United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

SUSTAINABLE DEVELOPMENT GOALS

Information Paper No. 54
September 2018
UIS/2018/ED/IP/54

Learning Divides: Using Data to Inform Educational Policy

UNESCO

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946. The Organization currently has 195 Member States and 11 Associate Members.

The main objective of UNESCO is to contribute to peace and security in the world by promoting collaboration among nations through education, science, culture and communication in order to foster universal respect for justice, the rule of law and the human rights and fundamental freedoms that are affirmed for the peoples of the world, without distinction of race, sex, language or religion, by the Charter of the United Nations.

To fulfil its mandate, UNESCO performs five principal functions: 1) prospective studies on education, science, culture and communication for tomorrow's world; 2) the advancement, transfer and sharing of knowledge through research, training and teaching activities; 3) standard-setting actions for the preparation and adoption of internal instruments and statutory recommendations; 4) expertise through technical cooperation to Member States for their development policies and projects; and 5) the exchange of specialised information.

UNESCO Institute for Statistics

The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science, technology and innovation, culture and communication.

The UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

This paper was written by J. Douglas Willms, University of New Brunswick and President, The Learning Bar Inc., ksi@nbnet.nb.ca

Published in 2018 by:

UNESCO Institute for Statistics
P.O. Box 6128, Succursale Centre-Ville
Montreal, Quebec H3C 3J7
Canada

Tel: +1 514-343-6880
Email: uis.publications@unesco.org
http://www.uis.unesco.org

Ref: UIS/2018/ED/IP54

Acknowledgements

I am grateful to Danielle Durepos, Elizabeth Fairbairn, Alma Lopez and Lucia Tramonte for comments on drafts of this manuscript, to Jennifer Morgen, who did the programming for the figures used in the report and to other members of The Learning Bar Inc. staff who supported this work at various stages. Many of the ideas in this report stem from a long-term collaboration with Lucia Tramonte, who helped clarify a number concepts such that they can be applied more directly to the research being conducted for the UN's Agenda for Sustainable Development.

The Educational Prosperity model was developed with the support of school leaders across Canada and Australia and was subsequently modified for the OECD's PISA for Development. During its development, I received many helpful comments from educators and directors in several jurisdictions. I am grateful for the contributions of Michael Ward and the OECD staff, members of the Questionnaire Expert Group and International Advisory Group and the national project managers and analysts from the eight countries participating in PISA for Development.

I would like to thank Eduardo Backhoff, David Berliner, Harvey Sanchez, Valérie Tehio and Servaas Van der Berg who conducted comprehensive reviews of the paper, providing many important insights.

Thanks also to Silvia Montoya, Director of the UNESCO Institute for Statistics, for supporting this research and encouraging me to write this report.

Table of contents

Introduction

Countries vary substantially in their levels of students' reading skills, even amongst the wealthiest countries of the world. Within countries, schools also vary considerably in their students' reading skills. These are two of the findings from *Learning Divides: Ten Policy Questions about the Performance and Equity of Schools and Schooling Systems* (Willms, 2006). The term 'learning divides' was used because the analyses revealed large variation among countries in the relationship between students' reading skills and their families' socioeconomic status (SES). The strength of this relationship is an indicator of inequality, or the 'learning divide' between students from low and high SES backgrounds. On average, countries with low levels of inequality had higher overall levels of reading skills. These differences among countries and among schools within countries, were evident for students at the end of the 4th grade and increased as they progressed through school.

The 2006 report was based on analyses of data from the 2001 Progress in International Reading Literacy Study (PIRLS) and the 2000 Programme for International Student Assessment (PISA). PIRLS is an assessment of pupils' reading skills in their fourth year of primary school, which has been conducted every five years since 2001 under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). PISA is an assessment of the knowledge and life skills of 15-year old youth conducted by the Organisation for Economic Co-operation and Development (OECD). It emphasises the reading, mathematics and science skills that students need in their everyday lives when they pursue post-secondary education or enter the workforce. It has been conducted every three years since 2000.

Socioeconomic gradients were used in *Learning Divides* as an over-arching structure to consider questions about educational policy. A socioeconomic gradient, or 'learning bar', simply describes the relationship between a schooling or social outcome and SES for individuals in a specific jurisdiction, such as a school, region, or country (Willms, 2003a). The main goal of educational policy is to improve student outcomes and reduce inequalities associated with family background; in other words, to 'raise and level the learning bar'. The report showed that a detailed analysis of socioeconomic gradients that takes account of the hierarchical structure of the schooling system can provide valuable information for policy-makers on how best to intervene to improve students' skills and reduce inequalities. The report also tackled the question of 'added value': "Is the variation among schools attributable to levels of school resources and to school and classroom policy and practice?" (Willms, 2006, p. 54). The analyses showed that after students' family backgrounds were taken into account, the observed relationships between student achievement and measures of school and classroom practices were small and in most cases, not statistically significant.

This new report is a sequel to the 2006 report. It is based on an assessment framework called Educational Prosperity that can be used to monitor the success of families, communities and public institutions in developing children's cognitive skills and their social, emotional, physical and spiritual well-being. Educational Prosperity embraces a life-course approach, which considers the processes that determine how children's outcomes develop from conception through to late adolescence. The framework includes a core

set of outcomes, called 'Prosperity Outcomes,' for each of six stages of development and a set of family, institutional and community factors, called 'Foundations for Success', which drive these outcomes. It also distinguishes between equality and equity and explicitly links the monitoring of educational systems to national and local policy. The Educational Prosperity framework is described in greater detail a separate paper (Willms, 2018a).

The role of large-scale international studies for informing educational policy has mainly relied on two approaches. One is to collect data on a myriad of school and classroom factors and determine the relationships of these factors with schooling outcomes. This 'quest for school effects' has been a central feature of all large-scale international studies, including PISA and the IEA studies. The contextual frameworks have been based on an input-process-output 'production function' paradigm, attempting to capture the most salient student, family, classroom and school factors that explain student achievement. The analyses of data have been based on multilevel regression models that examine the relationships between a student outcome, such as reading performance and a long list of school and classroom factors. Quite often these statistical models are estimated separately for each country, with the idea that the classroom and school factors relevant to students' skills vary among countries. The results of these analyses are used to support various national policies.

The second approach is for countries to compare their results with those of other countries. The factors considered relevant to student success are grouped into a number of policy themes; the main ones are school resources, accountability, school governance, teaching practices and selective schooling. A country can then ask whether its policies differ from those of countries with a similar social and economic context. However, the international policy community tends to look mainly at success stories. For example, Finland's strong results in the first PISA study spawned numerous accounts about why its schooling system was successful, pointing mainly to the expertise of Finnish teachers and their approach to assessing students (Grek, 2009; Simola, 2005).

Both approaches are problematic. The main issue is that students' performance on the PIRLS tests at the end of 4th grade, or on the PISA tests at age 15, are the *cumulative result* of countless factors that affect children's development, beginning at conception and continuing through to the time of the assessment. Students' cognitive and language skills upon entering the 1st grade are strong predictors of whether they become successful readers two to three years later (Scarborough, 1989; Schatschneider et.al., 2004). The reading skills at the end of primary school are a strong predictor of reading skills at age 15 (Adlof, Catts and Lee, 2010). Therefore, we should not expect measures of school or classroom practices, derived from questionnaires administered at the same time as the achievement tests, to have strong relationships with reading performance.

Moreover, the measures of the key school factors that do affect student performance tend to be inter-correlated and strongly correlated with the average SES of the school. It is virtually impossible to isolate the 'school effects' attributable to particular resources or processes with a cross-sectional study (Raudenbush

and Willms, 1995). The second approach is also problematic for the same reasons. One country may have better reading results than another because it has strong foundations for children's development during the early years and thus any comparison of the effects of school policies can yield spurious results. Simply put, one cannot make causal claims based on findings from national or international studies.

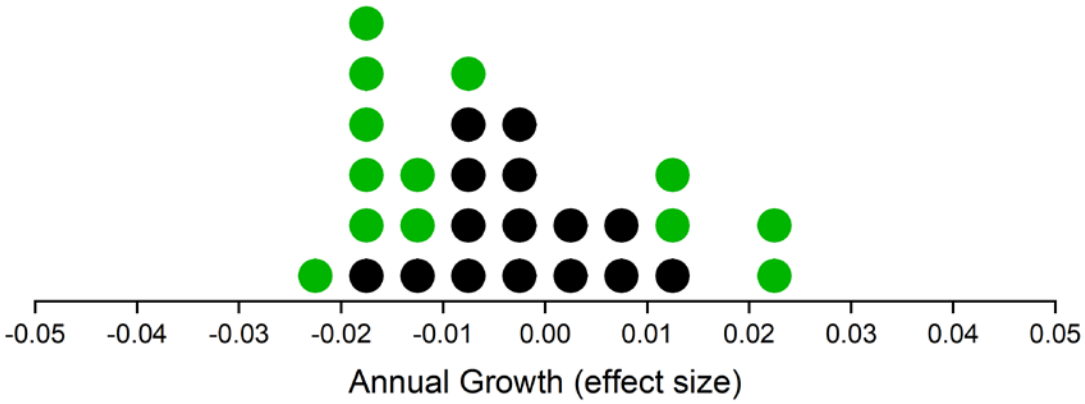
Students' reading skills have not improved over the past 15 years. This is the most compelling reason for adopting a new approach for using data to inform educational policy. **Figure 1** shows the annual rate of growth for the 28 countries that participated in PISA from 2000 through to 2015. The scores for reading performance for the 2000 assessment were scaled to have a mean of 500 and a standard deviation of 100 for all OECD countries. The scales were equated across cycles enabling one to estimate the changes in performance on a common scale.[1] The dots in green indicate that a country's rate of growth was statistically significant. The detailed results are provided in Appendix Table 1.

On average, the annual rate of growth was slightly negative, -0.004% of a standard deviation. The annual growth rates for fifteen of the twenty-eight countries were not statistically significant. Among those with significant annual growth, nine had negative annual growth rates and four had positive annual growth rates. For all countries, the annual rates of growth were less than 2.5% of a standard deviation, either positive or negative. This is only 2.5 points on the PISA scale, which has a mean of 500 and a standard deviation of 100. To put these findings in a broader context, a well-designed reading intervention can bolster students' reading proficiency by 50 or 60 points; that is, an effect size of 0.50 to 0.60. This estimate is based on Hattie's (2009) summary of 50 meta-analyses of reading interventions. The average effect size was 0.51 and for the 14 studies that involved direct phonics instruction, the average effect size was 0.60.

Some of the observed changes in PISA scores may be attributable to measurement and sampling error, resulting in a 'regression to the mean': countries with high scores in 2000 were more likely to have negative annual growth rates, while those with low scores in 2000 were more likely to have positive growth rates. The correlation between countries' PISA scores in 2000 and annual growth rates is -0.39. Finland, the country with the highest score in 2000, had the greatest decline: its SES-adjusted scores fell by almost 2.5% of a standard deviation per year.

[1] The estimates in Figure 1 were based on the full samples of data collected for each country at each cycle and adjusted for SES to take account of changes in the SES of the samples. The analyses were based on a two-level hierarchical 'stability' model fit separately for each country, with SES as a covariate at Level 1 and 'year' at Level 2 (see Willms and Raudenbush, 1989).

Figure 1. Annual growth rates in PISA reading proficiency, 2000-2015



Source: PISA, 2015.

Three premises underlie the approach taken in this report. First, the development of children's reading skills needs to be the primary focus of educational monitoring systems. It is a pre-requisite for the development of strong academic skills at the lower and upper secondary levels and is essential to school completion and social justice (Beswick, Sloat and Willms, 2008; Snow, Burns and Griffin, 1998; Willms, 2006). Second, the literature on classroom and school effects has provided the knowledge we need to build informative educational monitoring systems. We do not require the large-scale national or international studies to continue with the quest for school effects, with numerous measures of classroom and school factors. Instead, we need these studies to focus on a small number of factors, to measure them in greater detail and to track them longitudinally. Third, the results from the large international studies, combined with national studies and small controlled experimental studies, can provide educational administrators with information for setting achievable goals, for allocating resources and for assessing the effects of policies that alter one or more of the structural features of schooling.

This research is not a call for the abandonment of large-scale international studies; indeed, many of the examples used in this report are based on PISA data. The majority of low- and middle-income countries have not yet participated in an international assessment and would benefit by understanding how well their students fare compared with students in other countries. Moreover, the results of comparative studies often serve to increase a country's political will to invest resources in education (Singer and Braun, 2018). Instead, it is intended to shift attention away from the rank-ordering of countries in their outcomes or making causal claims based on cross-sectional data.

The examples presented in Sections 3 and 4 of this report use data for Mexico from PISA 2015. However, it is not intended as an evaluation of the Mexican educational system. Backhoff, Bouzas, Hernández and García (2007) provide a detailed evaluation of results for Mexico following the structure set out in the 2006 report on learning divides (Willms, 2006). Martínez and Díaz (2016) provide detailed results of the PISA 2015 results for Mexico.

The over-arching aim of this report is to provide a structure for using monitoring data to inform two types of educational policies: those concerned with improving schooling outcomes and reducing inequalities; and those pertaining to strategies for achieving educational goals. Throughout this report, the term 'strategies' is used in a broad sense to include courses of action such as reallocating resources among sub-populations of students or schools, the adoption of new curricula or instructional approaches, or changes in a key structural feature of schooling. Fifteen key statistics and five graphical approaches – the 'tools of the trade' – are used for this purpose. The technical details for the estimation of the fifteen statistics are provided in Appendix 2 and examples of the graphical approaches are provided throughout the report.

The first section of this report provides a summary of the Educational Prosperity framework. Section 2 discusses the two most critical transitions for the development of literacy skills. Section 3 is concerned with policies about student performance. How can monitoring data be used to address the question, "How are we doing?". A case is made for identifying vulnerable students and setting realistic goals for their successful development. Section 4 considers strategies for achieving educational goals. Five types of strategies are described alongside a discussion about how monitoring data can be used to discern which types of strategies are most likely to improve student performance. A strategy or intervention can aim to strengthen the Foundations for Success, reduce inequities in the provision of school resources, or alter one or more of the structural features of schooling. Each of these is discussed in this section. Section 5 discusses a method for using monitoring data to set goals and provides an example of a monitoring system based on Educational Prosperity. The section concludes with a discussion of the implications for monitoring schools and school systems based on the Educational Prosperity framework.

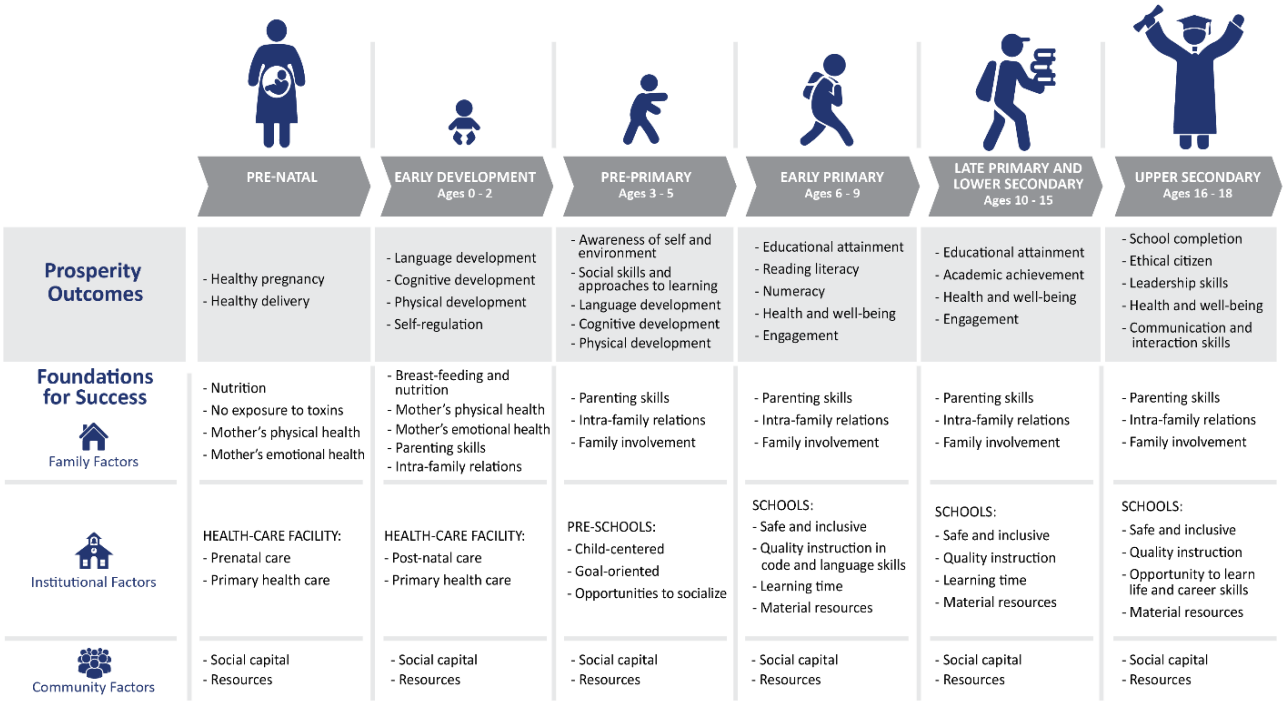
I. Educational Prosperity: A life-course approach

Thriving

Thriving means developing one's best potential. The goal is *to give all children the opportunity to thrive*. Educational Prosperity means that families, communities and public institutions are successful in realising that goal by enabling children to develop academic, social, emotional, physical and spiritual well-being. Educational Prosperity is an assessment framework that countries, school jurisdictions and schools can use to monitor progress towards that goal.

The Educational Prosperity framework (Willms, 2018a) is shown in **Figure 2**. The framework includes six stages of development, covering the period from conception to upper secondary. Each stage includes a small set of outcomes, called 'Prosperity Outcomes' and a number of family, institutional and community factors that drive these outcomes, called "Foundations for Success".

Figure 2. The Educational Prosperity framework



Prosperity Outcomes

The Prosperity Outcomes are considered universal in the sense that they are key markers of child development and are necessary for all children to thrive. Countries may differ in their priorities for the outcomes at various stages, especially in the last stage, but the outcomes included are considered desirable in all contexts. They are also consistent with the Sustainable Development Goals (SDGs) set out by UNESCO (UNESCO Institute for Statistics, 2017).

Foundations for Success

The Foundations for Success are also considered universal in that a large body of research confirms that they are necessary conditions for success at each stage of development. They were chosen based on three criteria: they had to be potent, pervasive and proximal.

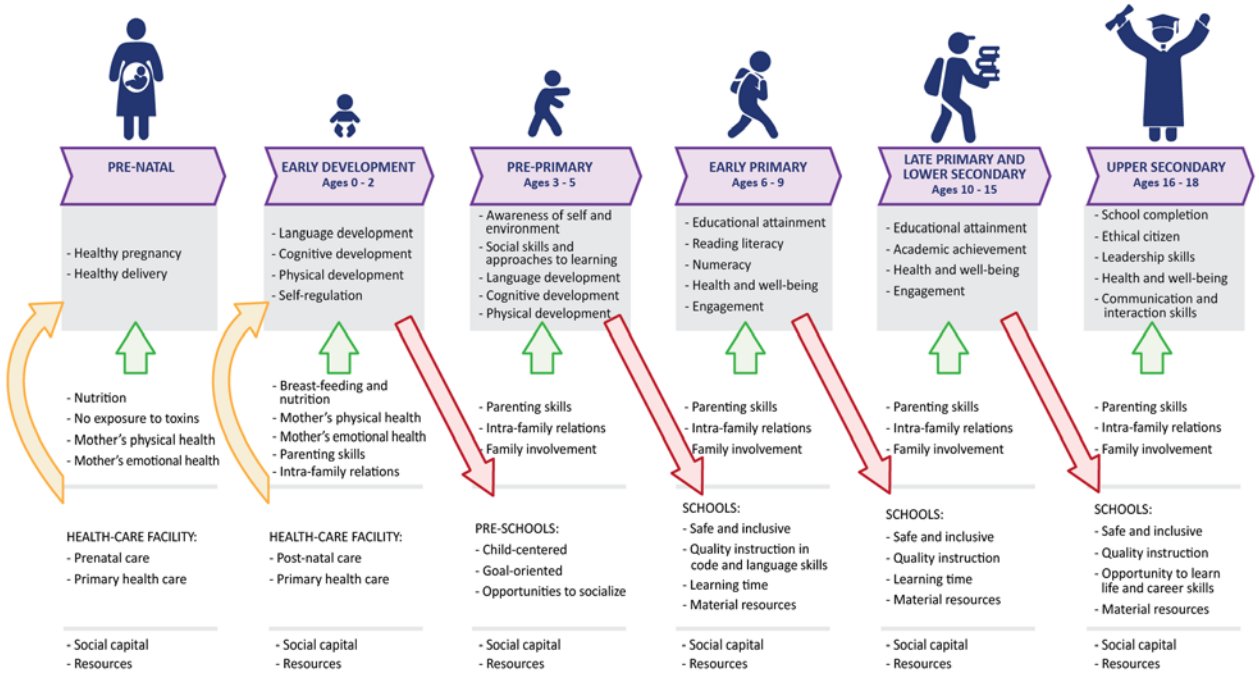
A 'potent factor' is one that has a strong correlation with an outcome and prior research supports claims that it has a *causal* relationship with the outcome. A factor is considered a *causal* factor if it has been shown that it temporally precedes an outcome, is correlated with the outcome and that a change in the factor results in a change in the outcome (Kraemer et. al., 1997). Quality instruction is potent in that it is correlated with academic achievement throughout the schooling period and strategies that improve quality instruction result in better academic achievement (Anderson, 2004; Creemers and Kyriakides, 2006; Kyriakides, Christoforou and Charalambous, 2013; Rosenshine, 2010).

A 'pervasive' factor is positively correlated with a wide range of outcomes and ideally, prior research supports claims of a causal relationship with each outcome. For example, a 'safe and inclusive' school, which is included as a foundation for the last three stages of the framework, affects a wide range of outcomes, including academic achievement, educational attainment, student engagement and health and well-being.

A 'proximal' factor has a direct relationship with an outcome; its effect is not mediated through one or more other factors. Learning time has a direct impact on student outcomes; its effect is not mediated through other factors. Teacher professional development is not considered a foundation factor because it is mediated by several other factors. Its effect is only realised, for example, if it results in improved quality instruction, a more inclusive context, or an increase in family and community support.

Figure 3. Four ways that success accumulates

Orange arrows: biological embedding
Green arrows: foundations for success
Purple arrows: cumulative effects
Red arrows: institutional selection effects



Four ways that success accumulates

The Educational Prosperity framework is based on a developmental model which includes four ways that success accumulates over the life-course from conception to late adolescence. These are shown in **Figure 3**.

