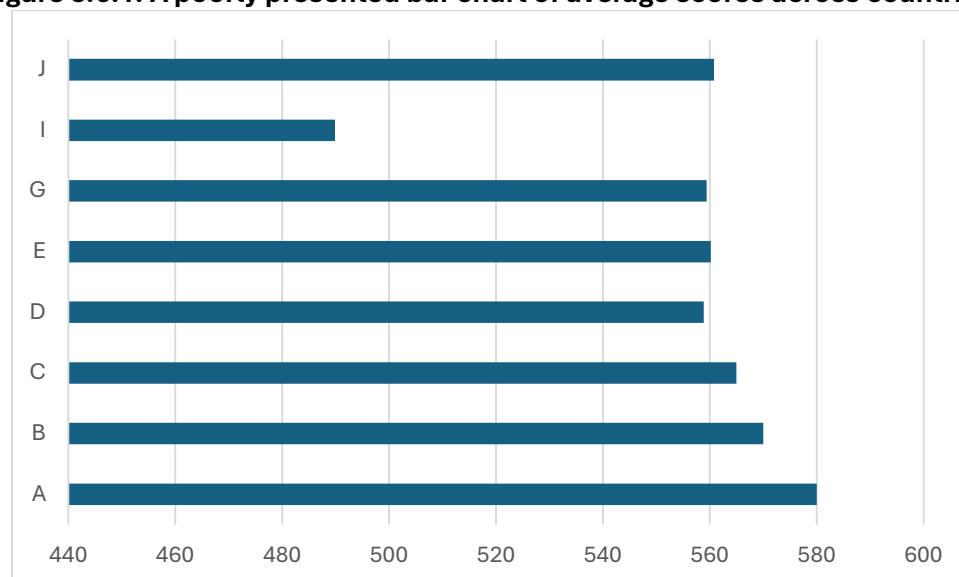
- Order bars meaningfully. Arrange countries in a logical order rather than alphabetically. Where higher values represent better outcomes, placing the strongest results at the top usually makes the chart easier to read and helps readers compare countries’ relative positions.
- Highlight focal countries. Use a different bar colour to emphasise any country of particular interest. This is especially useful when benchmarking one country against its international peers, as readers can identify the focal country immediately and compare it with the others.
- Choose country labels carefully. Full country names are generally preferable, but long names may leave less space for the bars themselves. A mixture of full names and familiar abbreviations can be used where necessary, provided that any shortened labels are explained clearly in the figure notes.
- Use horizontal bars. Bar charts are often most effective when the bars run from left to right and countries are listed from top to bottom. This layout accommodates country names more easily and remains readable when a relatively large number of countries is included.

- Select countries sensibly. It may not be practical to show every participating country. Choose a relevant and clearly justified group based on the research question and the available space. Any exclusions should be documented, with the complete results provided elsewhere where possible.
- Add value labels. Readers may want to know the precise value represented by each bar rather than estimating it from the axis. Adding labels directly to the bars makes this information easy to access without requiring a separate table.
- Show statistical uncertainty. Include confidence intervals where possible. These communicate the uncertainty surrounding each estimate and can support comparisons between countries (recall the discussion in section 4.3 and 4.5 on related matters).
- Apply consistent formatting. Follow the broader principles introduced for cross-country scatterplots in section 8.5. Use informative axis titles, concise figure notes and plenty of white space. Remove unnecessary borders, gridlines and other features that do not help communicate the results.

To illustrate these points, we return to the data presented previously in Table 8.2.1, focusing on the results at time point 2. These results were presented in terms of an international “ranking” and (while average scores were also presented) the differences and similarities across countries were not instantly striking.

Figure 8.6.1 presents a deliberately poorly designed bar chart illustrating these results. As with the earlier examples, this initial visualisation is in several respects less effective than the initial presentation of results in Table 8.2.1.

Figure 8.6.1. A poorly presented bar chart of average scores across countries.