Biological embedding. Children's outcomes at birth are affected by the Foundations for Success (green arrow) of the prenatal period: nutrition, no exposure to toxins and the mental and physical health of the mother. To some extent, these outcomes are biologically embedded (orange arrow) through epigenetic processes in which chemical signatures are attached to genes that predispose the child to either vulnerability or resilience (Boyce and Kobor, 2015). Children's early experiences interact with their genetic disposition in ways that affect brain development as well as other neurological and biological systems associated with healthy child development (Boyce, Sokolowski and Robinson, 2012).

At birth, children have billions of neurons. During the course of early development, the neurons form connections called synapses in response to environmental stimuli. As this occurs, many of the neurons that are not being used are pruned away, such that the remaining connections become stronger. The pathways

for vision and hearing are established early, followed closely by those for emotional control, language and cognitive functioning (Center on the Developing Child, 2007; Knudsen, 2004; Shonkoff and Phillips, 2000). This process of synapse formation and neuron pruning – the sculpting of the brain – is more rapid during certain *critical periods* of the first two or three years of life (McEwen and Schmeck; 1994; Cynader and Frost, 1999; Hertzman, 1999).

Foundations for Success. After birth, children's ongoing development is supported by the "Foundations for Success". During the period from birth to age 2, for example, interactions with parents and other caregivers is critical for children's development of language and self-regulation skills (McClelland et. al., 2010). Therefore, parenting styles and intra-family relations are considered to be two of the foundations for this stage as well as for subsequent stages. During the pre-school period and the three schooling periods that follow, a safe and inclusive environment, quality instruction, learning time, material resources and family and community support are Foundations for Success.

Cumulative effects. Children develop Prosperity Outcomes in a cumulative way as they progress from one stage to the next. Language skills develop throughout childhood, but some skills are prerequisites for the development of other skills. For example, the development of strong cognitive skills during the pre-school years determines whether a child will to learn to read well during primary school (Leppänen et. al., 2004; Nation and Snowling, 2004). While family factors play a significant role in the development of Prosperity Outcomes, after age 5 schools play an important and ever-increasing role. Thus, a failure to develop strong skills during the early years increases the risk of school failure.

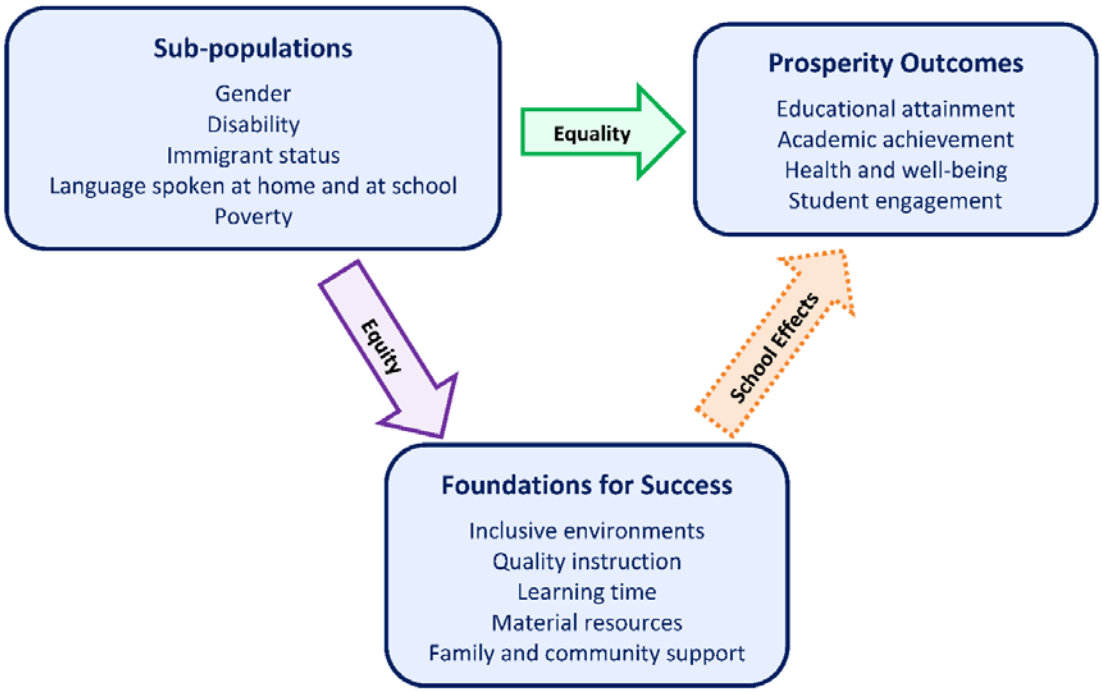
Institutional selection. When students are successful at one stage of development, their life-course can be altered if they are selected into certain classes, school programmes, or schools. In many school systems, children who have strong reading and language skills are streamed into classes or school programmes with strong foundations. These children are more likely to benefit from positive peer interactions, a higher quality of instruction and other factors that enable them to develop their skills at a faster pace (Willms, 2006).

Equality, equity and access

The 1989 UN Convention on the Rights of the Child included statements entitling children to a standard of living adequate for their physical, mental, spiritual, moral and social development; the highest attainable standards of health care; and a quality education, with a view to achieving these rights progressively and on the basis of equal opportunity. Most countries recognise that social and economic development requires universal provision of education from early childhood to adolescence and accordingly governments have established constitutional and legal guarantees for universal provision (Levin, 2009). However, even in the world's wealthiest countries, children from higher socioeconomic backgrounds have greater access to economic, social and educational resources than their peers from lower socioeconomic families (Willms, 2006). Consequently, educational leaders have become increasingly interested in quantifying and monitoring inequalities in educational outcomes among sub-populations and determining the factors associated with these inequalities.

The terms, 'equality', 'equity' and 'access' have been used by researchers in various ways and quite often inter-changeably. This report, following Willms, Tramonte, Duarte and Bos (2012) and Willms (2011), distinguishes between equality and equity, with access treated as an aspect of equity. The Educational Prosperity model was adopted for PISA for Development (PISA-D), an initiative of the OECD which aims to make PISA assessments more relevant to low- and middle-income countries (OECD, 2017). The path model for Educational Prosperity used in PISA-D is shown in **Figure 4** to illustrate the concepts.

Figure 4. Equality, equity and school effects for Educational Prosperity



Equality. Equality refers to differences among sub-populations in the distribution of their educational outcomes. In Figure 4, it is the path linking demographic characteristics, such as gender or disability, to the Prosperity Outcomes (green arrow). The measurement of equality is relatively straightforward. For example, the difference between girls and boys in their average PIRLS reading scores is a measure of equality. Note, however, that the definition refers to differences in the *distribution* of outcomes. For an outcome such as reading scores, an important difference between the sexes is in their distribution of scores. A relevant marker of equality is the prevalence of boys and girls that do not attain some minimum standard. Differences among socioeconomic groups in their outcomes is also a measure of equality. Later in this report, socioeconomic gradients are used to assess inequalities.

Equity. Equity is concerned with fairness – a just treatment of people from different sub-populations. It refers to differences among sub-populations in their access to school and to the resources and schooling processes that affect schooling outcomes. In Figure 4, it is the path linking demographic characteristics to

the Foundations for Success (purple arrow). The measurement of equity has been challenging as one must first identify the resources and schooling processes that have an effect on schooling outcomes. In Figure 4, this is called 'school effects' (orange arrow). These effects are presumed to be 'causal effects'. Large-scale cross-sectional studies cannot be used to determine which factors have strong school effects. The Educational Prosperity framework relies on the broader literature to identify the Foundations for Success.

This distinction between equality and equity is useful as measures of equality indicate the magnitude of differences among sub-populations – how big is the problem? – while measures of equity call for policies that address the problem – what needs to change to reduce inequalities?

Access. Access in education refers to whether schooling is freely available to children in a jurisdiction. The Educational Prosperity model treats access as an aspect of material resources, which calls for questions about equity. For example, "Do children with a disability have equal access to schooling as those without a disability?" The emphasis is on the *provision* of schooling and it is incumbent upon governments and educational institutions to ensure that schools are available locally, they are safe and they have adequate learning materials. Also, access means that educational policies do not create barriers for children to attend school. As a Foundation for Success, access is considered a supply-side factor.

The Educational Prosperity model considers school attendance and whether students complete successive levels of schooling as an aspect of attainment – one of the Prosperity Outcomes. A measure of school attendance incorporates demand-side factors: given there is adequate provision of schooling, not only material resources, but also a safe and inclusive environment, quality instruction, learning time and family and community support, then the question becomes, "To what extent do students attend school?" This depends on a several cultural, social, religious, political and economic factors.

The UN Sustainable Development Goal 4.1 states: "By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes" (United Nations, 2015). This goal includes access, as defined in this report as schooling that is freely available (material resources), school quality (quality instruction, safe and inclusive) and achievement, a Prosperity Outcome. In monitoring progress towards this goal, the distinction between equity and equality is important, as one can ask first about the equity of provision and second, whether equity of provision is related to equality of achievement and attainment. Lewin's (2015) CREATE model is useful as it stresses the importance of secure enrollment that leads to achievement and attainment, as well as the 'zones of exclusion' associated with students falling off track as they proceed through school.

Sub-Populations. The sub-populations of interest can vary among countries and may vary depending on the outcome or the stage of development. For example, the goals of the 2030 Agenda for Sustainable Development (UNESCO Institute for Statistics, 2017) focus on inequalities and inequities associated with gender. The factors listed in Figure 4 – gender, disability, immigrant status, language spoken at home and at school and poverty – were identified by the countries participating in PISA for Development to be the most important in their contexts.

The Role of SES

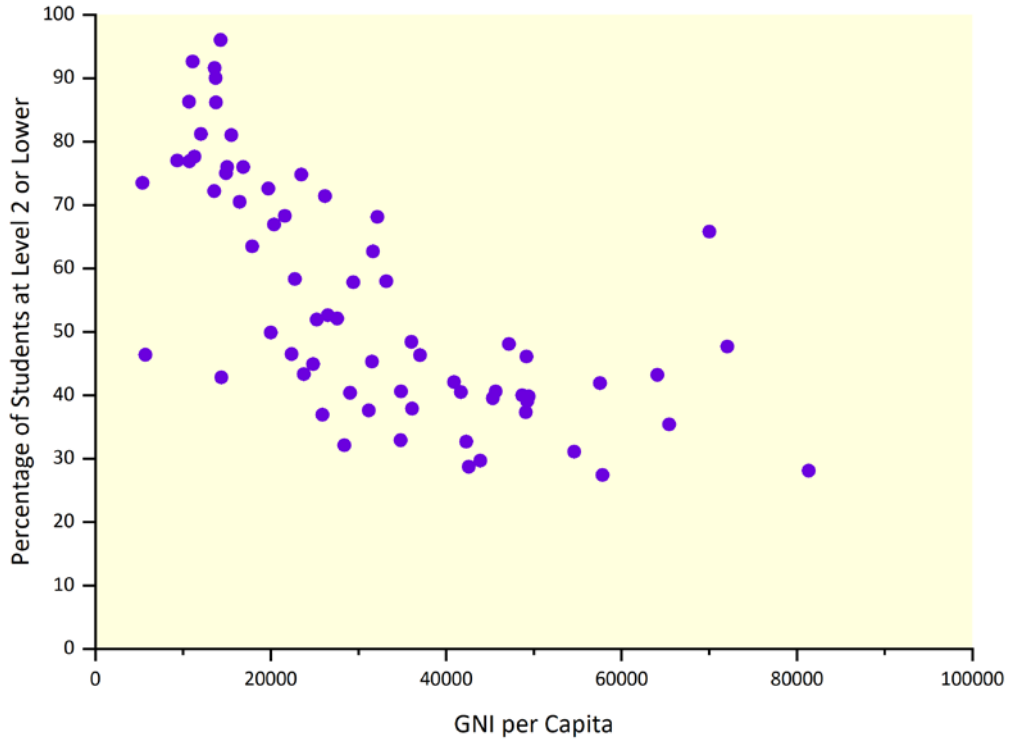
Family SES and poverty play a key role in children's development throughout the life-course, as they are related to the family and community foundations at every stage. SES and poverty also affect whether children have access to strong institutional foundations, especially after age 5 when institutional selection is more prominent. **Figure 5** shows the relationship between the percentage of students who achieved reading scores at Level 2 or lower on the PISA 2015 test and Gross National Income per capita. It clearly shows that levels of vulnerability increase sharply when GNI falls below $30,000.

The life-course approach taken with Educational Prosperity suggests that vulnerability begins early, for many children before they start school. For these children, their potential to become successful readers was likely evident when they entered primary school. In the majority of the low- and middle-income countries, more than 60% of students have poor reading skills at age 15. In high-income countries, about 30% to 50% of students have poor reading skills at age 15. We do not know the extent to which the potential to develop literacy skills is biologically embedded during the prenatal stage or the early development and pre-primary stages, but the evidence increasingly suggests that the early years play a dominant role. PISA for Development includes a comprehensive measure of SES which extends to the lower end of the SES scale to identify children living in extreme or severe poverty.

One cannot assume that when children reach the pre-primary stage, their die is cast; rather, in all countries there is a wide range of pre-literacy skills as children are set to start primary school. The strength of the school system, along with the support of families and communities, affects the subsequent development of all children. The majority of children can learn to read if they receive quality instruction in an inclusive environment with adequate learning time.

All stages of the Educational Prosperity framework are arguably 'critical' for the successful development of literacy skills. However, for the purposes of monitoring the success of a schooling system for informing educational policy, two transitions are especially important. Entry to early primary school, which in most systems occurs at about age 6, marks a change from learning at home, or in an early childhood education and care programme or kindergarten, to the Grade 1 classroom, which tends be a more structured learning setting with a more formal curriculum. The second critical transition is from early primary to late primary, as it involves the transition from learning-to-read to reading-to-learn. These two transitions are discussed in the next section.

Figure 5. Percentage of students with reading skills at Level 2 or lower at age 15 versus gross national income (Atlas method in U.S. dollars)



Source: PISA 2015 and World Bank, 2018a.

II. Two critical transitions for the development of literacy skills

School entry to early primary education

Governments are increasingly recognising that investments in early childhood programmes yield high returns in children's development (Knudsen, Heckman, Cameron and Shonkoff, 2006). As the start of primary school is a key transition, measuring children's developmental skills as they make that transition is valuable for two reasons. It serves as a lagging indicator of the success of society in preparing children for school. It also serves as a leading indicator for determining which children may need extra support during the primary school years.

The term 'school readiness' is often used to indicate the extent to which a child is prepared for success in a more structured environment (UNICEF, 2012). A broader definition incorporates children's skills and the roles of families and communities (Rhode Island Kids Count, 2005):

Ready Families + Ready Communities + Ready Services + Ready Schools = Children Ready for School

A long-standing model for school readiness, stemming from the work of the U.S. National Education Goals Panel, includes five domains: physical well-being and motor development, social and emotional development, approaches to learning, language development and cognition and general knowledge (Barnett, Ayers and Francis, 2015; Kagan, Moore and Bredekamp, 1995; National Early Literacy Panel, 2008).

The link between cognitive and language skills at age 5 with reading skills at age 8 or 9 are well established (Deary, Strand, Smith and Fernandes, 2007; Duncan *et al.,* 2007; McClelland, Morrison and Holmes, 2000; Raver et al., 2011; Rose, 2006). Oral language skills and cognitive ability are especially important (Scarborough, 2001). Attention and self-regulation are also key and appear to be more important than problematic externalising behaviours (Trezesniewski, Moffit, Caspi, Taylor and Maughan, 2006).

The Early Years Evaluation (EYE) or Evaluación Infantil Temprana (EIT in Spanish), is an assessment used to identify the developmental skills of children aged 3 to 6 years as they prepare for and make the transition to formal schooling (The Learning Bar, 2011). It includes measures of the five domains set out by the National Education Goals Panel. It comprises two complementary tools that help educators monitor the overall development of children as they prepare for and make the transition to school. The EYE-Direct Assessment (EYE-DA) is a play-based, interactive assessment conducted by a trained evaluator. The EYE-Teacher Assessment (EYE-TA) provides a systematic framework that teachers can use to structure their observations and informal assessments. It is typically used by kindergarten teachers to provide them with formative, instructionally relevant information.

The EYE-TA was used in a population-based, national evaluation in Uruguay in 2017, following a three-year implementation process that entailed a pilot study to examine the psychometric properties of the assessment; modification of the assessment to fit the cultural context after consultation with curriculum

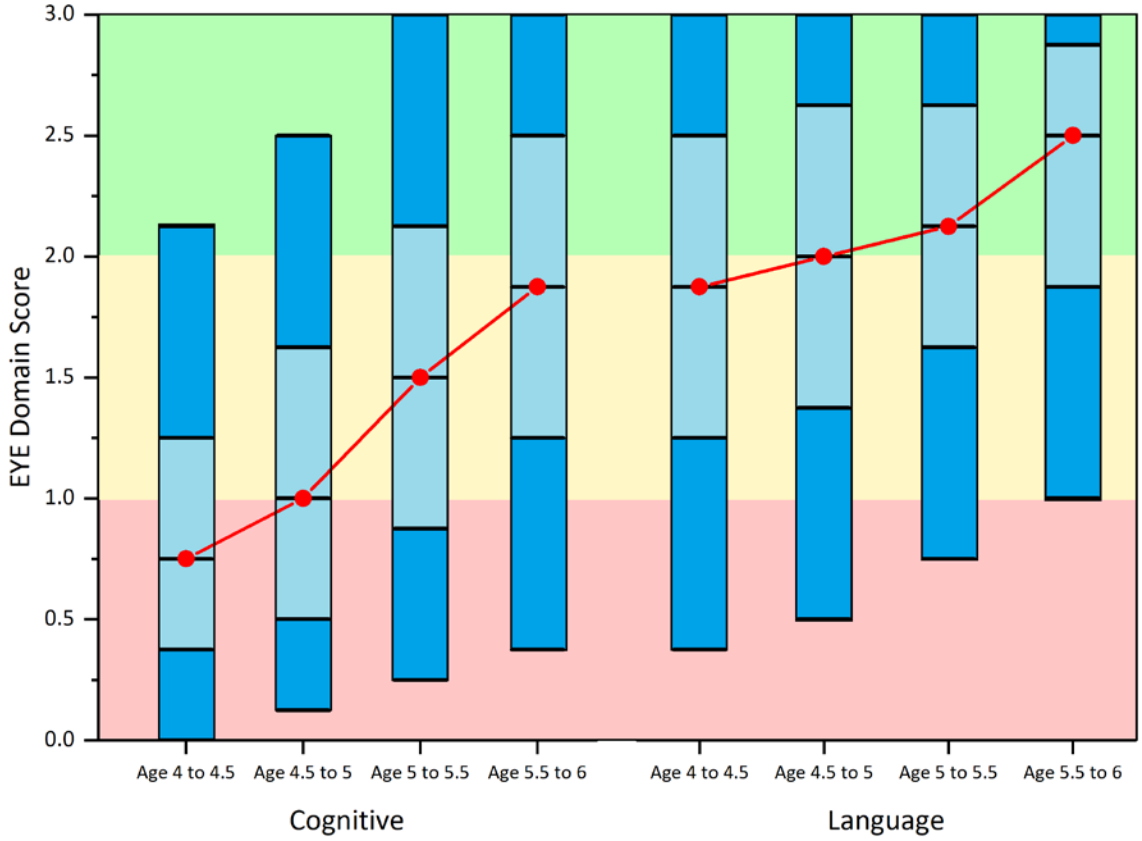
experts and teachers; the involvement of teachers in the development of the implementation process and its use for best practice; and reviews by the Government to ensure the content of the assessment was relevant to kindergarten children's development and contributed to improvements in teaching practices (Lopez, 2016; Willms, 2018b).

Figure 6 shows the range of scores for the full population of Uruguayan 4- and 5-year-old children in 2017. The EYE-TA is treated as a criterion-referenced test, indicating whether children have mastered the tasks required to succeed during primary school. The scores range from zero to three. Children with scores at or above 2 are at Level 1 (green). They are considered to have achieved the tasks in that developmental area. Level 2 (yellow, with scores at or above 1 but below 2) indicates that the child is experiencing some difficulty in achieving the tasks, while Level 3 (red, with scores below 1) indicates that the child is experiencing significant difficulty. The box plots indicate the range of scores at each age, which are shown in half-year increments. The lowest section of each box indicates the range from the 5th to the 25th percentile; the next section depicts the range from the 25th percentile to the median; the next section from the median to the 75th percentile; and the top section from the 75th to 95th percentile.

The findings indicate that for cognitive skills, the majority of children have lower skills at age 4 to 4.5 than for language skills. However, the cognitive skills develop at a faster pace. These results are consistent with recent research on brain development: the synapse formation for language skills peaks earlier than for cognitive function and is more stable by age 5 than that of cognitive function, which does not stabilise until about age 15 (Bhattacharjee, 2015).

These analyses show that for the cognitive skills of 5-year old children, over one-half are vulnerable in the sense that they are experiencing some difficulty or significant difficulty: for those age 5 to 5.5, 38% were experiencing some difficulty and 33% were experiencing significant difficulty, while for those aged 5.5 to 6, 34% were experiencing some difficulty and 49% were experiencing significant difficulty. If one considers all five-year old children, who are set to enter primary school, those in the bottom quartile have skills that are comparable to those aged 4 to 4.5. In other words, they are more than one full year behind their peers in their cognitive skill development. In the next sub-section, the role of cognitive skills in the development of decoding skills is discussed, which is the most important factor determining whether children become successful readers.

Figure 6. Early years evaluation cognitive and language skill domain scores for Uruguay, 2017



Source: The Learning Bar, 2017.

From 'learning-to-read' to 'reading-to-learn'

The timely transition from learning-to-read to reading-to-learn, which for most children occurs at about age 8 or 9, is essential to academic success, school attainment and well-being (Snow, Burns and Griffin, 1998). During primary school, children learn subject-matter content and acquire a variety of skills; however, the focus is on 'learning-to-read'. When children make the transition to late primary, there is a tacit assumption that children can read fluently and understand the content of school subjects such as health, social studies and science. They are expected to 'read-to-learn' – to learn the languages of subject domains and use those languages to think critically, solve problems and create new knowledge. As they progress through school and enter lower secondary, the demands for strong literacy skills increase and students who lack the fundamental reading skills fall further and further behind. A long-standing finding is that the majority of students who are struggling readers at the end of primary school continue to have problems into lower and senior secondary school (Francis et. al., 1996; Warwick, 2005).

Learning to read is a complex process involving phonological awareness, which is the ability to hear and manipulate the sound structure of language; alphabetic understanding, which involves the mapping of print to speech; and fluency or the recognition of words in text (Good, Simmons and Kame'enui, 2001; Storch and Whitehurst, 2002). The widely recognised "Simple View of Reading" (Gough and Tunmer, 1986) postulates that successful reading acquisition depends upon two components: decoding and linguistic comprehension. Decoding is the ability to efficiently recognise familiar and unfamiliar words. The ability to accurately retrieve the phonological code of a written word contributes not only to children's reading development, but also to their vocabulary development and reading comprehension (Verhoeven, van Leeuwe and Vermeer, 2011). Linguistic comprehension is the ability to understand and interpret spoken and written language when they are parts of sentences or other discourse. For a child to become a successful reader, he or she must not only master the ability to accurately decode written words but also understand the meaning of words and how they combine in phrases, sentences and paragraphs.

The ability to decode words is the 'critical filter' during the primary school years. Before children can understand what they are reading, they need to be able to identify words accurately and efficiently and hold the information of a phrase or sentence in their working memory (Perfetti, Landi and Oakhill, 2005; Vellutino and Scanlon, 1987). The development of decoding skills reinforces the development of comprehension skills; however, the majority of children who do not learn to read well have difficulties with decoding skills (Storch and Whitehurst, 2002; Verhoeven, van Leeuwe and Vermeer, 2011). Effective teachers develop a repertoire of instructional strategies for teaching phonemic awareness and word recognition; they achieve the best results when the teaching of decoding skills includes opportunities to read connected text (Torgesen, Otaiba and Grek, 2005).

Confident Learners is a whole-school literacy programme aimed at improving the literacy skills of Indigenous children during the primary grades (The Learning Bar Inc., 2016). It was developed with and for Indigenous educators in Canada, with the support of an Indigenous Advisory Circle. A key feature of the programme is a 'pathway approach' to assessment and instruction based on the simple view of reading. It includes a professional learning programme for teachers aimed at increasing their knowledge of children's literacy development and the use of high-yield teaching strategies to teach decoding and language skills. The programme includes 20 modules or 'steps' for decoding skills and 20 steps for language skills, with each step comprising 14 skill-based objectives. The objectives for each step are linked to one or more fun, engaging and culturally-relevant learning activities. Teachers monitor students' progress by regularly conducting assessments to determine whether students have mastered the objectives for each step as they proceed on the pathway.

During the initial stages in developing Confident Learners, three literacy experts amassed the curricula from all Canadian provinces, some US states and from Australia and the UK. They set out a model describing the skills at a micro-level that are required for decoding words and developing the language skills required for learning to read during the primary school years. The model resulted in over 300 'coding' skills and 300 'language' skills covering the period from Kindergarten to Grade 3. An unexpected finding was that over one-

half of the essential decoding skills are taught in most curricula from the beginning of Grade 1 through to the first half of Grade 2. The research team refers to this phenomenon as the 'reading mountain'.

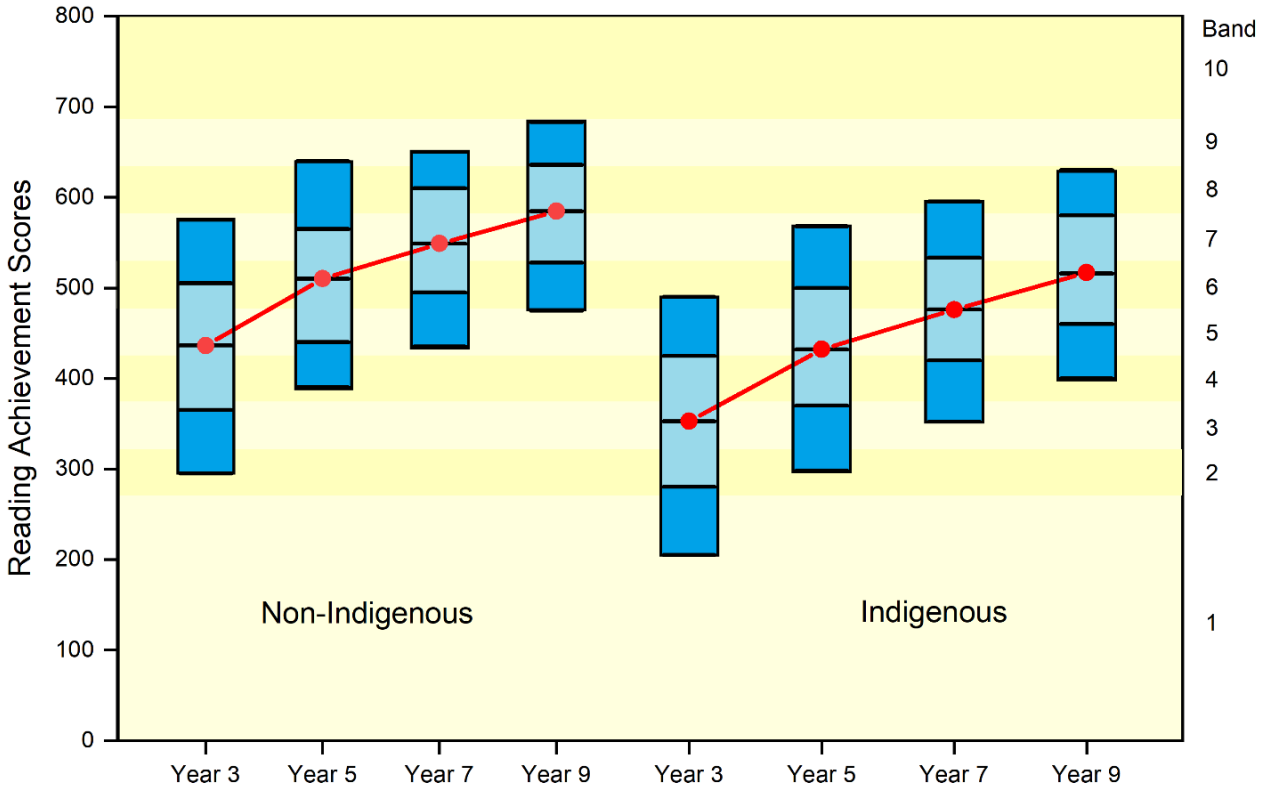
The problem facing students who begin primary school without the fundamental pre-literacy skills is that they cannot traverse the reading mountain. Results from the Early Years Evaluation indicate that in many jurisdictions and for certain sub-populations, about one-quarter to one-half of children enter primary school with cognitive and language skills that are one or more years behind their peers. When faced with a grade-based curriculum in the 1st grade, they are unable to master the skills required to proceed to the next grade. The three most prominent strategies to address this issue are grade repetition, classifying children as special needs and segregating them into special classes, or 'waiting to fail'; that is, having them proceed to Grade 2 with the hope they will catch up. None of these strategies are effective in increasing the prevalence of children who make a successful transition from learning-to-read to reading-to-learn.

The 'reading mountain' phenomenon and the slow rate of improvement of reading skills after primary school are evident in the results of Australia's National Assessment Program of Literacy and Numeracy (NAPLAN; Australian Curriculum, Assessment and Reporting Authority, 2017). **Figure 7** shows the range of reading scores from Year 3 (average age is 8 years, 3 months) to Year 9 (average age is 14 years, 7 months) for non-Indigenous and Indigenous students. The box plots indicate the range of scores for each year. The lowest section of each box indicates the range from the 5th to the 20th percentile; the next section from the 20th percentile to the mean; the next section from the mean to the 80th percentile and the top section from the 80th to the 95th percentile.

Three aspects of these findings are relevant to the development of students' reading skills. First, the rate of growth from Year 3 to Year 5 is slow and is even slower after Year 5. The average score at Year 3 for non-Indigenous students was 436.3 (SD= 84.1) and at Year 5 it was 510.2 (SD=74.4). The average growth per year, as an effect size, is 0.44 per year. The average growth per year from Year 5 to Year 7 (Mean=548.9, SD=66.7) was 0.26, while the average growth per year from Year 7 to Year 9 (Mean=584.6, SD=63.6.7) was 0.24.

Second, the rates of growth for Indigenous students were comparable: from Year 3 (Mean=352.8, SD=92.6) to Year 5 (Mean=432.1, SD=81.3) it was 0.43 per year, from Year 5 to Year 7 (Mean=476.0, SD=75.0) it was 0.27 per year and from Year 7 to Year 9 (Mean=516.9, SD=73.0) it was 0.27 per year. The major difference in the scores for Indigenous students compared with non-Indigenous students is in their scores at Year 3. This achievement gap may be attributable to differences in their life-course experiences before they entered primary school or to differences in their experiences at school during the early primary stage. Data are not available to discern the extent to which the achievement gap was evident when children were in kindergarten or increased during the period from kindergarten to Year 3. An understanding of this relationship and the provision of schooling during the primary years has important implications for the monitoring of the education system.

Figure 7. Reading achievement scores for Australia, based on NAPLAN 2017



Source: ACARA, 2017.

Students in Year 9 with reading scores in the bottom quintile had skills comparable to students who were at about the median in Year 5. This is the case for both Indigenous and non-Indigenous students. If students who were in the bottom quintile in Year 5 received an intervention that bolstered their scores by an extra one-quarter of a standard deviation per year, they would have scores in the middle of the distribution by Year 9.

Grade Effects in PISA. The relatively small growth in reading scores after Year 9 in Australia and more generally across OECD countries is also evident in PISA. The target age for students participating in PISA is 15. In most countries, among students who have not repeated a grade, there are 15-year old students in at least two grades. If one excludes the students who have repeated a grade, those who are in the grade that is one year lower than the modal grade are in that grade because of their birthdate; that is, they are representative of 15-year-olds who are progressing through the grades on schedule. In Australia, for example, among students who had not repeated a grade, 9% were in Year 9, 76% in Year 10 (the modal grade) and 15% in Year 11. Therefore, it is possible to estimate a 'grade effect' on reading scores associated

with a one grade increase, from one grade behind the modal grade to the modal grade and from the modal grade to one grade above the modal grade.

For all OECD countries, the grade effect associated with a student being in the modal grade versus one grade lower was 55 points, while the effect associated with being in the grade that was one grade higher was 33 points. Overall, the 'grade effect' was 43 points. For non-OECD countries, the corresponding grade effects were: 41 points for being in the modal grade versus the one grade lower, 24 points for being in the higher grade and an overall 'grade effect' of 29 points. The overall 'grade effect' for all participating countries was 36 points. [2]

When considering PISA results, PISA points translate directly to effect sizes, with 100 points representing an effect size of 1.0. We can also consider a difference of 43 points to represent one year of schooling, or 4.3 points to be about one month of schooling in a ten-month school year.

Learning to read in non-English languages. The rate at which children become proficient in decoding words depends on the orthographic depth of the language. English and French have a relatively deep orthography; that is, the relationship between letters and speech sounds (phonemes) is less consistent than in languages with shallow orthographies, such as Finnish, Italian and Spanish. Consequently, learning to read in English is slower than in a language with a shallow orthography, such as Spanish (Caravolas et. al., 2013). The successful decoding of words is a prerequisite for reading in all alphabetic languages, but in languages with a shallow orthography, linguistic comprehension plays a more important role in early reading development. Studies of early reading growth of Spanish students, based on the simple view of reading, confirm the importance of linguistic comprehension over decoding (Polo, Araujo and Salceda, 2017; Ripoll, Aguado and Castilla-Earls, 2014).

These findings have important implications for reading interventions designed for low- and middle-income countries. In many African countries, for example, children are taught to read in an official national language which may or may not be their first language, or they are taught in their first language during the primary years and then continue in the national language after the early primary stage. If the initial language of instruction is the children's first language, a reading intervention such as Confident Learners might

[2] This analysis used a technique called Hierarchical Linear Models (HLM), with students nested within countries. Thus, the estimated effect is an average effect across all countries. The analysis revealed that the grade effect varied considerably among countries; four countries had a grade effect of less than 5 points, while eight countries had a grade effect greater than 65 points. The PISA grade effects for Australia, from Year 9 to Year 10 was 16 points and from Year 10 to Year 11 was 12 points. These finding are consistent with the NAPLAN results discussed above, which show declining rates of growth after Year 5. (An HLM model was also fitted with students' age as a variable and the results revealed that the effects of age were negligible; students that were young for their cohort lagged behind those who were old for their cohort by less than 1 point.)

emphasise decoding skills over linguistic comprehension skills during the primary grades. The height of the 'reading mountain' will differ, depending on the orthographic depth of the language. If the language of instruction is not the children's first language, then an intervention would need to place greater emphasis on linguistic development skills, as these skills are less likely to be reinforced at home or in the local community. The relative emphasis to place on decoding skills versus linguistic comprehension skills is key to education policies in South Africa. After apartheid ended in 1994, the constitution gave official status to nine Indigenous languages, as well as Afrikaans and English (Heugh, 2013). In most cases, children are taught in their first language until Grade 3 and then switch to English (Manyike, 2012). This is likely to be a challenge for the majority of students who have not mastered the decoding skills required for learning to read in English.

Poverty also plays an important role in determining the most appropriate intervention for developing pre-literacy skills during the pre-literacy stage and for teaching reading during the early primary school years. The link between poverty and children's pre-literacy skills is well established (World Bank, 2018b). Children living in poverty are more likely to experience nutrient deficiency, infectious diseases and toxic environments during the first two stages of development, from conception to age 2 and the consequences of these risk factors can be biologically embedded (Black *et al*., 2008). As children mature during the early years, they are also at risk of receiving less stimulation, direction and support (Black *et al*., 2017). The data for Uruguay based on the Early Years Evaluation discussed above, indicate that when children enter the 1st grade, over one-quarter of them are more than one full year behind their peers in their cognitive skill development. The percentage of such vulnerable children is greater in Departments that have a higher percentage of children living in poverty. Thus, if one were to design a reading intervention for Uruguayan children, the development of decoding skills might play a dominant role in Departments with high levels of poverty, while the development of linguistic comprehension skills may be more important in the wealthier Departments.

III. Policies about student outcomes and equality

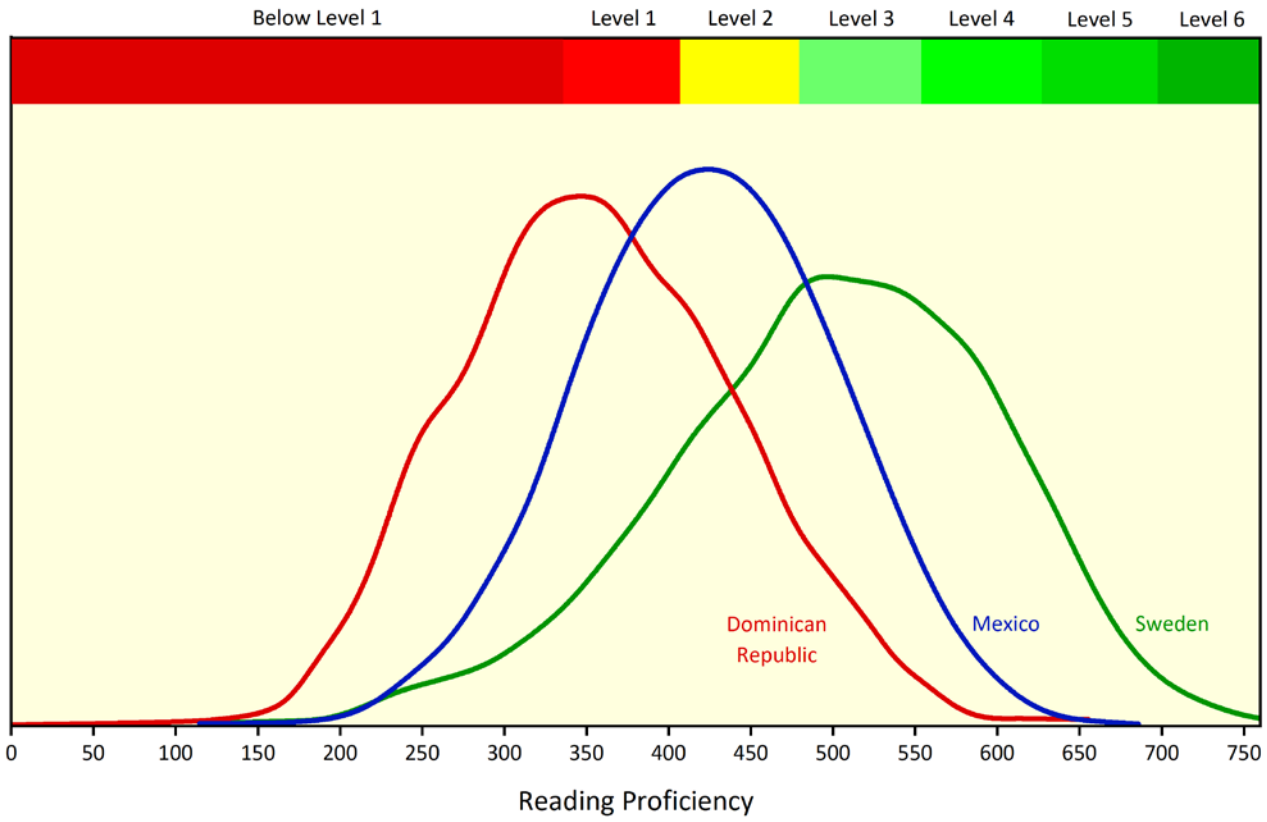
The over-arching goal of educational policy is to ensure all children are meeting their potential – *thriving*. Educational policies establish a course of action that entails setting goals and identifying strategies for achieving them. Monitoring data can inform the policy process by characterising the performance of the school system. The most basic question is, "How are we doing?". Social indicators typically derive their meaning in one of three ways: through comparisons to some standard, such as a national or international average; through comparisons among jurisdictions, such as comparisons of schools within a province or state; and by an examination of trends over time. For example, one can ask, "How are we doing compared with international norms?", "How are we doing compared with other countries that have similar levels of resources?", or "Are we improving our outcomes over time?" The answers to these questions inevitably lead to more questions, "Who is vulnerable – are there inequalities in outcomes?" and "Where are our most vulnerable children?" With answers to these questions, educational administrators can set realistic, measurable goals for improving student outcomes and reducing inequalities. This section is about Prosperity Outcomes. Its aim is to discuss the kinds of indicators that are most useful and the analyses and reporting tools that can inform policies about student outcomes and equality. Section 5 discusses how this information can be used to set goals.

How are we doing?

Focus on outcomes. Student attainment – how far students go in school – is the best place to start. For low- and middle-income countries, the first question is whether students attend pre-primary education or begin primary education. Most children enroll in primary school and successive cohorts tend to go further in school than those before them (World Bank, 2018b). However, some children are never enrolled in school, in some cases because they are in remote areas or because they have a disability. In many low-income countries, the enrollment rates for girls is substantially lower than that of boys (World Bank, 2018b). Thus, two key markers are simply the percentage of students who enroll in pre-primary and primary school. At this stage, measures of children's pre-literacy skills are also important; they serve as lagging indicators of children's development since conception and as leading indicators of the resources required during early primary school. With these indicators in hand, one can then ask questions about which sub-populations are vulnerable and where vulnerable children reside.

The next key marker of educational attainment is the percentage of children who make the transition from early primary to late primary. As students make that transition, one also wants measures of literacy and numeracy skills, health and well-being and engagement. However, an indicator of the percentage making the transition is required; without it, indicators of literacy or numeracy are of little value as they likely pertain to a select group of students. The same principle applies for the transition from late primary to lower secondary and from lower secondary to upper secondary. The scores from international assessments such as PIRLS, TIMSS or PISA have little utility for comparative purposes unless the results are considered alongside indicators of the percentage of children not attending school.

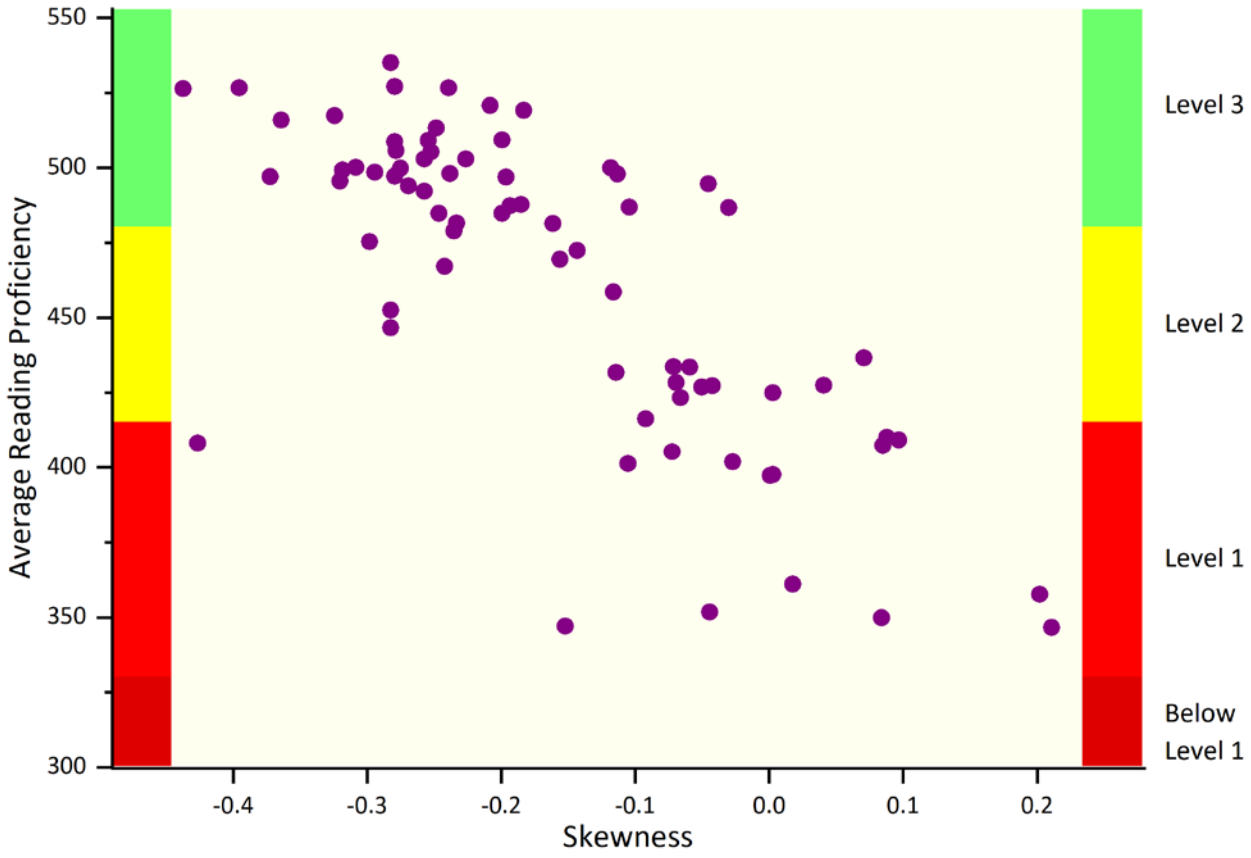
Figure 8. Distribution of reading proficiency at age 15 for Dominican Republic, Mexico and Sweden



Source: PISA, 2015.