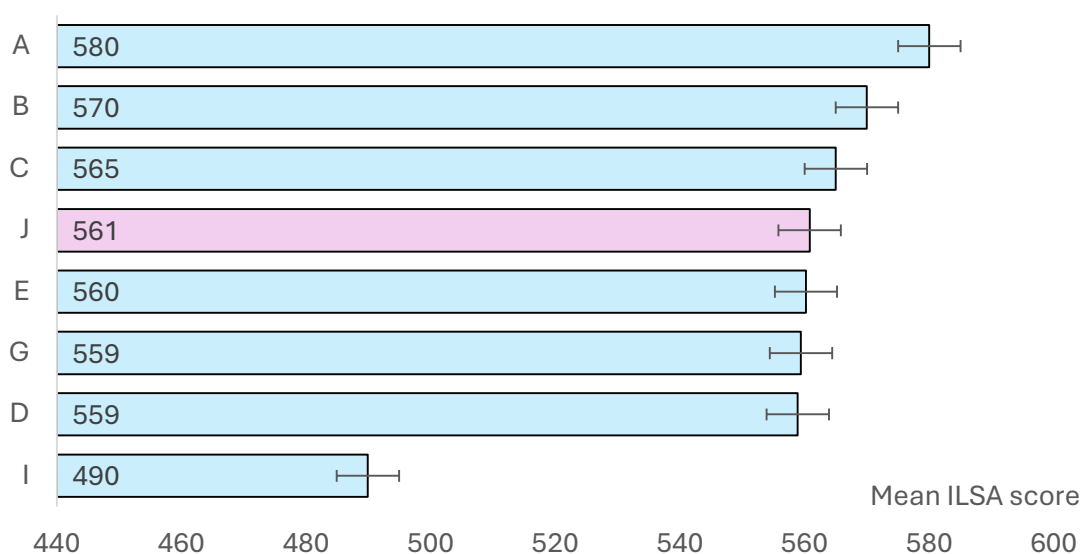


Following the advice above, how then can this initial version of the graph be improved? The following changes are suggested which in total take little more than five minutes to implement. (Excel is used to produce Figure 8.6.2, although the figure can also be produced in any other standard statistical software).

- Order bars meaningfully. Sort countries from the highest average score at the top to the lowest at the bottom, making comparisons much easier.
- Remove unnecessary clutter. Delete gridlines and borders and leave enough white space for the bars and labels to stand out clearly.
- Improve the look of the bars. Reduce the gap between the bars. Use less intense bar colours. Add subtle borders to the bars where needed.
- Label the axis. Add a clear axis title so readers immediately understand what the values represent.
- Show values. Display the average on the chart, making precise results easy to identify.
- Highlight the focal country. Use a different colour for Country J so readers can quickly see the country being benchmarked against the others.
- Add confidence intervals. Confidence intervals show sampling uncertainty and help informed readers judge whether differences between countries are likely to be statistically meaningful.
- Include clear notes. The graph is already largely self-contained. However, it is worth clarifying that the lines running through the bars are confidence intervals.

The chart that emerges after making these changes is presented in Figure 8.6.2.

Figure 8.6.2. Hypothetical example of average scores across countries. A visually appealing bar chart.



Notes: Bars show hypothetical country mean scores at time point 2. Thin black lines running through the centre of bars indicate the 95% confidence intervals. Country J highlighted as the focal country.

This revised graph provides a much more striking way of communicating the results. It is now easy to see that Country J performs similarly to most of the countries included in the comparison. Only two or perhaps three of the comparators are likely to be significantly different. It is also easy to see that Country I stands out with an average score significantly below those of the other countries. At the opposite end of the distribution, Country A has the highest average score and is likely to differ significantly from many of the comparators, at least before applying any adjustment for multiple comparisons. Overall, the graph thus now helps readers to identify the key points very quickly (i.e. Country J is broadly typical, Country I performs unusually poorly and Country A performs a bit more strongly). These messages are both easier to understand and more likely to be remembered than when presented in a table (e.g. Table 8.2.1) or where little effort has been put into the presentation (e.g. Figure 8.6.1).