For continuous measures, such as test scores, the aim is to describe the distribution of scores with a small set of summary statistics that can be used to track progress from year to year or be used to gauge performance compared to a standard or with other jurisdictions. **Figure 8** shows the distributions of scores for Dominican Republic, Mexico and Sweden. The two most common markers used to track progress are the mean and the median. However, three other statistics are also useful: the standard deviation, the skewness and the percentage of students who are 'vulnerable'. The standard deviation is an indicator of the amount of variation in a set of scores, or their spread. The inter-quartile range – the range from the 25th to the 75th percentiles – which was depicted for the EYE and NAPLAN scores above, can also be used for this purpose. Skewness indicates the extent to which a distribution is asymmetrical. Distributions that are negatively skewed have low scores that extend further below the mean than the high scores extend above it; the reverse is the case for positively skewed distributions.

Figure 9. Relationship between average reading proficiency and skewness



Source: PISA, 2015.

Skewness is important because as school systems develop they do not shift the entire distribution of scores, which may be positively skewed or even symmetrical (i.e., not skewed); rather, they tend to shift the scores of the more able students, leaving behind a sub-population that is vulnerable. For the three countries shown in Figure 8, the scores for Dominican Republic are positively skewed (skewness = 0.20), those of Mexico are almost symmetrical (skewness = -0.07), while those of Sweden are negatively skewed (skewness = -0.31). **Figure 9** shows the relationship between mean reading scores versus the skewness of the distribution for countries that participated in PISA 2015. The correlation between mean reading scores and the standard deviation is quite small, 0.12, but the correlation between mean reading scores and skewness is -0.74. Jordan is an outlier in this analysis with a low mean score (408) and negative skewness (-0.43). When it is excluded, the correlation between mean scores and skewness is -0.80. Most of the countries with low scores tend to be the low- and middle-income countries *(see Figure 5),* which tend to have lower transition rates into lower secondary. If one had reading scores for the out-of-school students, the negative relationship between mean scores and skewness would undoubtedly be stronger.

Focus on vulnerability. In the monitoring of children's development, a subject of debate has been whether to emphasise children's strengths versus their deficits (Ben-Arieh and Frønes, 2007). The literature on child development focuses mainly on children's deficits, such as behavioural and health problems. Efforts to monitor the effectiveness of schools and school systems have mainly stressed children's strengths with measures of academic achievement, although sometimes the focus is on 'reading failure' or 'dropping out of school'.

Closely aligned to the deficit *versus* strengths debate is whether to use dichotomous or continuous measures. For policy purposes, dichotomous measures are often preferred because they bring attention to the social and economic issues relevant to vulnerable children. They are also easier for people to understand; for example, most people would be able to interpret a statement such as "The prevalence of youth suffering depression increased from 8% to 10%" rather than "The average depression score increased from 8.5 to 8.9 on a 10-point scale."

The approach taken with Educational Prosperity is to favour strength-based measures, as the overarching goal is giving all children the opportunity to thrive. Policy-makers may find it easier to rally public support for increasing the prevalence of school-age children who are physically fit than for reducing childhood obesity. The goals of the 2030 Agenda for Sustainable Development (UNESCO Institute for Statistics, 2017) are couched in positive terms. However, some outcomes, such as poor mental health or physical health problems, are best discussed as deficits. Suffering anxiety or depression is not the same as having a low score on an index of happiness. Similarly, policy-makers may find it easier to focus on the prevalence of children with poor reading scores than those who succeeded in passing some threshold.

Similarly, continuous measures are generally preferred over dichotomous measures as they include more information. However, a country may be successful in reducing the prevalence of children with low reading scores, which may not be evident in changes in mean scores. Also, in many of the countries that participate in international assessments, a substantial number of students score at or near the floor of the test. The 'true score' of these students may be even lower than their estimated test score (Nonoyama-Tarumi and Willms, 2010). In this case, a focus on vulnerability is preferable. Some measures have well-established cut-points for defining vulnerability, such as low birth weight or childhood obesity. The international educational studies include cut-points for benchmarks or levels of achievement. The decision on which cut-point is appropriate depends on the outcome considered and its distribution for the population being considered.

Data from PISA 2015 show that the mean reading score for Mexico is 423. The standard deviation is 78 and the skewness is -0.07. The percentage of students considered vulnerable – those who did not achieve at least Level 3 – is 76.0%. The sampling designs of the major international studies typically entail sampling schools in a first stage and then students at the second stage from each of the sampled schools. Also, the samples for most countries are stratified to ensure that particular regions or sectors are adequately represented. These studies use a rotated test design in which students complete only a portion of the assessment tasks. Consequently, the estimation of statistics requires the use of a set of *replicate weights* to account for the

sample design and when the statistics pertain to test scores, the estimation also requires the use of *plausible values* to account for the test design. The techniques for estimating these statistics and a brief description of replicate weights and plausible values are provided in Appendix 2.

Who is vulnerable?

Socioeconomic gradients. A *socioeconomic gradient*, or 'learning bar', describes the relationship between a social outcome and SES for individuals in a specific jurisdiction, such as a school, a community, a province or state, or a country (Willms, 2003a; 2006). The outcome can be any social outcome, such as children's health status, test scores, or educational attainment. SES generally refers to the relative position of a family or individual on a hierarchical social structure, based on their access to, or control over, wealth, prestige and power (Mueller and Parcel, 1981). Measures of SES play an important role in educational monitoring as they provide a context for setting attainable goals and measuring progress towards them. They also provide an approach for assessing equality of outcomes and equity of provision; considering the potential of strategies aimed at improving student outcomes and reducing inequalities; and strengthening the validity of research studies (Willms and Tramonte, 2018).

Figure 10 shows the socioeconomic gradient for students' reading scores at age 15 in Mexico, based on the PISA 2015 data. The measure of SES is the composite measure used in PISA, called Economic Social and Cultural Status. A socioeconomic gradient is comprised of three components: the level, the slope and the strength of the outcome-SES relationship.

a. The *level* of the gradient is defined as the expected score on the outcome measure for a person with average SES. The level of a gradient for a country (or for a province, state or school) is an indicator of its average performance, after taking account of students' socioeconomic status. The level of the Mexican gradient is 450.3. For this analysis, the measure of SES was scaled such that zero represented the OECD mean. Therefore, the level refers to the expected score of a student whose SES was comparable to a hypothetical student with an average SES equal to the OECD mean.

b. The *slope* of the gradient indicates the extent of inequality attributable to SES. Steeper gradients indicate a greater impact of SES on student performance – that is, more inequality – while more gradual gradients indicate a lower impact of SES – that is, less inequality. The slope of the Mexican gradient is 21.2 (in the centre of the data), which indicates that the expected reading performance increases by 21.2 points for a one standard deviation increase in SES. The Mexican gradient is slightly curvilinear, with the slope decreasing slightly with rising levels of SES. However, the coefficient for the curvilinear component is not statistically significant.

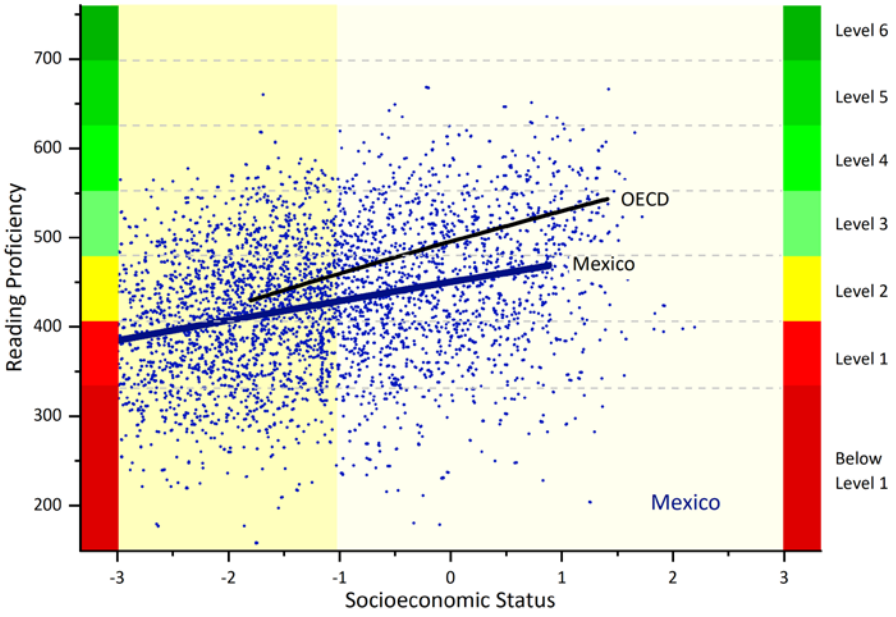
c. The *strength* of the gradient refers to the proportion of variance in the social outcome that is explained by SES. If the strength of the relationship is strong, then a considerable amount of the variation in the outcome measure is associated with SES, whereas a weak relationship indicates that relatively little of

the variation is associated with SES. The most common measure of the strength of the relationship is a measure called R-squared, which for Mexico is 0.12.

The gradient line for Mexico, shown in blue, is drawn from the 5th to the 95th percentile of the SES scores. For Mexico, the 5th and 95th percentiles are -3.08 and 0.88 respectively. Therefore, 90% of Mexican students fall within this range. Students in Mexico on average have a lower SES than those in other OECD countries. The graph also shows the reading performance and SES for a representative sample of 5,000 Mexican students. These are shown with the small blue dots above and below the gradient line. They show that there is considerable variation in reading performance at all levels of SES. The gradient line for all OECD countries is shown in black. The 5th and 95th percentiles for OECD students are -1.78 and 1.40 respectively.

The term 'learning bar' is used as a metaphor for the socioeconomic gradient. The central question facing most schools and countries is "How can we raise and level the learning bar?" Increasing educational performance and reducing inequalities among students from differing socioeconomic backgrounds can be achieved in a number of ways. The approach that may work best depends on social and political issues, but it also depends on the distribution of students' outcomes and SES within and among schools and how these factors are related to and interact with the Foundations for Success.

Figure 10. Socioeconomic gradient for reading proficiency for Mexico



Source: PISA, 2015.

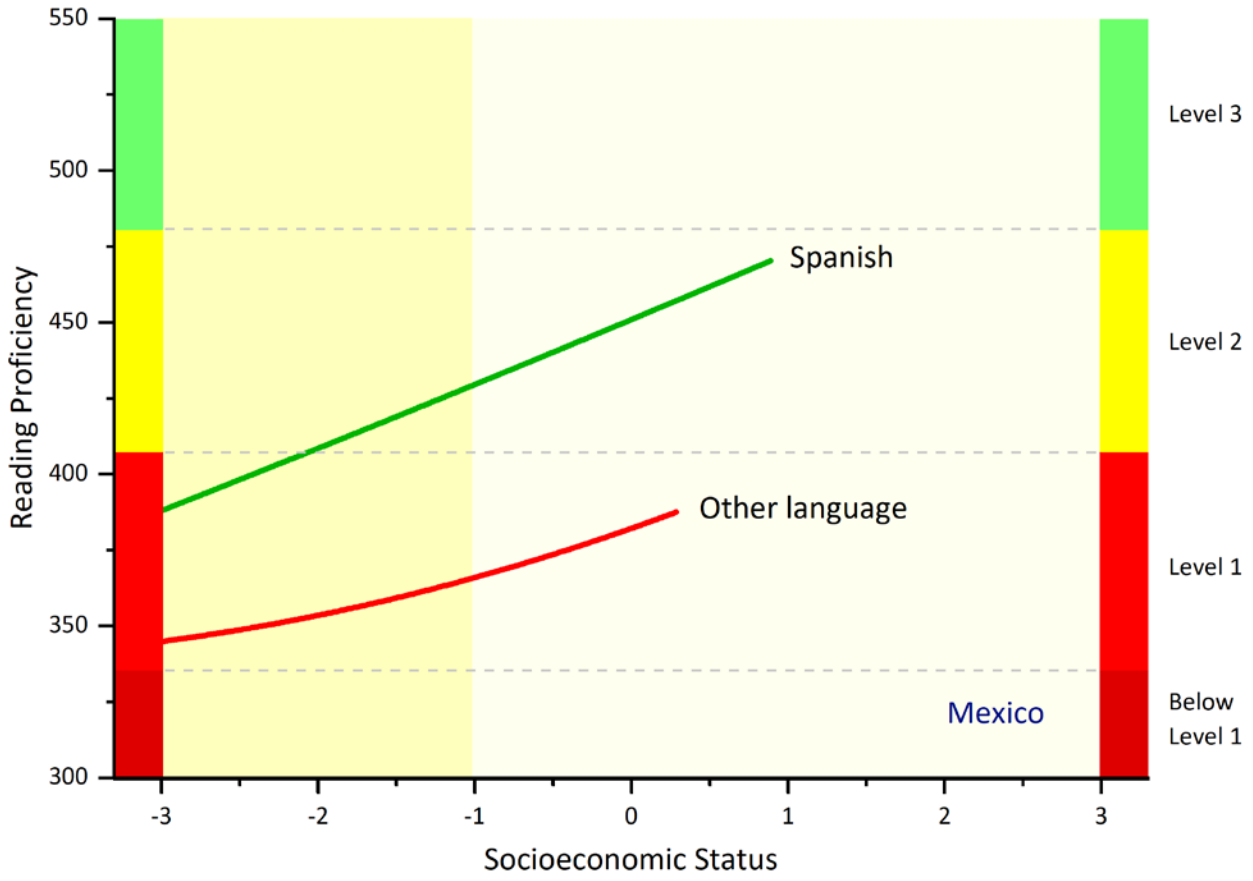
Diminishing returns. The "hypothesis of diminishing returns" holds that there are weaker effects on social outcomes associated with SES at higher levels of SES. One might predict, for example, that above a certain level of SES, there would be little or no increase in students' reading achievement associated with SES. This does not appear to be the case for Mexico; the gradient line shows that average reading performance increases linearly with increasing SES. The estimation of the gradient line includes a quadratic term for SES, which is the statistic used to gauge whether returns to SES increase or decrease with rising levels of SES. For Mexico, the coefficient for the linear term (i.e., the slope) was 21.2, while the quadratic term was -0.3. This means that there is a diminishing return for increasing levels of SES, but in this case, it is not statistically significant.

Appendix 2 describes the statistical techniques for the estimation of the statistics for socioeconomic gradients.

Willms and Somers (2001) found *increasing* returns for SES for reading and mathematics achievement of Grade 3 and 4 students in several Latin American countries. They suggested that there was a premium associated with parents having completed secondary school. For students at age 15, an indication of increasing returns may be the result of students not having made a successful transition from learning-to-read to reading-to-learn at the end of lower primary and therefore were unable to benefit as much from instruction in later grades. It may also be attributable to a floor effect on the PISA test: in most low- and middle-income countries, a substantial percentage of students with low SES are unable to successfully answer even the easiest test items.

Equality. A straightforward measure of equality is the difference in test scores between two sub-populations. The difference in reading proficiency for students whose first language is Spanish compared to those whose first language is not Spanish is approximately 74 points. The SES gradient approach allows one to assess how large the difference is after controlling for SES. It also allows one to discern whether there is a significant interaction associated with SES; that is, does the gap vary at differing levels of SES? **Figure 11** shows the gradient lines for these two sub-populations. In this case, SES only partially accounts for the differences between the two sub-populations; the difference for students with average SES is about 69 points. The figure also shows that the difference is fairly uniform across the range of SES.

Figure 11. Socioeconomic gradients for reading proficiency by first language for Mexico

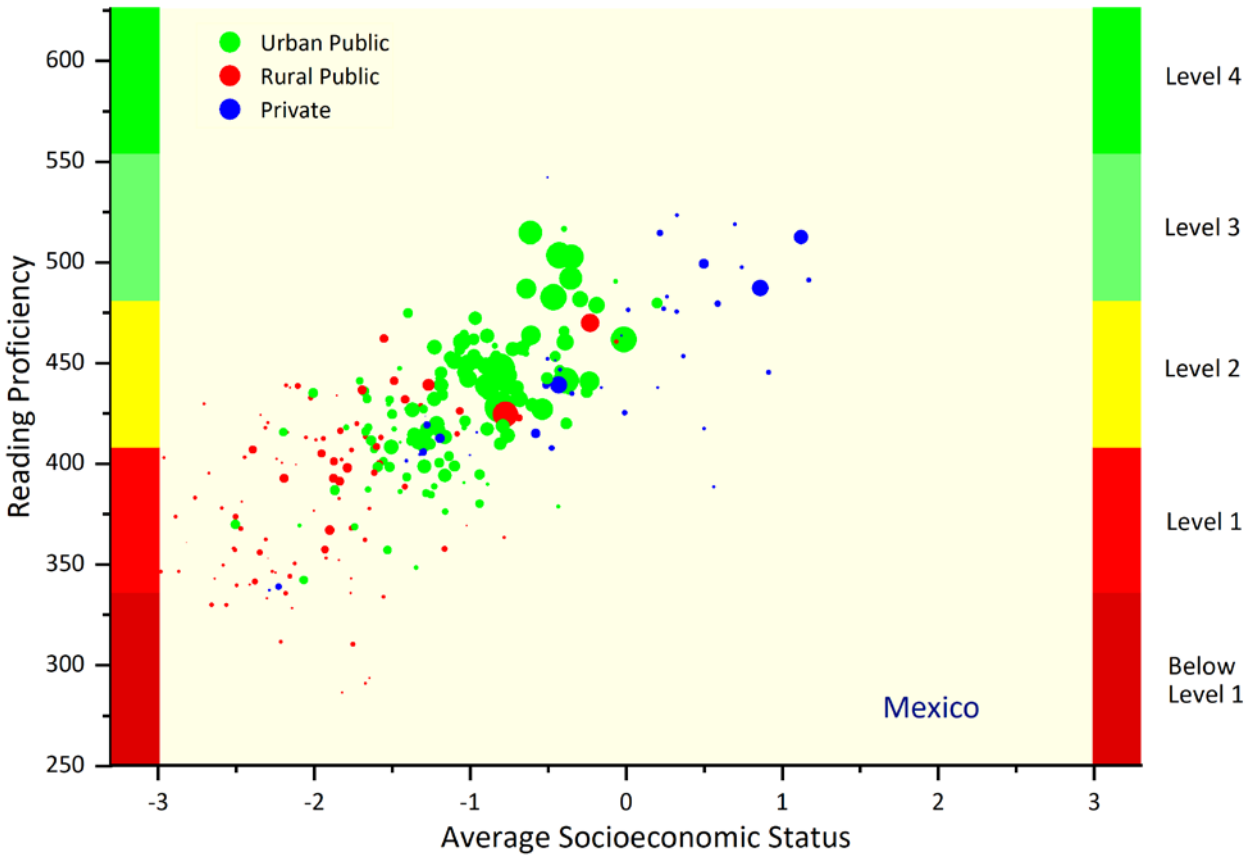


Source: PISA, 2015.

Where are the vulnerable children?

School profiles. A school profile is simply a scatter-plot of the school mean scores for an outcome plotted against school mean SES. **Figure 12** shows the school profile for reading scores for Mexico. Each of the dots represents one of the schools that participated in the PISA 2015 study. The size of the dots is proportional to the square root of a school's enrolment. The colours of the dots indicate whether the school is a rural public (red), urban public (green), or private (blue) school.

Figure 12. School profile for reading proficiency in Mexico



Source: PISA, 2015.

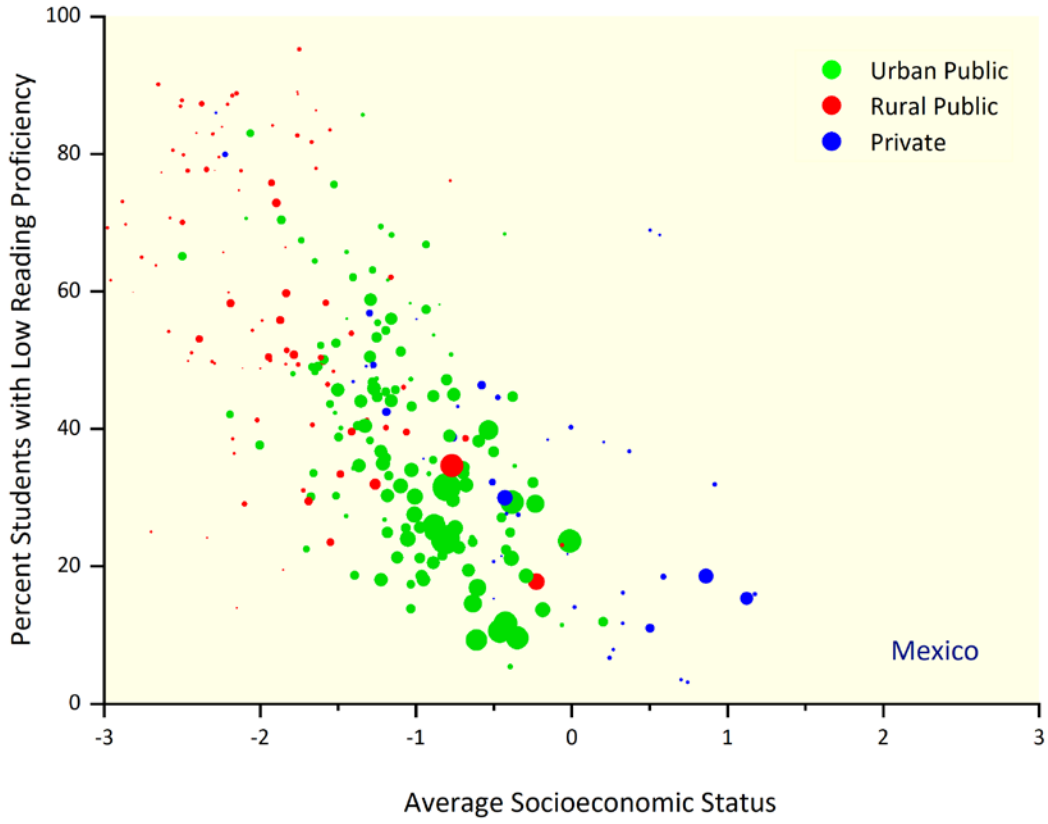
Three findings are immediately evident from this school profile.

a. Schools differ substantially in their average reading scores. At all levels of SES, the range between the schools with the lowest reading scores and the highest reading scores is about 100 points. This is one standard deviation, or an effect size of 1.00.

b. If we use an SES score of -1.0 as an informal 'poverty line', the school profile shows that almost two-thirds (65%) of the schools have an average SES score that is below the poverty line. The majority of these schools are small, rural schools. A more detailed measure of SES, as well a measure of poverty, have been developed for PISA for Development (Tramonte and Willms, 2018).

c. The average SES of most private schools is substantially higher than that of urban public schools; however, their average reading scores are comparable to urban public schools that have an average SES above -0.5.

Figure 13. School profile for low reading proficiency (Level 1 and lower) in Mexico



Source: PISA, 2015.

Figure 13 shows the school profile for low reading proficiency for Mexico. For this analysis, low reading proficiency is defined as having a PISA score at Level 1 or lower. This type of profile, which shows 'percent vulnerable', is useful for dichotomous outcomes. The outcomes can be negative, such as low achievement or poor mental health, or positive, such as school completion. This profile goes hand-in-hand with vulnerability concentration plots, which are discussed below.

Inclusive school systems. Inclusive school systems have schools in which all children can thrive. 'All' means all learners across the categorical boundaries of gender, disability, social class, ethnicity, national origin, sexual orientation and religion (Willms, 2009a). 'Thriving' means developing one's best potential in academic, social, emotional, physical and spiritual well-being. Inclusion embraces a set of values and beliefs that all children have a legal right to be educated in a safe environment with their neighbourhood peers. It involves

"a common vision which covers all children of the appropriate age range and a conviction that it is the responsibility of the regular system to educate all children" (UNESCO, 2005, p. 13). Achieving inclusion requires making accommodations for students with special needs or behavioural problems, but the over-arching principle is that students have a right to be educated in the least restrictive environment. Developing an inclusive school system requires parents, educators, politicians and other community leaders to take a stance on inclusion by countering discriminatory attitudes and supporting initiatives that increase the participation of learners with diverse needs (Riehl, 2000).

Inclusive school systems have better student outcomes and less inequality (Willms, 2010). When school systems are more inclusive, the Foundations for Success – quality instruction, learning time, material resources and family and community support – are more equally distributed among schools. Moreover, students themselves are a key resource. Students with low ability have a better chance of success when they are in mixed-ability classes with peers that have high expectations and are intellectually engaged.

Horizontal inclusion. At the system level, one can consider two aspects of inclusion: horizontal and vertical inclusion. In school systems that are horizontally inclusive, the variation in SES intakes is relatively small. The range in school mean SES in Mexico, based on the school profile, is from about -3.0 to 1.2. Residential segregation is the main factor contributing to 'horizontal segregation'. In Mexico, for example, there are marked socioeconomic differences among neighbourhoods in the large cities as well as between urban and rural areas.

Vertical inclusion. In a vertically inclusive school system, the variation in student outcomes for schools with similar SES intakes is small. The school profile for Mexico indicates that the range of reading proficiency scores for the majority of schools is from 350 to 550. In school systems that are 'vertically segregated', that is, not vertically inclusive, students are allocated into certain types of schools or school programmes based on their prior academic achievement or on some measure of cognitive or language ability. Students can also be selected into particular classes or programmes within schools, such as classes for students with special needs. Grade repetition, language immersion programmes and curricular tracking can also contribute to vertical segregation.

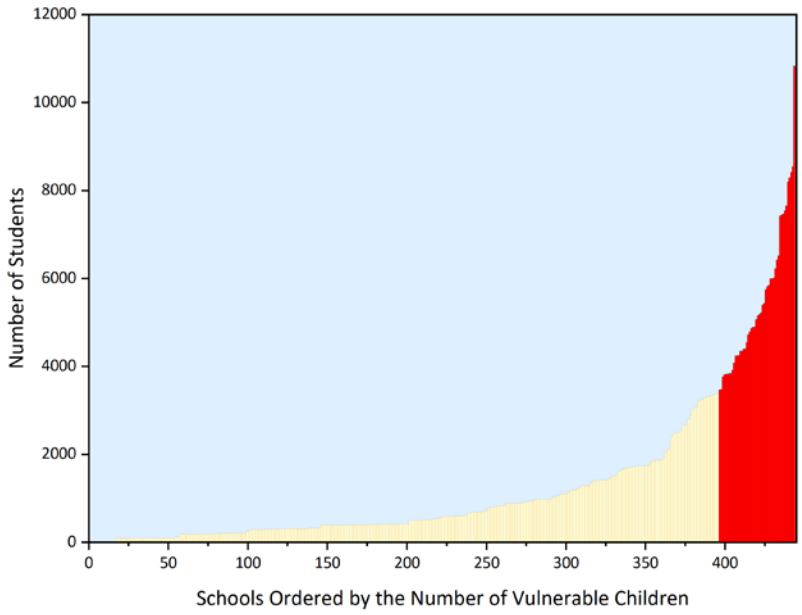
Private schools can contribute to both kinds of segregation. If they select students based on prior educational performance, they contribute to vertical segregation. If they have high school fees, which make them prohibitively expensive for low SES families, they contribute to horizontal segregation.

Two statistics are used to summarise the level of vertical and horizontal inclusion. The variation in student performance is partitioned into a within-school component and a between-school component. Vertical inclusion is the proportion of variation in a student outcome that is *within* schools. Similarly, the variation in students' SES is partitioned into within-school and between-school components. Horizontal inclusion is the proportion of variation in SES that is *within* schools. Theoretically, these indices can range from 0 to 100%, but most countries have levels in the range of 10% to 50% (Willms, 2010). The statistical techniques for the estimation of these indices are described in Appendix 2.

Note that the measure of vertical inclusion depends on the outcome being considered. Average levels of student well-being tend to be quite similar across schools and thus estimates of vertical inclusion based on a measure of self-esteem, for example, would have a much larger vertical inclusion index than one based on reading performance. Given the importance of reading skills, a vertical inclusion index based on a national or international reading test would be useful for monitoring vertical inclusion. The level of vertical inclusion can also vary with the level of schooling, especially in school systems that select students on the basis of achievement or ability at certain ages.

Vulnerability concentration. A vulnerability concentration plot is a rendition of a Lorenz plot, which is used in economics to represent the distribution of income or wealth (Lorenz, 1905). Vulnerability concentration plots show the concentration of vulnerable students among schools in the country. **Figure 14** provides an example. In Mexico, the PISA 2015 data indicate that there are 44,335 schools with 15-year old students in Grade 7 or higher. The population size is 1,392,995 students. Of these students, 41.8% have reading skills at Level 1 or below Level 1. Figure 14, which is a histogram, has 443 bars, with each bar representing 100 schools. The height of the bars indicates the estimated number of vulnerable students in each set of schools, with the schools ordered by the number of vulnerable children in a set of 100 schools. For example, the height of the 400th bar indicates that there are 3,826 students in that set of 100 schools, or on average about 38 students per school with Level 1 or lower reading proficiency. The group with the highest number has 10,838 students, or about 108 students per school.

Figure 14. The concentration of students with low reading proficiency in Mexican schools



Source: PISA, 2015.

This vulnerability concentration plot reveals that 50% of the students with low reading skills are in about 11% of the schools, which is less that 5,000 schools. 75% of the vulnerable students are in 18% of the schools, slightly less than 8,000 schools. In Section 4 of this report, one of the five types of strategies pertains to strategies that are outcome-targeted. In this case, a viable strategy would be to implement an intervention in 5,000 to 8,000 schools. This may seem daunting, but perhaps not if one considers that there are over 44,000 schools in Mexico serving 15-year old students.

The vulnerability concentration plots can be used also to assess the concentration of any sub-population, such as Indigenous students or students with a disability. The sub-population does not need to be 'vulnerable'. For example, in Mexico, about 4.5% of students had reading skills at Level 4 and higher. 50% of these students attended less than 2% of the country's schools.

Socioeconomic gradients within and between schools. The socioeconomic gradient shown in Figure 10 is the overall gradient for Mexico and these can be estimated for any jurisdiction, such as a country, region or school district. However, within each jurisdiction, the schools also have their own SES gradients, each with its own level, slope and strength. A multilevel gradient model describes the gradients for all schools simultaneously and the relationships among them. For Mexico, the average within-school gradient is 7.7, while the between-school gradient is 41.0. This suggests that within each school, students tend to have similar levels of reading proficiency. Inequalities in reading proficiency associated with SES are mainly associated with differences among schools.

School composition. A 'school composition effect' refers to the effect on students' outcomes associated with the aggregate characteristics of a school, such as the average SES of the school, the percentage of students who have a disability, or the percentage of students whose home language differs from the language of instruction (Alexander and Eckland, 1975; Bryk and Driscoll, 1988; Willms, 1986; Willms, 2010). The presence of a composition effect indicates that in addition to the effects associated with SES, there is an additional effect associated with school composition. If the composition effect for the mean SES of the school is positive, it indicates that students attending high SES schools tend to have better outcomes than those attending low SES schools, even after taking account of students' SES at the individual level. The composition effect associated with school mean SES for Mexico is 35.2.

The composition effect should be considered a proxy for a number of factors that are correlated with school composition, because with cross-sectional data one cannot separate the effects associated with school composition from effects attributable to foundation factors or other schooling processes (Raudenbush and Willms, 1995). To some extent, composition effects are attributable to peer effects: when bright and motivated students work together, they learn from each other and set higher standards for performance (Robertson and Symons, 1996; Zimmer and Toma, 1997). Also, schools with high SES intakes tend to attract and retain talented and motivated teachers, have more instructional resources and receive greater support from parents (Willms, 1986; Willms and Somers, 2001; OECD, 2001). In other words, they have stronger

Foundations for Success. The term 'effects' connotes a causal relationship, but one cannot infer a causal relationship from cross-sectional studies such as PISA (Alexander, Fennessey, McDill and D'Amico, 1979).

Intake variability. A number of educational policies and practices are based on the argument that when students are taught in schools or classrooms with more homogeneous intakes they learn at a faster pace. This argument underlies policies supporting the tracking of students into academically- and vocationally-oriented schools or school programmes, as well as the practice of having students repeat a grade when their academic performance is significantly below grade level.

Like the school composition effect, the effect of intake composition is estimated by including a measure of intake variability, such as the standard deviation of the SES of the school, as a school-level variable in a multilevel model. For Mexico, the effect of intake variability is 5.9. This is a small effect, which is not statistically significant.

Converging gradients. Earlier research based on the 2001 PIRLS data and the 2000 PISA data found that students' reading skills for high SES students did not vary substantially among countries compared with the variation for low SES students (Willms, 2006). In other words, gradients tended to converge at higher levels of SES. This pattern was also evident among youth aged 16-25 that participated in the 1997 International Adult Literacy Study (Willms, 2003b). For individual countries, an important question is whether the gradients for schools converge at higher levels of SES. If they do, it suggests that students from high SES backgrounds tend to fare well in their skills in most schools, whereas those from low SES backgrounds vary considerably in their skills, depending on the school they attend.

The statistic for estimating the extent of the convergence is the correlation between the levels of the gradients and their slopes. This statistic can be estimated with a two-level multilevel regression model that includes a measure of students' family SES at the student level and allows both the intercepts and the SES slopes to vary among schools. For Mexico, the correlation is 0.37, indicating that the slopes diverge with increasing levels of SES.

The estimation of the statistics based on a multilevel gradient model is described in Appendix 2.

IV. Policies about strategies and their execution

Monitoring data can inform policy questions about the performance of the school system: How are we doing? Who is vulnerable? and Where are the vulnerable children? The previous section identified several key statistics and graphs that can be used to summarise the performance of a school system. They provide a reliable and consistent approach for assessing levels of student performance and the equality of performance among sub-populations of students. Monitoring data can also be used to set realistic, measurable goals for improving student outcomes and reducing inequalities.

Educational policy entails setting goals and developing a course of action for achieving them. The 'course of action' requires the identification of a small set of strategies for achieving the outcomes and a plan for their execution. It involves setting priorities, identifying short- and long-term targets aligned with the goals and monitoring progress towards achieving these targets. It also requires policies about how best to allocate available resources. Monitoring data are at the heart of developing a set of strategies and making plans for their execution.

This section considers three ways to frame questions about strategies and their execution. First, what types of strategies are most likely to lead to system improvement? Second, how can resources be allocated to strengthen the Foundations for Success and develop a more equitable school system? Third, how can monitoring data be used to assess the effects of policies that change one or more of the key structural features of the school system?

Five types of strategies. Willms (2006) described five types of strategies that can be implemented by a country, province or state, jurisdiction or school.[3] The relationships between student outcomes and SES, which are depicted with the socioeconomic gradients and school profiles, can help one discern which type of strategy or combination of strategies is most likely to raise and level the learning bar. The five types are: universal, performance-targeted, SES-targeted, compensatory and reallocation. These are discussed below using the PISA 2015 data for Mexico as examples. For each strategy, the potential effect of a hypothetical strategy is considered. For the graphs displaying the hypothetical effects, the red gradient line displays the 'before intervention' status, which is set to the gradient observed with the PISA 2015 data. The green line displays the hypothetical 'after intervention' gradient.

Universal. A universal strategy strives to improve the outcomes of all students in a jurisdiction. Curriculum reforms, reducing class size, changing the age-of-entry into kindergarten, or increasing the time spent on reading instruction are all universal strategies as they are targeted towards all students, irrespective of their SES.

[3] In the earlier version of Learning Divides (Willms, 2006), the term 'interventions' was used, but this term is often used to connote specific strategies, such as introducing a new reading programme or pedagogical technique.

Figure 15 shows the effect of a universal strategy which has a uniform effect for students at all levels of SES. The effect size of the hypothetical strategy is 0.5, equivalent to 50 points on the PISA scale. Some universal strategies have stronger effects for low SES than high SES students, or vice-versa.

Performance targeted. A strategy that is targeted towards students with low levels of performance on an outcome is a performance-targeted intervention. For example, the Early Years Evaluation is used in several countries to identify the developmental skills of children aged 3-6 years as they prepare for and make the transition to formal schooling (The Learning Bar, 2011). The data collected are used to classify students into three groups, based on their scores in five domains. The classification provides teachers with information regarding the *type and amount* of support required for each child.

A performance-targeted strategy can also be implemented at the school level. For example, a reading programme may be administered in a sample of schools that has low average performance. In school systems with a low vertical inclusion index it is efficient to implement a whole-school strategy in a small number of schools. A vulnerability concentration plot can be used to estimate the number of children that would be reached with an intervention in a particular number of schools.

Figure 16 shows the effect of a performance-targeted strategy provided for all students who scored at Level 2 or lower. This hypothetical strategy has an effect size of 0.50. In this case, the strategy raises and levels the learning bar as there are disproportionately more students with low performance at lower levels of SES.

Risk-targeted. A risk-targeted strategy aims to provide additional support or resources for children deemed at risk of school failure, such as those with low levels of SES. The distinction between a risk-targeted and an outcome-targeted strategy is that risk-targeted strategies select and intervene with children who are deemed 'at risk' rather than those who have already been identified as having a poor developmental outcome. A Head-Start preschool programme for children from low-income families is a good example of a risk-targeted intervention. This type of strategy can also be directed towards a sub-population that is considered vulnerable, such as new immigrants or students who are members of an ethnic minority.

Figure 17 shows an example of a risk-targeted strategy with an effect size of 0.50, for students with an SES score below -2.0. This strategy raises the bar for low SES students and therefore also levels it, reducing the gap between low and high SES students. One issue concerning risk-targeted strategies is that high SES students with low scores do not receive the intervention. This issue can be addressed if a performance-targeted strategy can be implemented in combination with the risk-targeted strategy. However, performance-targeted strategies are not always possible. For example, during the early stages of the life-course, one cannot easily measure children's developmental outcomes and therefore it is more practical to implement a strategy designed for children deemed at risk. The risk factor might be low SES families or children who are born prematurely or with a low birth weight. Similarly, for some types of outcomes, such as engaging in risky sexual behaviour or dropping out of school, performance-targeted strategies are not possible. Therefore, one must use available information on risk and protective factors to identify the target population.

Figure 15. Effects of a universal strategy with an effect size of 0.50

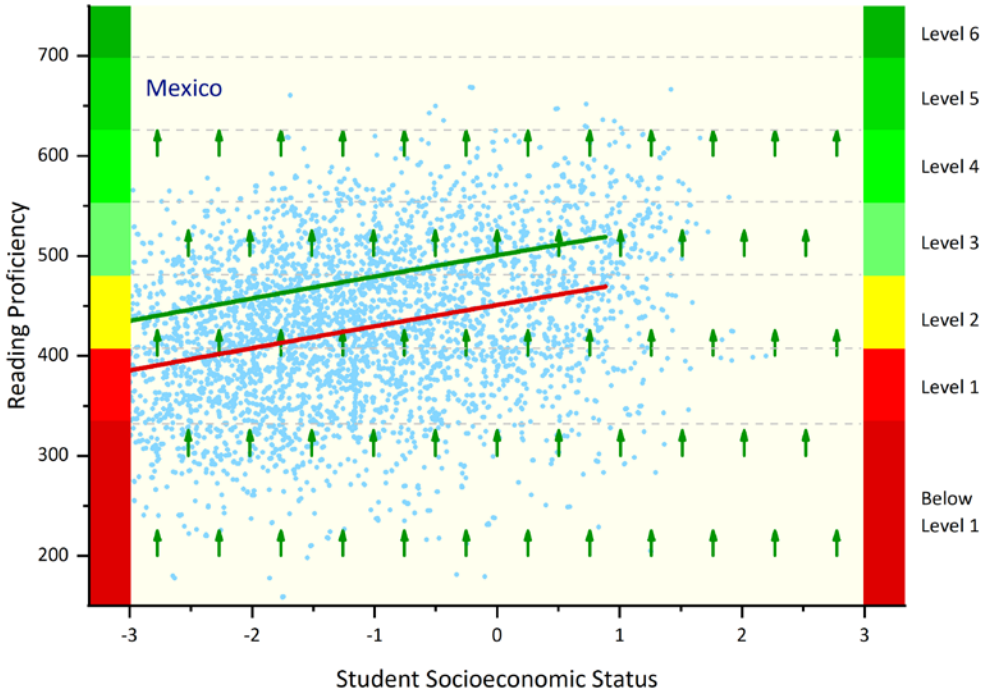


Figure 16. Effects of a performance-targeted strategy with an effect size of 0.50 for students with reading proficiency at Level 1 or lower

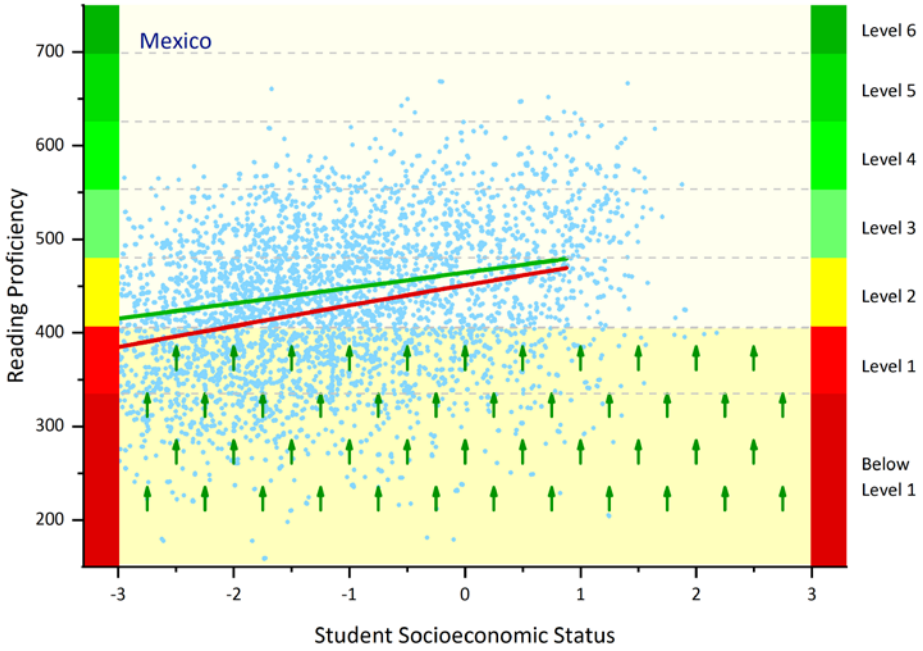
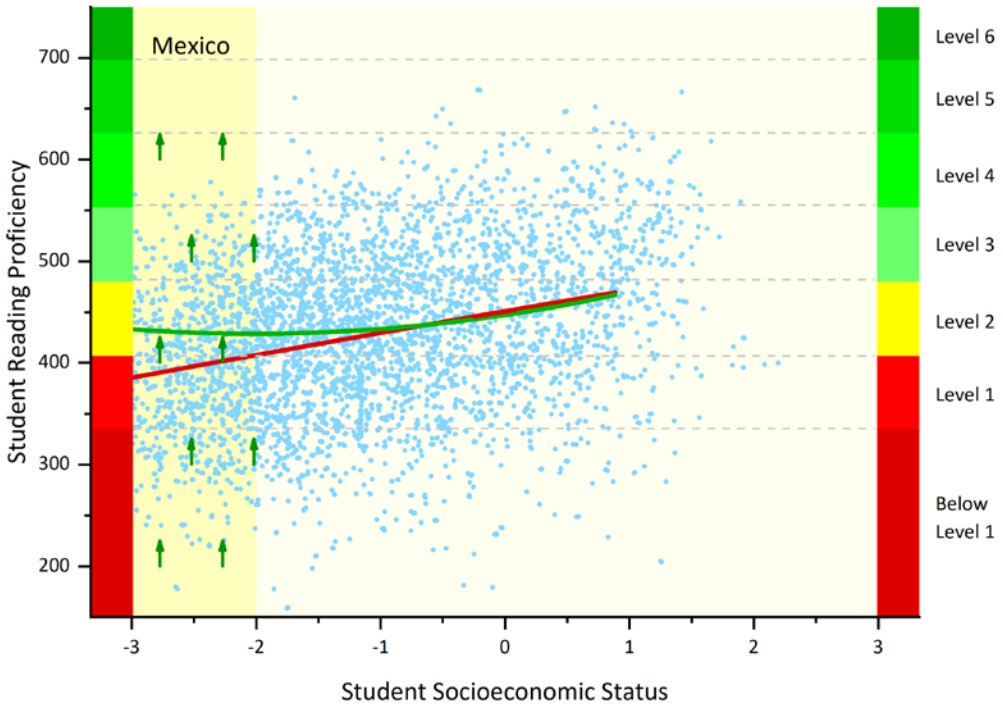


Figure 17. Effects of a risk-targeted strategy with an effect size of 0.50 for students with an SES at -2 or lower



Another concern raised about risk-targeted strategies is that students who are considered at risk but have high outcome scores also receive the intervention. One could argue that this is not a cost-effective strategy. However, one could rebut this argument by claiming that as those children further improve their skills, they serve as role models for other students with similar family backgrounds.