8.7. Beautiful line graphs

The final type of chart considered is the line graph, where year is usually plotted on the horizontal axis with the outcome of interest shown on the vertical axis. In ILSAs line charts are mainly used to show how outcomes (e.g. average test scores) change over time. Their particular value is to illustrate how trends in one country compare with those in other countries. In such instances, the international comparative angle can help provide important context to the results. For instance, following the COVID-19 pandemic average PISA scores fell in many countries. Viewed in absolute terms, the picture was hence rather negative almost everywhere. Yet the comparative perspective illustrated how average scores in some countries held up better than others. Thus, when designed well, line charts can communicate trends in cross-country differences quickly and clearly. While they contain less detailed information than a table, the benefit is hopefully that readers grasp, remember and share the key findings much more readily.

When producing line charts using data from ILSAs, we offer the following the advice:

- Limit the number of lines. Too many lines will quickly clutter the graph and make the main message difficult to identify. Where possible, show trends for only a small number of countries, perhaps alongside an international benchmark such as the OECD average.
- Emphasise the focal country. When many comparators are genuinely needed, highlight the main country using a thick, solid black line. Plot the remaining countries using thin grey lines in the background, without labels. Exact figures can be provided separately in an appendix table.
- Mark breaks in series. Clearly identify any year in which the assessment framework, scale, sample coverage or data collection method changed substantially. An example from

ILSAs is when PISA moved from paper to computer assessment in 2015. A vertical line with annotation can warn readers that results before and after that point may not be directly comparable.

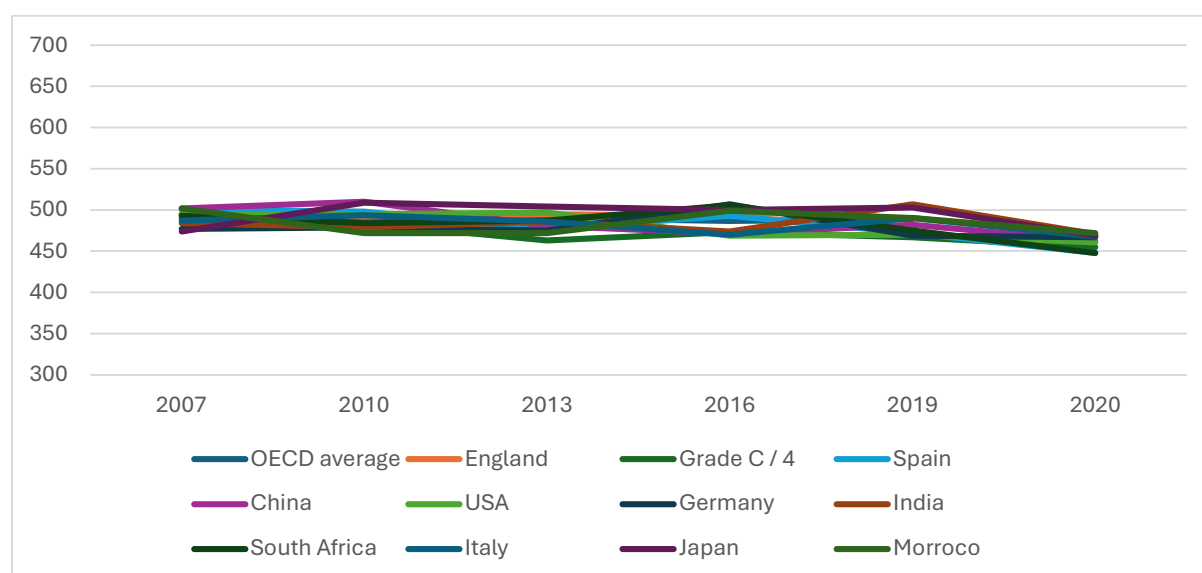
- Show uncertainty carefully. Confidence interval bands can help communicate uncertainty caused by sampling variation, but they can also make the graph difficult to read. Where confidence bands would create too much clutter, statistical significance can instead be discussed in the text or reported in an appendix.
- Label points selectively. Value labels and short annotations can help readers identify exact results or understand important contextual information. However, they work best when only a limited number of lines and data points are being displayed.
- Space years accurately. Make sure that the distance between years on the horizontal axis reflects the actual time gaps. For example, a four-year interval should appear twice as wide as a two-year interval.
- Distinguish comparison groups. Use different colours, markers and/or line styles to help readers separate the countries being compared. These choices should ideally remain clear when printed in greyscale and should not rely on colour alone.
- Position the legend carefully. Place the legend in a sensible location, ideally within unused white space inside the plotting area rather than taking up extra space above or below the graph. Labels should be concise, informative and easy to match to the lines.
- Use small multiples. When several countries must be presented, separate panels using the same axis scales may be clearer than placing every line on one graph. Small multiples allow each trend to remain visible while still supporting cross-country comparison.
- Apply the wider guidance discussed previously. The general advice for other graph types still applies. For example, keep the design clean, use white space effectively, add informative figure notes and seek feedback from AI tools and non-specialist readers.