Compensatory. A compensatory strategy provides additional educational resources to students from low SES backgrounds or students deemed 'at risk' for other reasons. The term, 'at risk', can refer to being at risk for not successfully achieving a particular development outcome or more generally at risk of poor development for a range of developmental outcomes. The targeted sub-population can be the same as for an SES-targeted intervention; however, the difference is that a compensatory strategy strives to improve children's socioeconomic circumstances with the view that it will improve their educational outcomes. Providing free breakfast or lunch programmes or free textbooks for low SES students are compensatory strategies.

Figure 18 shows the effects of a compensatory strategy that raises SES scores by 0.25 standard deviations. The assumption in this example is that children receiving the strategy will improve their reading scores commensurate with the slope of the SES gradient. The slope of the gradient for Mexico was 21.2 points, suggesting that scores increase 21.2 points with each one standard deviation increase in SES. Thus, if a

student's SES increased by 0.25 standard deviations, we could expect his or her reading score to increase by about 5 or 6 points. A compensatory strategy can also be implemented at the school level. However, neither strategy has a marked effect in improving children's outcomes.

Reallocation. A reallocation strategy strives to include marginalised or disadvantaged children into mainstream schools. The argument is that when they are in mainstream schools they will benefit from the school composition effects discussed earlier. One strategy is to close small, low SES schools and reassign students to other schools. However, this strategy does not yield large effects on achievement and in some jurisdictions can have adverse effects on attainment. For example, children are more likely to remain in school longer if schools provide the full range of schooling from pre-primary to upper secondary. Another strategy, which can be possible in large urban school jurisdictions, is to redraw catchment areas with the goal of reducing SES segregation.

The school profile for Mexico, shown in Figure12, shows that a number of schools have an average SES below -2.0. **Figure 19** shows the effects of a reallocation strategy in which students attending schools with an SES below -2.0 were placed in higher SES schools. However, the majority of the schools in Mexico with very low SES are small, rural schools and thus neither of the reallocation strategies described above is likely to be viable. Moreover, the effect on raising and levelling the learning bar is minimal.

Figure 18. Effects of a compensatory strategy that increases the SES of low SES students by 0.25 standard deviations

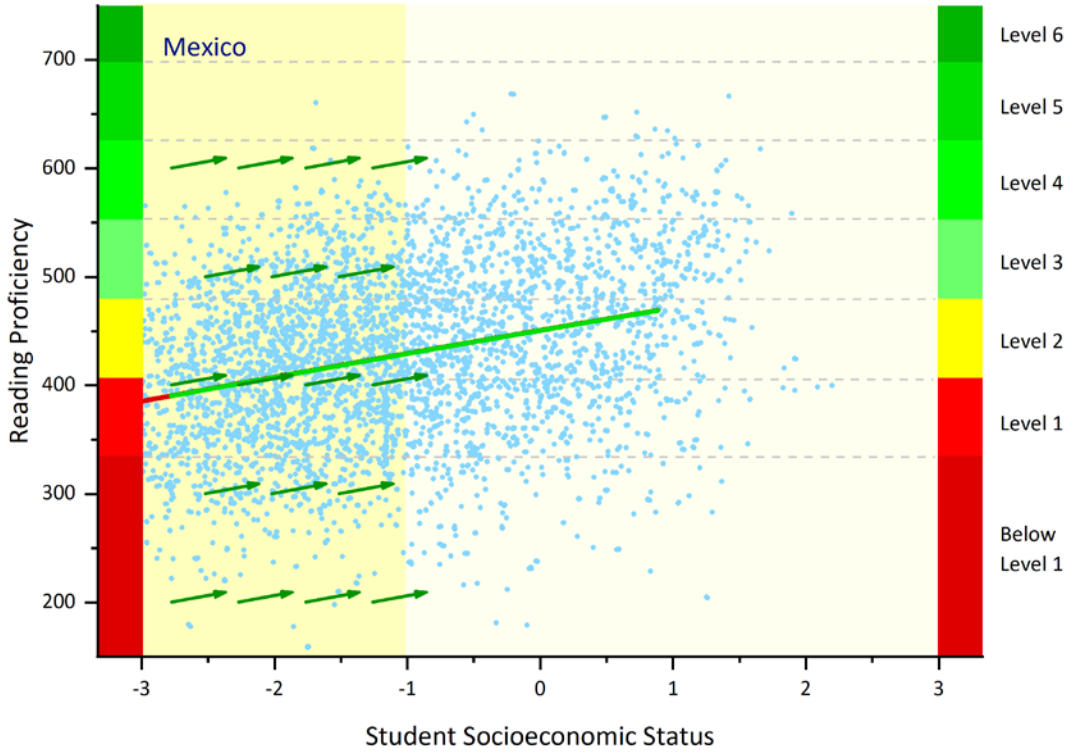
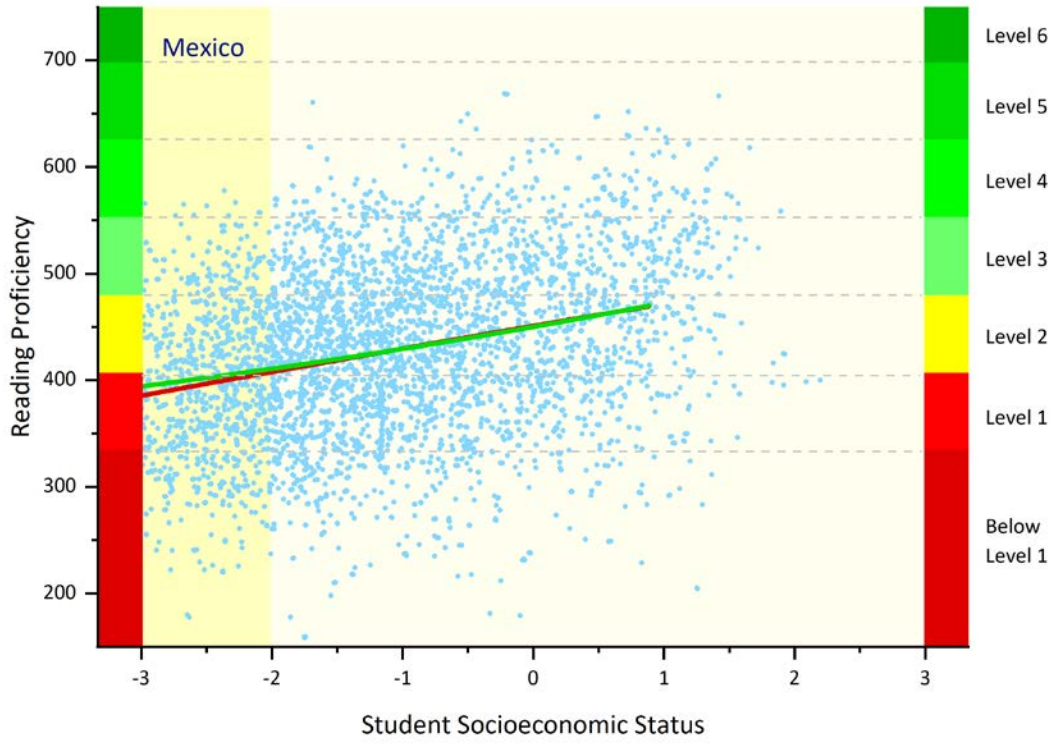


Figure 19. Effects of a reallocation strategy that reassigns students from low SES schools into mainstream schools



Strengthening the Foundations for Success

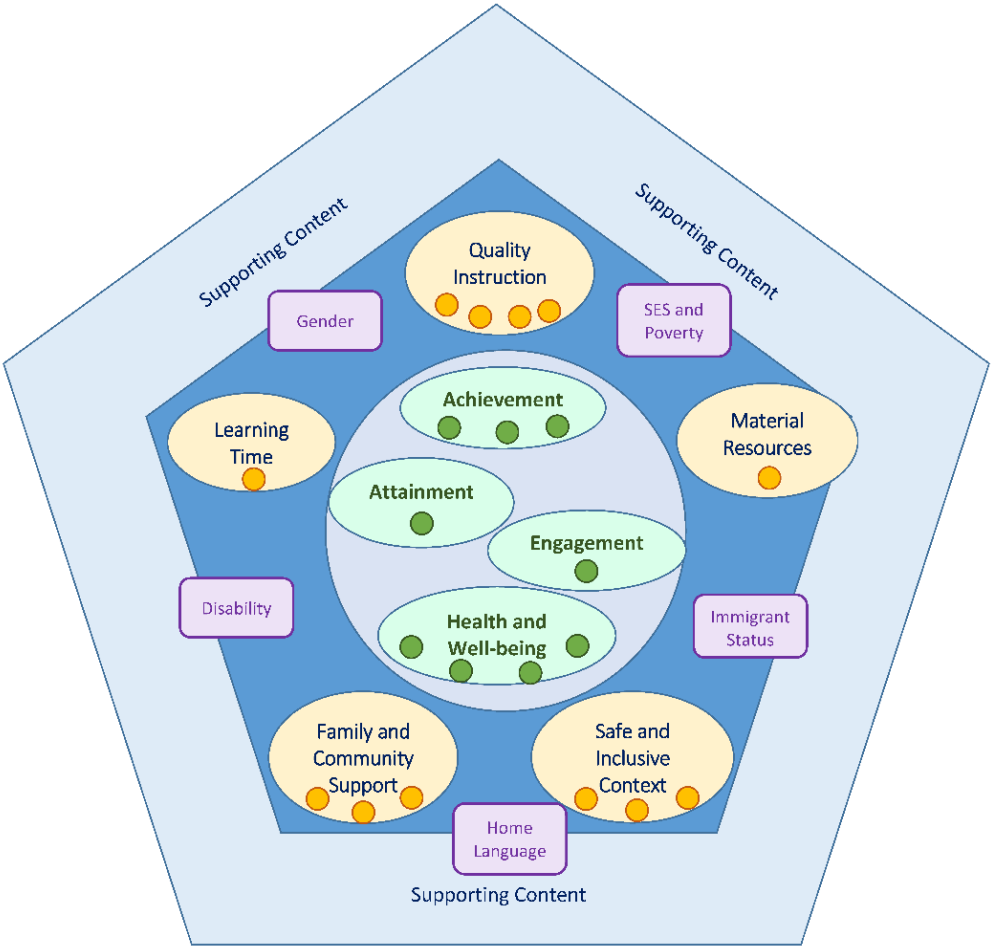
The Foundations for Success in the Educational Prosperity model comprise a set of family, institutional and community factors for each of the six stages of development. During the last three stages, when children are in formal schooling, the key school foundations include: safe and inclusive schools, quality instruction, learning time and material resources. When educators interact with and influence families and community leaders, they contribute to the social capital that is brought to bear on children's development. Thus, for this report, we include a fifth foundation, family and community support, as a school factor.

The Educational Prosperity model adopted for PISA for Development includes these five foundations (OECD, 2017). **Figure 20** shows the Prosperity Outcomes, the Foundations for Success and the equality-equity sub-populations. It summarises the set of measures included in the student, teacher and school questionnaires. The four Prosperity Outcomes are at the centre of the figure, shown with green ovals. They include one or more measures for each outcome. For example, achievement has three measures: reading, mathematics and science proficiency. The Foundations for Success, shown with light orange ovals, include one or more measures for each foundation. The sub-populations of interest to the participating countries are shown with the purple rectangles. The contextual questionnaires also include a number of other factors that are relevant

for the participating countries. These are referred to as supporting content. Altogether, the questionnaires for the main PISA-D survey include 49 questions in the student questionnaire, 33 in the teacher questionnaire and 28 in the school questionnaire (OECD, 2017).

The key idea in setting out the framework in this way is to focus attention on a small number of critical metrics for each stage of development and achieve alignment across all stakeholders of the school system. This could be the 'Minister's Dashboard'. The measures underlying each of the foundations provide information on the strengths and weaknesses of the school system. These metrics can then be used to set short- and long-term goals and develop an action plan to achieve these goals. It establishes the base for a communication plan; this is the direction of education policy, writ large, which is communicated to actors at all levels of the school system. The aim is to align goals with strategies at the Ministerial level, the jurisdiction level and the school level.

Figure 20. Educational Prosperity Framework for PISA for Development



An important difference between this approach and the traditional input-process-output 'logic' approach that underlies the tradition of the school effects research is that it does not presume to capture all of the relevant factors that contribute to student success. Clearly, 'principal leadership' is an important factor contributing to student success, but its effect is mediated through the core foundations. If a jurisdiction were to launch a training programme on principal leadership, its effects would be realised through principals' ability to effect change in quality classroom instruction, ensuring the school is safe and inclusive and so on. With this lens, the design of a principal leadership programme would focus on how the foundations are defined and measured, the key indicators associated with each of them and the strategies for building strong foundations. Also, the approach helps one identify strategies that are unlikely to effect change. Quite often, administrators have 'pet projects', which in many cases are good things to do, but they do not really support the foundations that lead to Prosperity Outcomes. The aim is for everyone to maintain a relentless focus on building the Foundations for Success.

Another important difference is that the 'effects' of the foundation factors are based on a broad literature that supports causal effects on the outcomes, rather than on estimates of their relationships to the outcomes, based on a recent cross-sectional study or on any single study. The foundations are stand-alone constructs with reliable and valid measures that represent them. These measures are metrics that can be used to set short- and long-term goals and develop an action plan to achieve these goals. An extreme example can help illustrate this point. Smoking is bad for one's health. We might have an indicator for health and well-being indicating the prevalence of youth who are non-smokers. We are not concerned with estimating the effects of smoking on the measures of health and well-being measured with the monitoring system, nor are we concerned with whether the 'relative risk' of smoking varies among countries. Reducing the prevalence of youth smoking is simply the right thing to do.

The Educational Prosperity approach is consistent with the principles of 'outcome mapping' (Earl, Carden and Smutylo, 2001). In particular, "it is not based on a cause-effect framework; rather, it recognises that multiple, nonlinear events lead to change. It does not attempt to attribute outcomes to any single intervention or series of interventions. Instead, it looks at the logical links between interventions and behavioural change" (Earl et al., p. 12). As a strategic planning tool, one can develop a chain of strategies that emanate from each foundation. One could ask, for example, "What is the set of conditions that are *necessary* and *sufficient* to improve quality of instruction". One could make a case that one necessary condition is that teachers need a deep understanding of the "simple view of reading" and develop a repertoire of teaching strategies associated with teaching children how to read. This condition could then be linked with one or more specific projects. One cannot attribute a causal connection for any single project or intervention, as there are always other factors that qualify. Rather, the aim is to monitor 'quality instruction' with a small set of key indicators.

In developing the measures for the Foundations for Success, there are advantages to reporting them at the school level. Although many of the drivers of schooling outcomes vary within and between schools (Rowan, Raudenbush and Kang, 1991), setting goals to improve the Foundations for Success is clearer and more

straightforward when one has a small number of measures. Reporting each of the measures on a ten-point scale is also beneficial, with the scoring rules made explicit. The acid test is whether each school principal knows how the foundations are measured and what it takes, say, to move his or her school from a four to a five, or from a seven to an eight, for each of the foundations.

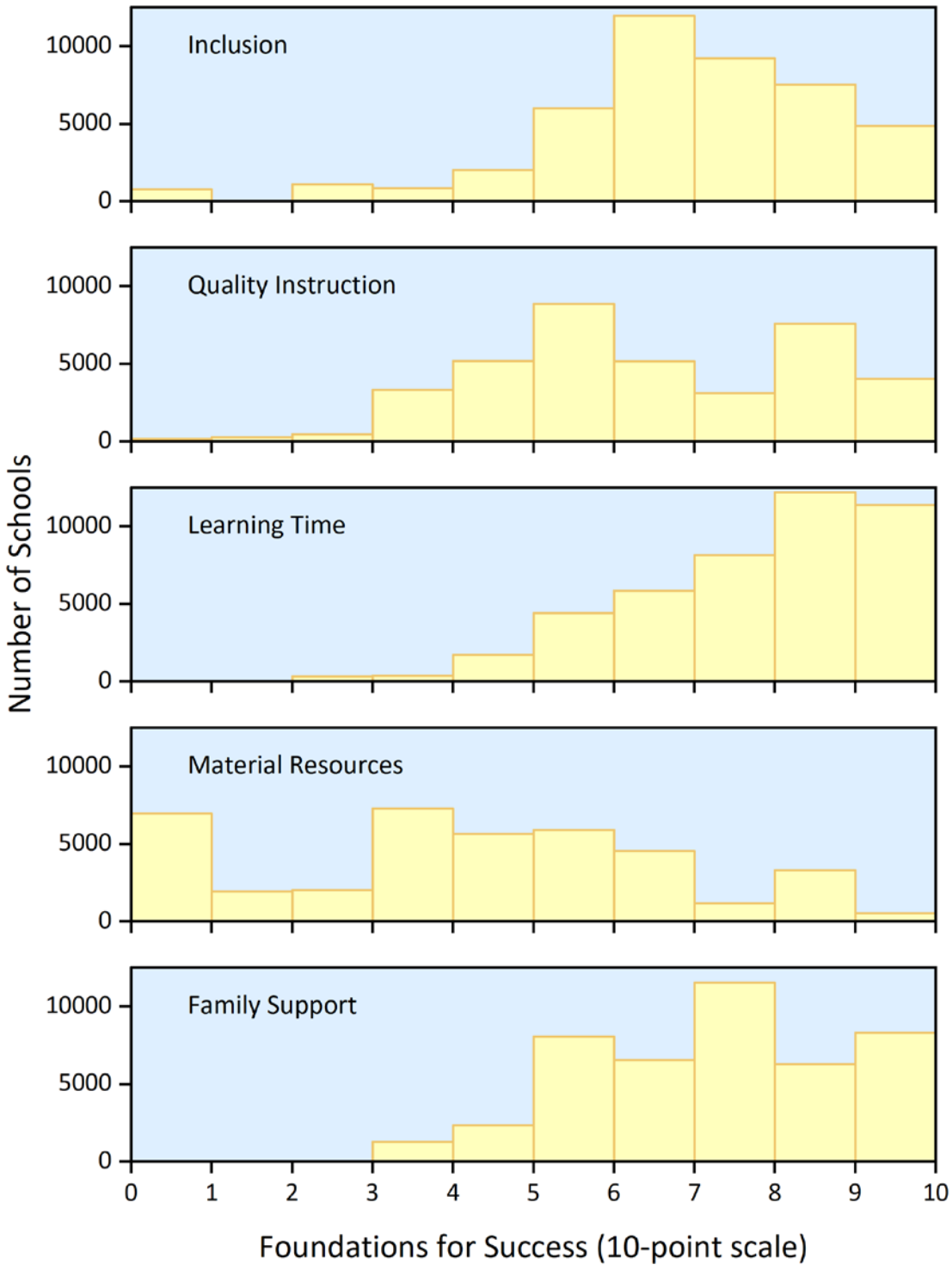
Figure 21 shows a school resource plot, again using PISA 2015 data for Mexico as an example. A school resource plot is simply a set of histograms summarising the key indicators for the Foundations for Success on a single page. This figure is for illustrative purposes, as PISA 2015 includes only a small number of measures that map onto the foundation factors. For this example, each school was assigned a score for inclusion derived from the student-level measure of 'sense of belonging at school'. Each point on the scale indicates that 10% of the students in the school have a positive sense of belonging. For example, a score of 6 indicates that a school has between 50% and 60% of its students with a high sense of belonging. A similar approach was used to determine the percentage of students in each school who were considered to have strong family support and for learning time, with positive class and school attendance. The measures of quality instruction and material resources were derived from the school-level variables of staff resources and material resources in the PISA 2015 school-level data.

Analyses pertaining to the equity of provision can be conducted using the techniques described earlier in this report. In particular, for measures at the student level, the use of socioeconomic gradients, school profiles and vulnerability concentration plots provide useful information for policies about the allocation of resources. For measures at the school level, school profiles are informative. For certain sub-populations, such as those defining ethnicity or immigrant status, one can construct school-level variables denoting low, middle and high percentages of the sub-population of interest and then develop school profiles similar to the example presented in Figures 12 and 13.

In this example, the school resource plot for Mexico indicates that there is considerable variation among schools in the levels of school resources and quality instruction, whereas the measures of learning time and family support have less variation and higher scores.[4] Clearly, the data from the main PISA study do not provide a comprehensive assessment of the Foundations for Success. Moreover, they provide information for only a sample of schools and on a three-year cycle.

[4] The PISA measure for material resources is based on four items for a question asked of school administrators, "Is your school's capacity to provide instruction hindered by any of the following issues?", with response categories of "not at all", "very little", "to some extent" and "a lot". PISA-D uses a more extensive measure based on a schema set out by Murillo and Román (2011), which asks teachers questions about the availability and condition of didactic resources and their use of these resources and questions of teachers about the availability and condition of basic services and didactic resources (OECD, 2017; Tramonte and Willms, 2018).

Figure 21. School resource plots for Mexico



Source: PISA, 2015.

Altering the structural features of schools

In most societies, schools and school systems have long traditions about how schooling is organised, what is taught and how instruction is delivered. These are 'structural features' of schools. They include aspects of school organisations at the system level, such as when students start school, the grade levels at each stage of schooling, school size, the formal curriculum, the language of instruction, assessment strategies, professional development and the ways students are selected into particular schools and school programmes. They include procedures within the school, such as how classes are organised, the length of the school day, the length of class periods, teaching arrangements and how students are grouped for instruction. School reforms typically entail altering one or more of the long-standing structural features of schools.

When governments implement a reform to restructure schools, they seldom have the luxury of conducting a *true* experiment, with a well-defined intervention and with schools or students randomly assigned to treatment conditions and a control group. Rather, reforms tend to transpire slowly with a gradual shift from more bureaucratic approaches, with specialised and differentiated work roles and top-down, formal lines of authority, to communal forms, with varied work roles and shared responsibility for a common set of goals (Lee and Smith, 1993; 1995). One can easily classify schools as traditional versus progressive and compare academic and social outcomes; however, one cannot easily attribute causation to any particular reform element, such as mixed-ability classes, flexible class schedules, or providing more time for teachers to plan lessons. Selection effects and the variation in implementation fidelity are too difficult to control. Moreover, as restructuring is proceeding apace, critical events or changes in direction can stymie or hinder reform efforts, rendering it impossible to evaluate their effects. Changes in government alongside interference from teachers' unions are common and often there are pressures from parents and community leaders to revert to the *status quo* (Willms, 2008).

A strong monitoring system with well-defined, reliable measures of the Prosperity Outcomes and the Foundations for Success is essential to assess the impact of structural innovations. Positive reforms improve outcomes and reduce inequalities. They strengthen foundations. Monitoring data can be leveraged to conduct strong quasi-experiments for assessing effects associated with altering structural features of schools.

Imagine a school jurisdiction that was convinced ungraded classes would be a better structure for primary schooling, especially for schools with a high percentage of vulnerable students. The jurisdiction has a strong monitoring system in place which provides data it can use for the design and evaluation of an intervention. Administrators decide to implement an intervention to assess the effects of ungraded classes in a small number of schools. The monitoring data can be leveraged in at least five ways.

First, a jurisdiction with a well-developed monitoring system can use its data as a baseline for conducting an intervention study. For example, a reading intervention such as Confident Learners could be implemented in a small sample of schools for which there were baseline data. The study could be a quasi-experiment in

which the intervention is implemented in one region of a country but not another, or with a randomised experiment with treatment and control groups. The core principle is that monitoring data constitutes an input-process-output 'shell' which provides baseline measures of the Prosperity Outcomes and the Foundations for Success. These measures provide the means to track changes over the course of the reform. An advantage not to be over-looked is that the start-up costs of the intervention are minimised as the jurisdiction does not need to spend one or two years developing outcome measures or identifying and defining sub-populations. Its energy can be spent on addressing the challenges inherent in implementing the reform.

Second, the monitoring data can be used for identifying strategic samples of schools. School profiles and vulnerability concentration plots can be used to identify the target schools. School mean SES or the percentage of vulnerable students can be used to define a stratum for stratified random sampling (S*ä*rndal, Swensson and Wretman, 1992). For example, a sample of potential participating schools is divided into strata based on their profiles and then random samples are collected within each stratum.

Third, if it is feasible to conduct a relatively formal assessment, then the school profiles can be used in a randomised block design. For example, sets of schools with similar profiles are randomly assigned to treatment conditions associated with the intervention (Matts and Lachin, 1988).

Fourth, the monitoring data can be used after data collection to strengthen the internal validity of the evaluation. Demographic measures can be used as statistical controls when estimating intervention effects. They can also be used to assess treatment-by-subject interactions: for example, do the effects of ungraded programmes compared with graded programmes have stronger effects for low SES students than for high SES students?

Fifth, the monitoring system provides a reliable strategy for assessing changes in the equality of outcomes for key sub-populations.

Finally, one can ask whether the intervention has any desirable or undesirable side effects. While the main aim is to improve students' reading skills, does it also result in students being more engaged? How does it affect attendance? Are teachers more or less likely to use high-yield teaching strategies?

V. Monitoring for Educational Prosperity

The Educational Prosperity framework is a powerful tool that can be used by Ministries of Education to monitor a core set of student outcomes and the drivers of those outcomes. The framework has proven to be useful in low-, middle- and high-income countries (Willms, 2018b). It has implications for tracking progress towards achieving the 2030 Sustainable Development Goals (SDGs). This section of the report discusses strategies for setting goals, provides an example of a comprehensive monitoring system and discusses new directions for educational assessment arising from Educational Prosperity.

Setting goals

A framework for setting goals in the fields of management and education uses the acronym, SMART, which is generally attributed to Doran (1981). His definition included five criteria: goals should be specific, measurable, assignable, realistic and time-related. A revised version commonly used in education replaces 'assignable' with 'attainable' and 'realistic' with 'relevant'. The revised version is used in this discussion.

Specific. A specific goal is one that clearly states what is to be accomplished. Goals are more likely to be achieved if there is a small number of well-defined goals. For monitoring purposes, specific goals can be written for each of the Prosperity Outcomes. They can include targets for both improving school performance and reducing inequalities.

Consider Goal 4.1 for the 2030 agenda for sustainable development: "By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes." This goal clearly requires greater specificity, which can be achieved by breaking it into a number of separate goals that define effective learning outcomes. Also, the Educational Prosperity framework distinguishes between outcomes and the drivers of those outcomes. Thus, free, equitable and quality education are embodied in the Foundations for Success.

Measurable. Goals need to be stated in absolute terms and linked to time-invariant scales. For example, a country might set goals for reducing the prevalence of children who are born with low birth weight (LBW). LBW has well-defined criteria: infants who are LBW are those born with a weight of less than 2,500 grams, regardless of gestational age (World Health Organisation, 2010). For most educational variables, the outcomes are defined on standardised scales or as dichotomous variables. An 'effect size' is a useful metric if the scale is time-invariant; for example, the mean and standard deviation are fixed for a baseline year and not rescaled to have a new standard deviation each year or for each assessment cycle. The NAPLAN scale is exemplary in that it covers a wide age range and is not rescaled each year. Measures based on a set criterion are often preferable, such as the percentage of students achieving the Intermediate International Benchmark in PIRLS or Level 2 or higher in PISA. The measures of anxiety and depression used with The Learning Bar's OurSCHOOL surveys (The Learning Bar Inc., 2009), which are being used in PISA for Development, have fixed, time-invariant criteria for moderate and severe levels of anxiety and depression.

Goals stated in terms of a rank order of countries or other jurisdictions are useless because even a small change in an outcome can have a dramatic effect on a country's rank. Also, a country's rank can change, not because its results improved, but because the results for other countries declined.

Attainable. Setting attainable goals is perhaps the most challenging aspect of goal-setting. They must be realistically achievable in a specified period and yet challenging for educators at all levels of the school system. As was noted earlier, educational measures derive their meaning by comparison to some standard, through comparisons among jurisdictions and by an examination of trends over time. The value of international assessments such as PIRLS, PISA and TIMSS in goal-setting is that they provide a broader context to discern how well the students in a country are doing compared with countries with comparable social and economic contexts. They also provide an indication of the magnitude of inequalities in performance, inequities in provision and the extent to which schools vary in their outcomes at varying levels of socioeconomic status.

Individual schools require student-level data on the prosperity outcomes for all students in the school, collected at regular intervals. The results from PISA can be used to gauge what might be possible, given the right combination of national and local support and the capacity and will of the teaching staff. For example, the school profiles for Mexico in Figures 12 and 13 provide an indication of the range of PISA scores and the range of the prevalence of students with low scores, for all schools at varying levels of SES. Visually, if one ignores some of the outliers, the profiles suggest that PISA scores vary among schools at each level of SES by about 80 to 100 points and the prevalence of vulnerable children varies by about 40% to 60%. The HLM models from which the profiles are derived provide estimates of the variation in results after controlling for the mean SES of the school. As a way to gauge whether a goal is attainable, a useful measure is the inter-quartile range. For Mexico, the inter-quartile range of PISA scores is 36 points and the range of the prevalence of vulnerable students is 20%. Based on these results, an attainable goal for a school could be to increase standardised scores in reading performance over a five-year period by about 9% of a standard deviation and reduce average levels of vulnerability by 5%. These targets are approximately one-quarter of the inter-quartile ranges.

Tracking progress towards such goals requires the collection of data on individual students at regular intervals, covering the full range of grades in the school, ideally with measures that can monitor each child's growth in the Prosperity Outcomes. The collection of data with background questionnaires that provide information on the Foundations for Success can enable schools to monitor their progress in providing an environment in which all students can succeed. Most schools cannot do this alone. Thus, the role of school districts and local and national governments is to provide the tools that enable schools to monitor their outcomes. In many contexts, this requires a shift from assessment being used to hold schools accountable to one in which data from tests and questionnaires are used as leading indicators to set goals and develop school improvement plans.

The Educational Prosperity model calls for educational administrators at the Ministry level to set goals based on a range of assessments that cover the lifespan from birth to late adolescence. It also calls for measures of the Foundations for Success for each stage of development. Improvements in the strength of the school system need to be gauged on a wide range of measures, with goals set accordingly. Monitoring at this level requires regular assessments, conducted annually or at least biennially. These assessments can be based on random samples of schools and students.

The international assessments are not useful for setting goals, as they are too infrequent and the results can fluctuate from cycle to cycle in ways that do not reflect real changes the structure of the school system or the provision of quality education. For example, the PISA reading scores in Mexico fell by 22 points from 2000 to 2003, from 422 to 400. Science scores declined from 422 to 405, but math scores fell by only two points, from 387 to 385. Vidal, Díaz and Jarquín (2004) suggested the decline in reading and science scores may have been partially attributable to an increase in the participation rate in PISA, from 52% in 2000 to 58% in 2003. But this would not explain the math results. Also, the average socioeconomic status of the PISA sample decreased only slightly, from -1.17 to -1.11. Another plausible explanation is that the changes are mainly attributable to sampling and measurement error. By 2009, the reading and science scores for Mexico had returned to their 2000 levels, while math scores increased to about the same level as the reading and science scores. These results emphasise the fact that international assessments are not useful for setting goals. If a country set out to increase its academic achievement by 9% of a standard deviation and reduce average levels of vulnerability by 5%, the fluctuations attributable to measurement and sampling error are too large to provide a reliable gauge of changes in system performance. However, international assessments can lend credibility to the results of national assessments and their validity can be strengthened with studies that link international assessments to national assessments (Singer and Braun, 2018).

Relevant. The Educational Prosperity framework identifies four or five Prosperity Outcomes for each of the six stages of development. Their relevance is not debated in this paper. Instead, the 'relevance' criterion is viewed in two ways. First, the goal must be understood and seen as important for actors at all levels of the school system, from the Minister and his or her staff, to the regional administrators and to teachers, parents and students. This alignment of goals across all stakeholders in the school system is critical for achieving success. Second, goals can be seen as a means of communication. Improvements in a school system requires strengthening the Foundations for Success and in most cases this is facilitated when there is increased funding. A public statement of a small number of goals can help develop a 'framework of understanding' of their importance for a wide range of stakeholders, including politicians, community leaders and donors. The 'relevance' of Prosperity Outcomes can be bolstered by appealing to their economic, health and social benefits. For example, improvements in literacy skills are related to earnings and tax revenue, reduced crime rates, less unemployment, less dependence on social welfare and lower health care costs (Hanushek and Woessmann, 2015; Ross and Wu, 1995).

Timeframe. The large-scale international studies such as PIRLS and PISA have a relatively long cycle. Moreover, the data do not become available until at least one year after data collection. These assessments

are useful for long-term planning; however, to monitor progress school systems also require monitoring data provided annually or even more frequently. The requirement is especially pertinent to monitoring literacy skills during the early primary school period.

An example monitoring programme

Figure 22 shows an example of a monitoring programme based on Educational Prosperity. It is a revised version of an assessment system designed for Indigenous populations in Canada (Willms, 2009b). The arrows represent assessments that are census-based, while the solid rectangles indicate sample-based assessments. The implementation of this design in a number of Canadian jurisdictions has used the Early Years Evaluation for the assessments at ages 4 and 5 (blue arrows), the Confident Learners assessment for Kindergarten to Grade 3 (orange arrows, with the exception of the numeracy assesment) and the OurSCHOOL surveys for several of the Prosperity Outcomes and the Foundations for Success (purple arrows). The data for sample-based measures (orange and green rectangles) are collected with provincial assessments. The tracking of educational attainment is done by the local school jurisdictions. This design provides sufficient data for monitoring the Prosperity Outcomes and the Foundations for Success as outlined in this report.

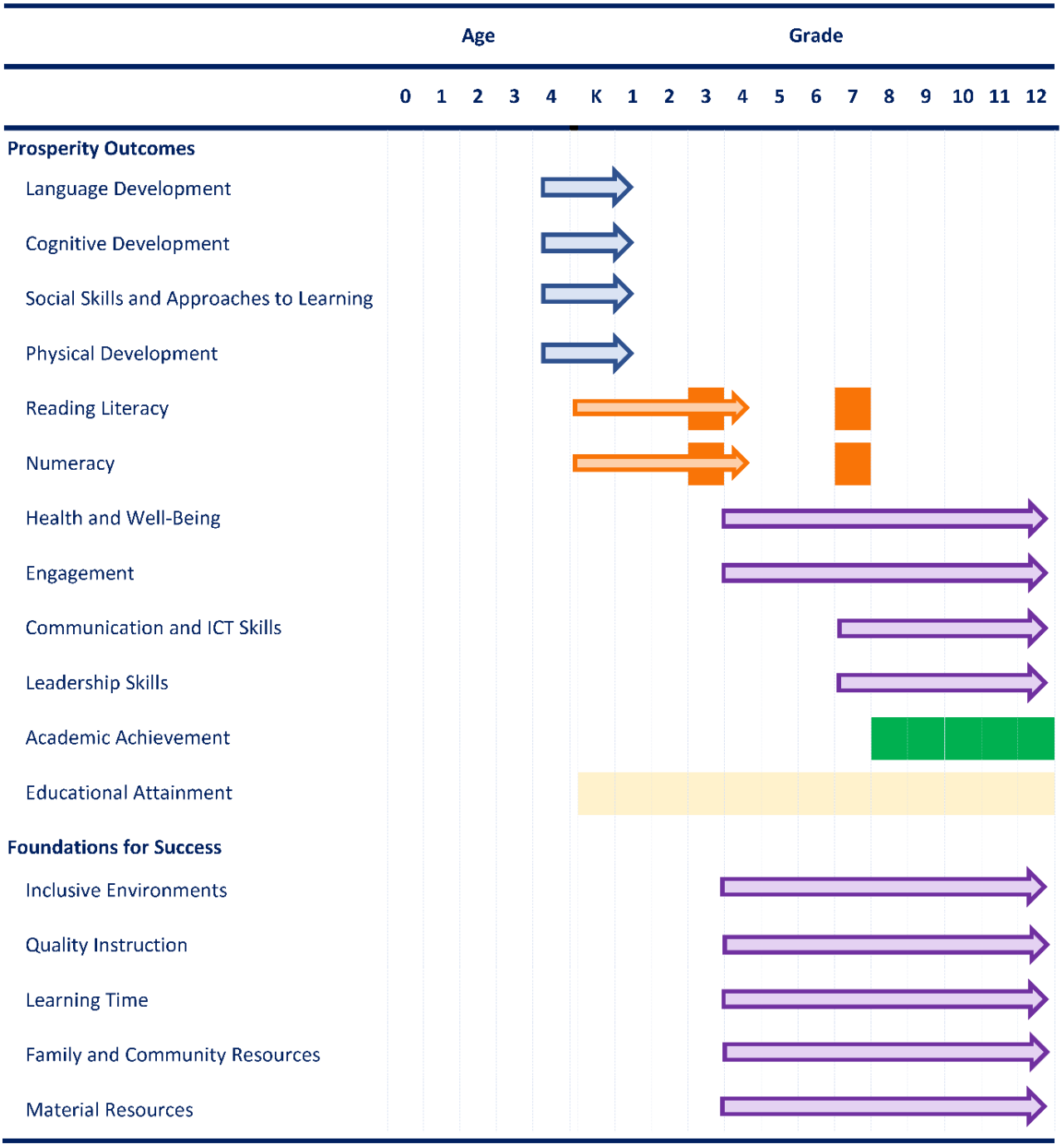
New directions

The Educational Prosperity framework calls for changes in the ways countries, jurisdictions and schools monitor their schooling systems. It also calls for new ways of using monitoring data to inform educational policy. Many countries administer student achievement tests at certain grades during the elementary and secondary school years. Alongside these efforts, some jurisdictions collect data from students, teachers, parents and school administrators on various 'contextual' aspects of the school system. In addition, many countries participate in large-scale international studies such as PIRLS, TIMSS and PISA.

The development of children's reading ability at the end of secondary school is the cumulative result of children's environments during the prenatal period and their learning experiences since birth. Findings of a national evaluation in Uruguay of children's early years skills in 2017 indicate that even before children begin formally learning how to read, their language and cognitive skills vary substantially. At least one-quarter of children at age 5, when they were set to enter the 1st grade, had cognitive and language skills that were one or more years behind those of their same-age peers.

Children who enter the 1st grade without the basic readiness skills for learning how to decode words are at increased risk of having low reading proficiency at the end of lower primary school. Moreover, most of the growth in reading proficiency occurs during the first few years of primary school. For example, findings from Australia's national assessment of reading skills, called NAPLAN, show that children vary substantially in their skills at the end of Year 3 and the rate of growth thereafter decreases year over year though to Year 9. After Year 9, the PISA results indicate that the annual rate of growth is only about 14% of a standard deviation.

Figure 22. Domains and measures for a national monitoring system



These two findings, that children vary substantially in their skills when they begin formal schooling and that learning to read during primary schooling is critical for students' long-term success, stress the importance of shifting the emphasis of educational monitoring to the early years. Therefore, Educational Prosperity emphasises collecting data on children's skills before or shortly after they enter school, at regular intervals with individual assessments during the first three years of school and at age 8 or 9, when students are expected to make the transition from learning-to-read to reading-to-learn. This shift in the emphasis of monitoring is required for low-, middle- and high-income countries alike. In their analysis of educational outcomes in South Africa, van der Berg, Spaull, Wills, Gustafsson and Kotzé (2016) state that "learning to read for meaning and pleasure in the Foundation Phase is the single most important goal for primary schooling" (p. 15).

Most monitoring systems have traditionally emphasised students' proficiency in reading, mathematics and science. Educational Prosperity calls for monitoring a wider range of outcomes which include educational attainment, engagement and health and well-being. For low- and middle-income countries and for students in low-income areas of high-income countries, the measures of attainment need to capture information on school attendance and whether students complete successive levels of schooling.

The emphasis for many school systems on monitoring contextual factors has been on estimating differences in school performance and using the results to hold schools accountable. In many cases, data are collected on a number of student background factors to enable analysts to estimate the 'value added' of schools after adjusting for students' family background. In addition, there has been a long-standing 'quest for school effects' – a search for the classroom and school factors that are associated with added value. A tenet of Educational Prosperity is that we already know which factors contribute to student success based on over thirty years of research on school effectiveness. These are the Foundations for Success. Educational Prosperity calls for a shift away from collecting data on a wide range of potential correlates of student achievement towards collecting detailed data on a small set of foundation factors. Data on the foundation factors can be used to assess equity of provision and set goals for system improvement.

Overall, these shifts in emphasis can be characterised as collecting 'leading indicators' to characterise the school system, rather than 'trailing indicators', which tend to be used mainly for accountability. Leading indicators can be used at the national and regional levels to inform policies aimed at improving outcomes, reducing inequalities in student outcomes, ensuring equity of provision, allocating resources and evaluating interventions. They can be used at the school and classroom levels to guide classroom practice, identify vulnerable students, set instructional goals and involve parents in meaningful ways.

Educational Prosperity is a simple model with a small number of foundational factors. As such, it provides a framework that can be used to set a vision for system improvement that can be shared by stakeholders at all levels of the school system – students, parents, teachers, principals and school administrators. System improvement will occur if all participants relentlessly strive to build a strong foundation with a single aim: giving all children the opportunity to thrive.

References

Adlof, S. M., H. W. Catts and J. Lee (2010). "Kindergarten predictors of second versus 8th grade reading comprehension impairments". *Journal of Learning Disabilities, 43*(4), 332-345.

Alexander, K. L. and B. K. Eckland (1975). "Contextual effects in the high school attainment process". American Sociological Revie*w, 4*, 402–416.

Alexander, K. L., J. Fennessey, E. L. McDill and R. J. D'Amico (1979). "School SES influences—Composition or context?". *Sociology of Education, 52*, 222–237.

Anderson, L. W. (2004). *Increasing Teacher Effectiveness* (2nd ed). Paris: UNESCO International Institute for Educational Planning (IIEP).

Australian Curriculum, Assessment and Reporting Authority (2017). *NAPLAN Achievement in Reading, Writing, Language Conventions and Numeracy: National Report for 2017.* Sydney: Australian Curriculum, Assessment and Reporting Authority.

Backhoff, E., A. Bouzas, E. Hernández and M. García (2007). *Aprendizaje y Desigualdad Social en México: Implicaciones de Política Educativa en el Nivel Básico*. Mexico: Instituto Nacional para la Evaluación de la Educación. ISBN: 968-5924-17-1

Barnett, S., S. Ayers and J. Francis (2015). *Comprehensive Measures of Child Outcomes in Early Years: Report to the OECD*. Paris: OECD.

Ben-Arieh, A. and I. Frønes (2007). "Indicators of children's well being: What should be measured and why?". *Social Indicators Research, 84*(3), 249-250.

Beswick, J.F., E. A. Sloat and J. D. Willms (2008). "Four educational myths that stymie social justice". *The Educational Forum, 72*(2), 115-128.

Bhattacharjee, Y. (2015). "Baby brains: The first year". *National Geographic, January 2015.* Retrieved from: http://ngm.nationalgeographic.com/2015/01/baby-brains/bhattacharjee-text

Black, R.E., H. A. Lindsay, A. Zulfiqar, L. E. Bhutta, M. Caulfield, E. Majid, Colin Mathers, et al. (2008). "Maternal and child undernutrition: Global and regional exposures and health consequences". *Lancet* 371 (9608): 243–60.

Black, M. M., P. Susan, L. C. H. Walker, C. T. Fernald, A. M. Andersen, C. L. D. DiGirolamo, C. McCoy, et al. (2017). "Early childhood development coming of age: Science through the life course". *Lancet, 389*(10064), 77–90.

Boyce, W. T. and M. S. Kobor (2015). "Development and the epigenome: The 'synapse' of gene-environment interplay". *Developmental Science, 18*(1), 1-23.

Boyce, W. T., M. B. Sokolowski and G. E. Robinson (2012). "Toward a new biology of social adversity". *PNAS, 109*(2), 17143-48.

Bryk, A. S. and M. E. Driscoll (1988). "The high school community: Contextual influences and consequences for students and teachers". Madison: National Center on Effective Secondary Schools, University of Wisconsin.

Caravolas, M., A. Lervåg, S. Defior, S. A. Málková and C. Hulme (2013). "Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies". *Psychological Science, 24*(8), 1398-1407.

Center on the Developing Child (2007). *In Brief: The Science of Early Childhood Development.* Cambridge, MA: Center on the Developing Child. Retrieved from http://developingchild.harvard.edu/resources/inbrief-science-of-ecd

Creemers, B. P. M. and L. Kyriakides (2006). "Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model". *School Effectiveness and School Improvement, 17*, 347–366.

Cynader, M. S. and B. J. Frost (1999). "Mechanisms of brain development: Neuronal sculpting by the physical and social environment". In D Keating, C. Hertzman (Eds.), *Developmental Health and the Wealth of Nations* (pp. 153–84). New York: Guilford.

Deary, I. J., S. Strand, P. Smith and C. Fernandes (2007). "Intelligence and educational achievement". *Intelligence, 35*, 13-21.

Doran, G. T. (1981). "There's a S.M.A.R.T. way to write management's goals and objectives". *Management Review. AMA FORUM. 70*(11), 35-36.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428-1446.

Earl, S., F. Carden and T. Smutylo (2001). "Outcome mapping: Building learning and reflection into development programs". Ottawa: International Development Research Centre.

Francis, D. J., S. E. Shaywitz, K. K. Stuebing, B. A. Shaywitz and J. M. Fletcher (1996). "Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis". *Journal of Educational Psychology*, *88,* 3-17.

Gonzales, E. (2016). *Calculating Standard Errors in International Large-Scale Studies.* Princeton: Educational Testing Service.

Good, R.H., III, D. C. Simmons and E. J. Kame'enui (2001). "The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for 3rd grade high-stakes outcomes". *Scientific Studies of Reading, 5,* 257-288.