An exemplar line chart is presented using PISA data from England linked to administrative records, including participants' performance in national GCSE examinations. These are high-stakes national exams taken by teenagers around six months after they take the PISA test. The line chart is designed to examine how the effective "floor standard" in GCSE mathematics (i.e. the minimum level that teenagers are generally expected to achieve) has changed over time. This was traditionally a grade C and is now broadly represented by grade 4. Interpreting trends is

difficult because England's national examinations changed during the period. Moreover, more generally, national tests may be affected by grade inflation (as discussed in Section 1.5.1) with PISA providing an external benchmark less susceptible to domestic political pressures. The graph hence illustrates the average PISA score of pupils who achieved the GCSE mathematics floor over time. It also includes trends for England overall and the OECD average to establish whether any change is specific to pupils reaching the floor standard or reflects a wider movement in PISA performance.

Consistent with the previous section, Figure 8.7.1 provides a contrived example where the initial illustration of the results is poor and which we will then go on to improve (applying the advice given above and in previous sections).

Figure 8.7.1. The decline in PISA scores associated with achieving grade C/4 in GCSE mathematics in England since 2007. A poorly presented line chart.



Notes: Comparator-country series are simulated for illustrative purposes only.

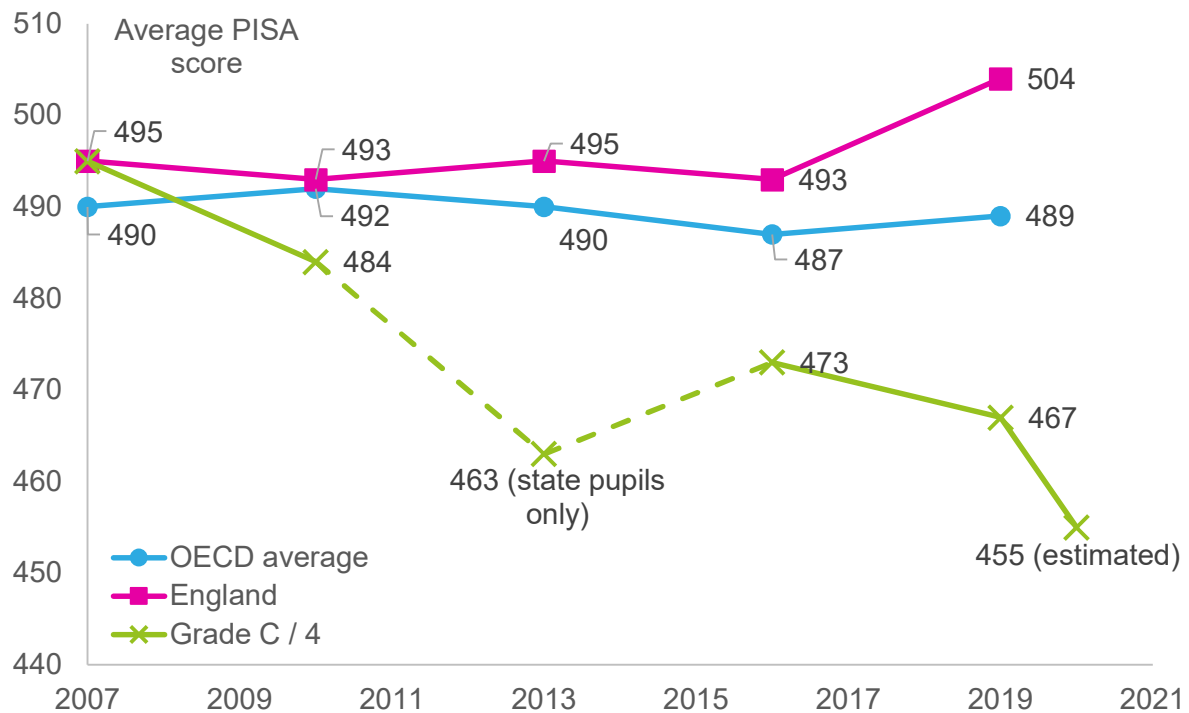
To then improve this graph visually, we follow the steps outlined below:

- Remove irrelevant information. Much of the information in the original graph does not help answer the research question (e.g. results for China, Spain and Germany). Only three lines are retained – (a) the average PISA score of those achieving a C/4 grade; (b) the average PISA score across England and (c) the OECD average. Remember that less is often more.

- Space years accurately. Most observations are three years apart. However, the final two are separated by only one year. The visual distance between points is changed to reflect the actual time interval rather than presenting every year as equally spaced.
- Choose an informative scale. The current vertical axis spans 400 points, equivalent to around four standard deviations. This makes the trends difficult to identify. A narrower (but still honest) scale of 70 points is thus reverted to in order to illustrate the trends visibly without exaggerating them.
- Distinguish groups clearly. Different markers are used for the different lines that remain. This helps readers identify each series quickly and makes the graph easier to interpret.
- Create a cleaner appearance. The unnecessary borders and gridlines are removed to make more effective use of white space.
- Position the legend carefully. The legend is moved to the bottom-left of the plotting area.
- Label the vertical axis. A concise label of “average PISA score” is added to the vertical axis. A horizontal-axis title is unnecessary here because the year labels already make the meaning of that axis clear.
- Add useful data labels. Now unnecessary information has been removed there is enough space to display the values beside each point. In this instance, these labels help readers obtain precise figures without needing to consult a separate table.
- Annotate important points. Some observations require additional explanation. Although this information could have been placed in the figure notes, brief annotations beside the relevant points make some important issues clearer and more transparent. In this instance, these additions do not make the graph cluttered.

The resulting graph is presented in Figure 8.7.2. The key findings are now much easier to see. Most striking is the decline over time in the average PISA maths score of pupils achieving grade 4/C. At the same time, average PISA maths scores for England as a whole slightly increased and remained above the OECD average. The decline amongst pupils at the grade 4/C threshold therefore does not appear to reflect a general fall in mathematical performance across the country. Rather, it suggests that the level of mathematics skills required to reach this floor standard may have been lowered over time. Further discussion of this issue is provided by Jerrim (2021).

Figure 8.7.2. The decline in PISA scores associated with achieving grade C/4 in GCSE mathematics in England since 2007. A nicely presented line chart.



Notes: 2012 data for the Grade C/4 results is based on state school pupils only. The 2020 figure is an estimated rather than a directly observed data point.

8.8. Test your knowledge questions

8.8.1. Task 1.

Task 1. Table 8.8.1.1 presents average test scores for in a mathematics test across several countries. The columns are as follows:

- Country = The three-letter ISO abbreviation code for the country. (You can find how these codes correspond to countries here: https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes)
- 2020_Mean = The average mathematics score achieved by pupils in the country in 2020.
- 2020_SE = The standard error of the average mathematics score achieved by pupils in the country in 2020.
- 2025_Mean = The average mathematics score achieved by pupils in the country in 2025.
- 2025_SE = The standard error of the average mathematics score achieved by pupils in the country in 2025.

Table 8.8.1.1 Average test scores in 2020 and 2025 (hypothetical data)

Country	Mean_2020	SE_2020	Mean_2025	SE_2025
ALB	501	3.20	501	3.20
ARG	493	3.52	513	3.48
AUS	508	3.82	515	4.12
AUT	490	3.67	511	4.02
BEL	490	4.07	512	3.77
BRA	493	3.20	512	3.42
BUL	488	4.11	487	3.45
CAN	516	4.15	506	3.78
CHE	498	3.31	488	3.24
TPI	492	3.76	509	3.29
CHL	493	3.45	492	4.01
COR	504	3.75	501	3.86
HRV	494	3.85	487	3.84
CZE	515	3.79	509	3.87
DEU	508	3.85	494	3.55
DNK	495	3.92	502	3.88
ESP	498	3.99	508	3.60
EST	502	3.76	509	3.62
FIN	509	3.26	499	3.69
FRA	510	3.55	512	3.71
GRC	499	3.54	497	3.39
HK	513	3.45	506	4.01
HUN	486	3.71	498	4.00
IND	505	3.69	499	3.79
IRL	488	3.53	494	3.23
ISL	491	3.27	505	4.12
ISR	503	3.44	498	4.09
ITA	499	3.41	486	3.46
JOR	500	3.52	500	4.01
JPN	491	3.87	508	4.19
KAZ	513	3.38	495	3.94
LAT	496	3.42	516	4.19
LTH	504	4.03	502	3.64
LUX	494	3.69	511	3.94
MCU	487	3.91	491	3.60
MAL	488	3.43	490	3.88
MEX	493	3.35	490	3.58
MON	490	3.68	493	3.62
NLD	495	3.26	515	3.38
NOR	498	3.40	502	3.98
NZL	495	3.49	512	4.20
PRE	496	3.75	507	3.53
POL	501	4.12	504	3.67
PRT	495	4.09	493	3.88
QAT	489	3.70	489	3.58
ROM	508	4.00	509	3.43
RUS	495	3.97	501	3.66
SGP	496	3.28	491	3.36
SVK	507	3.92	512	3.81
SVN	491	4.14	501	3.71
SWE	501	3.23	504	3.63
THL	502	3.87	507	3.82
TUN	506	4.01	499	4.17
TUR	515	3.48	508	3.52
UK	497	3.66	503	3.57
UAE	498	3.89	499	3.71
URG	490	3.63	486	4.00
USA	506	3.57	514	3.67
VNM	513	4.08	514	3.30

Task 1. Question 1. Using the data available, please calculate for each country whether the change in the average scores between 2020 and 2025 is statistically significant at the 5% level. (Hint: You may need to draw upon knowledge / material covered earlier in the course to do this!)

Task 1. Question 2. Should you make a Bonferroni adjustment when conducting these significance tests? Explain your rationale for doing so / not doing so, and the implications of this.

Task 1. Question 3. Your boss at work asks you to use the information in this sheet to create a table or graph to illustrate how the change in mathematics skills between 2020 and 2025 compares across industrialised countries. This includes knowing in which countries the change is statistically significant. Please create such a table / graph, presenting this information as effectively as possible. You are free to use whatever software you wish to create these graphs (R; STATA; Excel).

8.8.2. Task 2.

Table 8.8.2.1 presents average test scores for in a mathematics test across several countries at six time points.

Please create a (better presented) table or graph that helps to better visualise these results for your boss at work. The key thing that they are interested in is how countries compare to one another within a given time point.

You can use whatever software you like to answer this question.

Table 8.8.2.1 Average test scores across countries at six time points (hypothetical data)

cnt	2009	2012	2015	2018	2021	2024
ALB	501.00	503.00	507.00	504.00	512.00	519.00
ARG	493.00	503.00	520.00	491.00	525.00	526.00
AUS	508.00	510.00	516.00	493.00	520.00	517.00
AUT	490.00	494.00	511.00	495.00	516.00	532.00
BEL	490.00	512.00	502.00	491.00	519.00	513.00
BRA	493.00	505.00	507.00	494.00	503.00	505.00
BUL	488.00	488.00	517.00	519.00	527.00	515.00
CAN	516.00	492.00	498.00	493.00	522.00	506.00
CHE	498.00	502.00	505.00	516.00	526.00	509.00
TPI	492.00	498.00	498.00	495.00	513.00	526.00
CHL	493.00	517.00	506.00	498.00	524.00	508.00
COL	504.00	517.00	520.00	501.00	526.00	523.00
HRV	494.00	491.00	521.00	491.00	517.00	508.00
CZE	515.00	496.00	522.00	510.00	515.00	519.00
DEU	508.00	501.00	513.00	511.00	510.00	506.00
DNK	495.00	504.00	512.00	503.00	508.00	505.00
ESP	498.00	494.00	499.00	493.00	524.00	521.00
EST	502.00	498.00	493.00	490.00	514.00	522.00
FIN	509.00	505.00	519.00	491.00	501.00	516.00
FRA	510.00	504.00	511.00	514.00	499.00	524.00
GRC	499.00	504.00	522.00	517.00	520.00	515.00
HK	513.00	499.00	496.00	498.00	508.00	533.00
HUN	486.00	511.00	506.00	492.00	527.00	507.00
IND	505.00	491.00	515.00	505.00	527.00	518.00

IRL	488.00	510.00	507.00	515.00	513.00	519.00
ISL	491.00	506.00	492.00	502.00	503.00	504.00
ISR	503.00	518.00	500.00	505.00	526.00	506.00
ITA	499.00	498.00	521.00	493.00	510.00	528.00
JOR	500.00	505.00	494.00	499.00	498.00	524.00
JPN	491.00	505.00	507.00	519.00	516.00	504.00
KAZ	513.00	517.00	516.00	505.00	516.00	515.00
LAT	496.00	508.00	513.00	497.00	517.00	525.00
LTH	504.00	498.00	509.00	510.00	502.00	526.00
LUX	494.00	493.00	503.00	497.00	497.00	534.00
MAC	487.00	517.00	502.00	510.00	499.00	526.00
MYS	488.00	501.00	504.00	502.00	500.00	526.00
MEX	493.00	488.00	510.00	493.00	515.00	533.00
MNG	490.00	502.00	500.00	517.00	518.00	532.00
NLD	495.00	495.00	509.00	508.00	517.00	524.00
NOR	498.00	502.00	493.00	502.00	518.00	505.00
NZL	495.00	514.00	503.00	490.00	503.00	514.00
PRE	496.00	509.00	497.00	509.00	505.00	505.00
POL	501.00	489.00	498.00	510.00	524.00	519.00
PRT	495.00	495.00	492.00	495.00	522.00	505.00
QAT	489.00	497.00	504.00	499.00	498.00	505.00
ROM	508.00	517.00	497.00	515.00	511.00	529.00
RUS	495.00	511.00	493.00	508.00	502.00	534.00
SGP	496.00	518.00	503.00	506.00	511.00	527.00
SVK	507.00	515.00	512.00	513.00	504.00	529.00
SVN	491.00	514.00	492.00	491.00	507.00	516.00
SWE	501.00	488.00	495.00	499.00	513.00	504.00
THL	502.00	511.00	519.00	517.00	517.00	524.00
TUN	506.00	511.00	522.00	496.00	509.00	534.00
TUR	515.00	508.00	511.00	493.00	499.00	506.00
UK	497.00	505.00	515.00	508.00	504.00	522.00
ARE	498.00	497.00	499.00	502.00	516.00	504.00
URG	490.00	511.00	496.00	498.00	523.00	521.00
USA	506.00	492.00	508.00	492.00	520.00	524.00
VNM	513.00	514.00	522.00	491.00	524.00	514.00

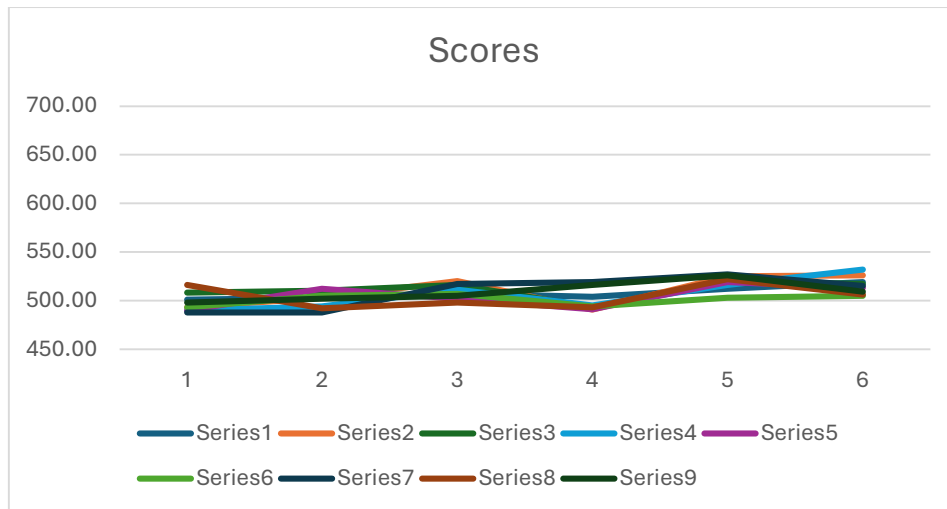
8.8.3. Task 3.

Figure 8.8.3.1 presents average test scores for in a mathematics test across several countries at six time points. This is based on the data presented in Table 8.8.3.1.

Table 8.8.3.1. Average test scores across countries. Hypothetical data in tabular form.

cnt	2009	2012	2015	2018	2021	2024
ALB	501.00	503.00	507.00	504.00	512.00	519.00
ARG	493.00	503.00	520.00	491.00	525.00	526.00
AUS	508.00	510.00	516.00	493.00	520.00	517.00
AUT	490.00	494.00	511.00	495.00	516.00	532.00
BEL	490.00	512.00	502.00	491.00	519.00	513.00
BRA	493.00	505.00	507.00	494.00	503.00	505.00
BUL	488.00	488.00	517.00	519.00	527.00	515.00
CAN	516.00	492.00	498.00	493.00	522.00	506.00
CHE	498.00	502.00	505.00	516.00	526.00	509.00

Figure 8.8.3.1. Average test scores across countries. Hypothetical data in a figure.



Task 3. Question 1. How do you think the presentation of this graph could be improved? Please make the changes to improve the presentation, listing out the steps you take step-by-step. Feel free to use whatever software you like to make these changes.

References

- Davies, H. T., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *BMJ (Clinical research ed.)*, 316(7136), 989–991. <https://doi.org/10.1136/bmj.316.7136.989>
- Department of Education and Youth. (2023). *Major international study shows Irish students are the best performers in reading literacy in the OECD and the EU*. Gov.ie. <https://www.gov.ie/en/department-of-education/press-releases/major-international-study-shows-irish-students-are-the-best-performers-in-reading-literacy-in-the-oecd-and-the-eu/>
- Jerrim, J., & Micklewright, J. (2011). Children's cognitive ability and parents' education: Distinguishing the impact of mothers and fathers. In T. M. Smeeding, R. Erikson, & M. Jäntti (Eds.), *Persistence, privilege, and parenting: The comparative study of intergenerational mobility* (pp. 261–285). Russell Sage Foundation.
- Jerrim, J. (2021, March 15). *Has a GCSE grade C/4 lost its value?* FFT Education Datalab. <https://ffteducationdatalab.org.uk/2021/03/has-a-gcse-grade-c-4-lost-its-value/>
- Kirk, A. (2024). *Data visualisation: A handbook for data driven design* (3rd ed.). SAGE Publications Ltd. <https://uk.sagepub.com/en-gb/eur/data-visualisation/book285882>
- Marks, G. N. (2008). Are Father's or Mother's Socioeconomic Characteristics More Important Influences on Student Performance? Recent International Evidence. *Social Indicators Research*, 85(2), 293–309. <http://www.jstor.org/stable/27734583>
- Zieger, L. R., Jerrim, J., Anders, J., & Shure, N. (2022). Conditioning: how background variables can influence PISA scores. *Assessment in Education: Principles, Policy & Practice*, 29(6), 632–652. <https://doi.org/10.1080/0969594X.2022.2118665>