Gough, P. B. and W. E. Tunmer (1986). "Decoding, reading and reading disability". *Remedial and Special Education, 7*(1), 6-10.

Grek, S. (2009). "Governing by numbers: the PISA 'effect' in Europe". *Journal of Education Policy, 24*(1), 23-37. DOI: 10.1080/02680930802412669

Hanushek, E. A. and L. Woessmann (2015). *The Knowledge Capital of Nations: Education and the Economics of Growth*. Cambridge and London: MIT Press.

Hattie, J. (2009). *Visible Learning: A Synthesis of over 800 Meta-Analyses relating to Achievement.* New York: Routledge.

Hertzman, C. (1999). "The biological embedding of early experiences and its effects on health in adulthood". *Annals of the New York Academy of Sciences, 896*, 85-85.

Heugh, K. A. (2013). "Multilingual education policy in South Africa constrained by theoretical and historical disconnections". *Annual Review of Applied Linguistics, 33*, 215.

Kagan, S. L., E. Moore and S. Bredekamp (Eds.). (1995). *Reconsidering Children's Early Learning and Development: Toward Shared Beliefs and Vocabulary.* Washington, DC: National Education Goals Panel.

Knudsen, E. I. (2004). "Sensitive periods in the development of the brain and behavior". *Journal of Cognitive Neuroscience, 16*(8), 1412-1425.

Knudsen, E. I., J. J. Heckman, J. L. Cameron and J. P. Shonkoff (2006). "Economic, neurobiological and behavioral perspectives on building America's future workforce". *Proceedings of the National Academy of Sciences, 103*(27), 10155-10162.

Kraemer, H. C., A. E. Kazdin, D. R. Offord, R. C. Kessler, P. S. Jensen and D. J. Kupfer (1997). "Coming to terms with the terms of risk". *Archives of General Psychiatry, 54*, 337-343.

Kyriakides, L., C. Christoforou and C. Y. Charalambous (2013). "What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching". *Teaching and Teacher Education, 36*, 143–152.

Lee, V. E. and J. B. Smith (1993). "Effects of school restructuring on the achievement and engagement of middle-grade students". *Sociology of Education, 66* (July), 164-187.

Lee, V. E. and J. B. Smith (1995). "Effects of high school restructuring and size on early gains in achievement and engagement". *Sociology of Education, 68*(4), 241-270.

Leppänen, U., P. Niemi, K. Aunola and J.-E. Nurmi (2004). "Development of reading skills among preschool and primary school pupils". *Reading Research Quarterly*, 39, 72–93.

Levin, H. (2009). "The economic payoff to investing in educational justice". *Educational Researcher, 28*(1), 5-20.

Lewin, K. (2015). *Educational Access, Equity and Development: Planning to Make Rights Realities.* Paris: UNESCO International Institute for Educational Planning (IIEP)..

Lopez, A. Y. (2016). „Steps Towards a National Implemetnation of the Early Years Evaluation in Uruguay: Actors and Actions". Paper presented at the Inaugural Symposium of the Comparative and International Education Society, Scottsdale, AZ.

Lorenz, M. O. (1905). "Methods of measuring the concentration of wealth". *Publications of the American Statistical Association, 9*(70), 209-219.

Manyike, T. V. (2012). "A comparison of reading and writing proficiency in home language among Xitsonga speaking learners in South African primary township schools". *International Journal of Education Science, 4*(2), 143-152.

Manyike, T. V. (2013). "Bilingual literacy or substantive bilingualism? L1 and L2 reading and writing performance among Grade 7 learners in three township schools Gauteng Province, South Africa". *Africa Education Review, 10(*2), 187-203. DOI: 10.1080/18146627.2013.812271

Martínez, F. and M. A. Díaz (2016). *México en PISA 2015*. Mexico: Instituto Nacional para la Evaluación de la Educación. Retrieved from: http://publicaciones.inee.edu.mx/buscadorPub/P1/D/316/P1D316.pdf

Matts, J. and J. Lachin (1988). "Properties of permuted-block randomization in clinical trials". *Control Clinical Trials, 9*, 327–344.

McEwen, B. S. and H. M. Schmeck Jr. (1994). *The Hostage Brain.* New York, NY, US: Rockefeller University Press.

McClelland, M. M., F. J. Morrison and D. L. Holmes (2000). "Children at risk for early academic problems: The role of learning-related social skills". *Early Childhood Research Quarterly, 15*(3), 307-329.

McClelland, M. M., C. C. Ponitz, E. E. Messersmith and S. Tominey (2010). "Self-regulation: The integration of cognition and emotion". In R. Lerner (Series Ed.) and W. Overton (Vol. Ed.), *Handbook of lifespan human development: Vol. 1. Cognition, biology and methods* (pp. 509-553). Hoboken, NJ: Wiley.

Mislevy, R. J., A. E. Beaton, B. Kaplan and K. M. Sheehan (1992). "Estimation population characteristics from sparse matrix samples of item responses". *Journal of Educational Measurement, 29*, 133-161.

Mueller, W. M. and T. L. Parcel (1981). "Measures of socioeconomic status: Alternatives and recommendations". *Child Development, 52,* 13-30.

Murillo, F. J. and M. Román (2011). "School infrastructure and resources do matter: Analysis of the incidence of school resources on the performance of Latin American students". *School Effectiveness and School Improvement*, *22*(1), 29-50.

Nation, K. and M. J. Snowling (2004). "Beyond phonological skills: Broader language skills contribute to the development of reading". Journal of Research in Reading, Vol 27(4), 342-356.

National Early Literacy Panel (2008). *Developing early literacy: Report of the National Early Literacy Panel.* Washington, DC: National Institute for Literacy. Retrieved from http://lincs.ed.gov/publications/pdf/NELPReport09.pdf

Nonoyama-Tarumi, Y. and J. D. Willms (2010). "The relative and absolute risks of disadvantaged family background and low levels of school resources on student literacy". *Economics of Education Review, 29*(2), 214-224.

OECD (2001). *Knowledge and Skills for Life – First Results from PISA 2000.* Paris: OECD.

OECD (2017). *PISA for Development Assessment and Analytical Framework: Reading, Mathematics and Science, Preliminary Version* Paris: OECD Publishing.

Perfetti, C. A., N. Landi and J. Oakhill (2005). "The acquisition of reading comprehension skill". In M. J. Snowling and C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 227-247). Oxford, UK: Blackwell.

Polo, D. S. Z., N. P. A. Araujo and J. C. R. Salceda (2017). "La concepción simple de la lectura en alumnos de 4º de primaria de una escuela fiscal de Quito". *Alteridad, 12*(1), pp. 228-235.

Raudenbush, S. W. and J. D. Willms (1995). "The estimation of school effects". *Journal of Educational and Behavioural Statistics. 20*(4), 307-335.

Raver, C. C., S. M. Jones, C. Li-Grining, F. Zhai, K. Bub and E. Pressler (2011). "The Chicago School Readiness Project's impact on low income preschoolers' preacademic skills: Self-regulation as a mediating mechanism". *Child Development, 82,* 362-378. doi:10.1111/j.1467-8624.2010.01561.x

Rhode Island Kids Count (2005). *Getting Ready: Findings from the National School Readiness Indicators Initiative, a 17 State Partnership*. Providence, RI: Rhode Island Kids Count.

Riehl, C. J. (2000). "The principal's role in creating inclusive schools for diverse students: A review of normative, empirical and critical literature on the practice of educational administration". *Review of Educational Research, 70*(1), 55-81.

Ripoll, J. C., G. Aguado and A. P. Castilla-Earls (2014). "The simple view of reading in elementary school: A systematic review". *Revista de Logopedia, Foniatría y Audiología, 34*, 17-31.

Robertson, D. and J. Symons (1996). *Do Peer Groups Matter? Peer Group versus Schooling Effects on Academic Attainment.* London: London School of Economics, Centre for Economic Performance.

Rose, J. (2006). *Independent Review of the Teaching of Early Reading.* London, HMSO: Department for Education and Skills.

Rosenshine, B. (2010). *Principles of Instruction.* International Academy of Education. Geneva: UNESCO International Bureau of Education (IBE).

Ross, C. E. and C. Wu (1995). "The links between education and health". *American Sociological Review, 60* (5), 719-745.

Rowan, R., S. W. Raudenbush and S. J. Kang (1991). "School climate in secondary schools: A multilevel analysis". In S.W. Raudenbush and J.D. Willms (Eds.), *Pupils, Classrooms and Schools: International Studies of Schooling from a Multilevel Perspective.* New York: Academic Press.

Rust, K. (1985). "Variance estimation for complex estimators in sample surveys". *Journal of Official Statistics, 1*(4), 381-397.

Särndal, C. E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scarborough, H. S. (1989). "Prediction of reading disability from familial and individual differences". *Journal of Educational Psychology*, *81*(1), 101-108.

Scarborough, H. S. (2001). "Connecting early language and literacy to later reading (dis)abilities: Evidence, theory and practice". In S. B. Neuman and D. K. Dickinson (Eds.), Handbook of early literacy research (pp. 97–110). New York: Guilford Press.

Schatschneider, C., J. M. Fletcher, D. J. Francis, C. D. Carlson and B. R. Foorman (2004). "Kindergarten prediction of reading skills: A longitudinal comparative analysis". *Journal of Educational Psychology*, *96*(2), 265-282.

Shonkoff, J. P. and D. A. Phillips (2000). *From Neurons to Neighborhoods: The Science of Early Childhood Development.* Washington, DC: National Academy Press.

Simola, H. (2005). "The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education". *Comparative Education, 41*(4), 455-70.

Singer, J. D. and H. I. Braun (2018). "Testing international education assessments". *Science, 360*(6384), 38-40. http://science.sciencemag.org/content/360/6384/38

Snow, C. E., M. S. Burns and P. Griffin (1998). *Preventing Reading Difficulties in Young Children.* Washington DC: National Academy Press.

Storch, S. and G. Whitehurst (2002). "Oral language and code-related precursors to reading: Evidence from a longitudinal structural model". *Developmental Psychology, 38*(6), 934-947. doi:10.1037//0012-1649.38.6.934

The Learning Bar Inc. (2009). OurSCHOOL Student Survey [Measurement instrument]. Retrieved from http://www.thelearningbar.com/

The Learning Bar Inc. (2011). Early Years Evaluation [Measurement instrument]. Retrieved from http://www.thelearningbar.com/

The Learning Bar Inc. (2016). Confident Learners [Measurement instrument]. Retrieved from http://www.thelearningbar.com/

Torgesen, J., S. Otaiba and M. Grek (2005). "Assessment and instruction for phonemic awareness and word recognition skills". In Catts, H. and Kamhi, A. (Eds.), *Reading and Language Disabilities* (pp. 112-145). Boston, MA: Allyn and Bacon.

Tramonte, L. and J. D. Willms (2018). *New Measures for Comparative Studies of Low- and Middle-Income Countries.* Manuscript submitted for publication.

Trzesniewski, K. H., T. E. Moffitt, A. Caspi, A. Taylor and B. Maughan, B. (2006). "Revisiting the association between reading ability and antisocial behavior: New evidence from a longitudinal behavior genetics study". *Child Development*, 77, 72-88.

United Nations (2015). *Transforming our World: The 2030 Agenda for Sustainable Development. https://sustainabledevelopment.un.org/post2015/transformingourworld*

UNESCO (2005). *Guidelines for Inclusion: Ensuring Access to Education for All.* Paris: UNESCO.

UNESCO Institute for Statistics (UIS) (2017). *The Quality Factor: Strengthening National Data to Monitor Sustainable Development Goal 4*. Montreal: UNESCO Institute for Statistics.

UNICEF (2012). *School Readiness: A Conceptual Framework*. New York, NY: United Nations Children's Fund.

van der Berg, S., N. Spaull, G. Wills, M. Gustafsson and J. Kotzé (2016). "Identifying binding constraints in education: Synthesis report for the Programme to support Pro-poor Policy development (PsPPd)". University of Stellenbosch: Department of Economics.

Vellutino, F. R. and D. M. Scanlon (1987). « Phonological coding, phonological awareness and reading ability: Evidence from a longitudinal and experimental study". *Merrill-Palmer Quarterly, 33*(3), 321-363.

Verhoeven, L., J. van Leeuwe and A. Vermeer (2011). "Vocabulary growth and reading development across the school years". *Scientific Studies of Reading, 15*(1), 8-25.

Vidal, R., M. A. Díaz and H. Jarquín (2004). *Resultados de las Pruebas PISA 2000 y 2003 en México: Habilidades para la Vida en Estudiantes de 15 Años*. Mexico: Instituto Nacional para la Evaluación de la Educación. Retrieved from http://publicaciones.inee.edu.mx/buscadorPub/P1/D/202/P1D202.pdf

Warwick, L. (2005). "Words to grow on". *Bulletin of the Centre of Excellence for Early Childhood Development, 4(*1), 2-4.

Willms, J. D. (1986). "Social class segregation and its relationship to pupils' examination results in Scotland". *American Sociological Review*, *51,* 224-241.

Willms, J. D. and S. W. Raudenbush (1989). "A longitudinal hierarchical linear model for estimating school effects and their stability". *Journal of Educational Measurement*, *26(3),* 209-232.

Willms, J. D. and M. A. Somers (2001). "Family, classroom and school effects on children's educational outcomes in Latin America". *International Journal of School Effectiveness and Improvement*, *12*(4)*,* 409-445.

Willms, J. D. (2011). "Measures of educational equality and equity: A methodological note for the INES Network for the Collection and the Adjudication of System-Level Descriptive Information on Educational Structures, Policies and Practices (NESLI)". Paper presented at the March 2011 OECD NESLI meeting, Netherlands.

Willms, J.D. (2003a). *Ten Hypotheses about Socioeconomic Gradients and Community Differences in Children's Developmental Outcomes*. Ottawa, ON: Applied Research Branch of Human Resources Development Canada.

Willms, J.D. (2003b). "Literacy proficiency of youth: Evidence of converging socioeconomic gradients". *International Journal of Educational Research, 39*(3), 247-252.

Willms, J. D. (2006). *Learning Divides: Ten Policy Questions about the Performance and Equity of Schools and Schooling Systems*. Montreal: UNESCO Institute for Statistics.

Willms, J. D. (2008). "The case for universal French instruction". *Policy Options, 29*(7), 91-96.

Willms, J. D. (2009a, October). "Classroom diversity and Inclusion: The Educational Advantage"**.** Plenary presentation at the Return to Salamanca – Global Conference on Inclusive Education. Salamanca, Spain.

Willms, J. D. (2009b). "Feasibility study for a First Nations assessment system". Ottawa: Indian and Northern Affairs Canada.

Willms, J. D. (2010). "School composition and contextual effects on student outcomes".

Willms, J. D., L. Tramonte, J. Duarte and S. Bos (2012). *Assessing Educational Equality and Equity with Large-Scale Assessment Data: Brazil as a Case Study.* Washington: Inter-American Development Bank.

Willms, J. D. and L. Tramonte (2018). *The Measurement and Use of Socioeconomic Status in Educational Research.* Manuscript submitted for publication.

Willms, J. D. (2018a). *Educational Prosperity.* Fredericton, NB: The Learning Bar Inc.

Willms, J. D. (2018b). "Educational Prosperity: An assessment strategy for supporting student learning in low-income countries". In D. A. Wagner, S. Wolf and R. F. Boruch (Eds.), *Learning at the Bottom of the Pyramid: Science, Measurement and Policy in Low-Income Countries.* Paris: UNESCO-IIEP.

World Bank (2018a). World Bank national accounts data and OECD National Accounts data files. Washington, D. C.: World Bank. Retrieved from https://data.worldbank.org/indicator/NY.GNP.MKTP.KD

World Bank (2018b). *World Development Report 2018: Learning to Realize Education's Promise.* Washington, DC: World Bank. DOI:10.1596/978-1-4648-1096-1

World Health Organisation (WHO) (2010). "Disorders related to short gestation and low birth weight, not elsewhere classified". *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) version for 2010*. Retrieved from: http://apps.who.int/classifications/icd10/browse/2010/en#/P07

Zimmer, R. W. and E. F. Toma (1997). *Peer Effects in Private and Public Schools: A Cross-Country Empirical Analysis.* Lexington: University of Kentucky.

Appendix 1. Growth rates in PISA reading proficiency, 2000-2015

	Baseline (2000)	(SE)	Annual growth	(SE)	SES	(SE)
Australia	**517.1**	(2.4)	-1.49	(0.24)	**43.8**	(1.0)
Austria	**500.1**	(3.9)	-1.79	(0.44)	**46.0**	(1.8)
Belgium	**507.2**	(3.0)	-0.81	(0.32)	**48.2**	(1.1)
Canada	**520.0**	(2.8)	-0.62	(0.25)	**33.9**	(1.5)
Czechia	**488.5**	(4.4)	0.37	(0.39)	**49.0**	(1.6)
Denmark	**491.4**	(3.0)	-0.74	(0.27)	**37.3**	(1.7)
Finland	**548.3**	(2.5)	**-2.45**	(0.30)	**31.7**	(1.6)
France	**521.2**	(8.8)	**-2.00**	(0.84)	**50.7**	(3.4)
Germany	**490.0**	(5.5)	0.80	(0.70)	**45.4**	(3.2)
Greece	**478.1**	(4.1)	-0.16	(0.31)	**34.4**	(0.6)
Hungary	**491.4**	(5.2)	-0.12	(0.52)	**47.9**	(2.1)
Iceland	**487.6**	(3.7)	**-1.87**	(0.33)	**26.7**	(0.9)
Ireland	**524.0**	(3.6)	-1.00	(0.40)	**38.0**	(0.6)
Italy	**483.8**	(6.2)	0.30	(0.48)	**32.1**	(1.0)
Japan	**504.6**	(2.2)	**2.20**	(0.72)	**37.9**	(3.6)
Korea	**530.1**	(7.2)	1.04	(0.89)	**32.0**	(4.1)
Luxemburg	**465.6**	(2.9)	**1.00**	(0.19)	**40.4**	(1.0)
Mexico	**456.5**	(37.7)	-1.16	(2.54)	**26.6**	(13.0)
Netherlands	**520.4**	(6.2)	**-1.78**	(0.58)	**38.9**	(1.1)
Norway	**486.1**	(6.1)	-.07	(0.69)	**37.3**	(1.9)
New Zealand	**523.8**	(3.0)	**-1.10**	(0.26)	**47.7**	(1.5)
Poland	**496.6**	(4.0)	**2.18**	(0.51)	**40.2**	(1.4)
Portugal	**491.3**	(2.4)	**1.14**	(0.19)	**31.5**	(1.7)
Slovak Republic	**478.1**	(3.6)	-0.99	(0.61)	**47.7**	(2.3)
Spain	**486.8**	(13.0)	0.96	(1.27)	**29.5**	(1.0)
Sweden	**506.1**	(2.7)	**-1.65**	(0.51)	**39.8**	(1.0)
Switzerland	**504.8**	(3.0)	-0.92	(0.39)	**42.1**	(2.1)
United Kingdom	**511.6**	(11.6)	-1.80	(1.95)	**43.0**	(2.0)
United States	**494.6**	(6.7)	-0.20	(0.55)	**40.5**	(3.5)

Appendix 2. Core statistics for informing educational policy

Estimation of descriptive statistics in large-scale international studies

Replication weights. The international studies use one of two techniques to take account of the sample design. The jackknife technique is used for PIRLS and TIMSS, while balanced repeated replication (BRR) is used for PISA *(see Rust, 1985).* The data set for PISA 2015, for example, provides an overall study design weight and a set of 80 BRR weights. Statistics such as a mean or standard deviation can be estimated using the overall design weight, however, to obtain accurate estimates of the standard error of a statistic, one must use the BRR weights. When the statistic involves a test score, the BRR weights are used together with the plausible values.

Plausible values. The test content for international studies is organised into various booklets and each student is randomly assigned to a booklet. In PISA 2015, for example, the study included 810 minutes of test items in reading, science, mathematics and collaborative problem solving, while each student only completed 120 minutes of test content. Each test is conceived to have a proficiency dimension associated with the full set of items. The statistical techniques used in PISA provide estimates of the proficiency distributions of the tests as well as the likelihood of a correct response for each student on each item, had they completed the full set of test items (Mislevy, Beaton, Kaplan, Sheehan, 1992). The student data set includes a set of ten plausible values for each student. These values are randomly selected from the estimated ability distribution for students with a similar response pattern and family backgrounds. In essence, the technique provides ten estimates or *plausible values* that indicate how well each student might have performed if he or she had completed the full test.

Statistics such as a mean score, a standard deviation, or regression coefficients are calculated in the usual way, using the overall design weight. The standard error of the statistic of interest, if it does not involve plausible values, is given by:

$$SE_\varepsilon = \sqrt{\frac{\sum_{r=1}^{R}(\varepsilon_r - \varepsilon_0)^2}{R(1-0.5)^2}}$$

where ε_0 is the statistic of interest, computed using the full sampling weight, ε_r is the statistic of interest computed using the replicate weight, *r* and R is the number of replicates, which for PISA is 80.

The standard error of the mean, if it does involve plausible values, is given by:

$$SE_\varepsilon = \sqrt{\left[\sum_{p=1}^{P}\left(\frac{\sum_{r=1}^{R}\left(\varepsilon_{r,p} - \varepsilon_{0,p}\right)^2}{R(1-0.5)^2}\right) * \frac{1}{P}\right] + \left[\left(1+\frac{1}{P}\right)\frac{\sum_{p=1}^{P}\left(\varepsilon_{0,p} - \bar{\varepsilon}_{0,P}\right)^2}{P-1}\right]}$$

where $\varepsilon_{r,p}$ is the statistic of interest computed using plausible value, p and replicate weight *r* and P is the number of plausible values, which for PISA is 10.

These formulae seem rather complicated for the non-statistical reader. However, if one constructs an 80 by 1 matrix of the replicate weights for statistics that do not involve plausible values, or an 80 by 10 matrix (80 BRRs by 10 PVs) for statistics that involve plausible values, then a simple formula can be applied. This approach can be used with any statistic, such as an estimate of skewness, a segregation index, or regression coefficients.

Gonzales (2016) provides a useful document which sets out the formulae for most of the major international studies.

[1] Mean. The formula for the mean is: $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, where *n* is the sample size and Y_i is the score for the i^{th} student. The mean is calculated using the overall design weight and the standard error is calculated using one of the two formulae above, depending on whether or not the mean is for a test score.

[2] Standard deviation. The formula for the standard deviation is: $SD = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$, where *n* is the sample size, y_i is the score for the i^{th} student and $\bar{y}$ is the mean. The standard deviation is calculated using the overall design weight and the standard error is calculated using one of the two formulae above, depending on whether or not the standard deviation is for a test score.

[3] Skewness. The formula for the skewness is: $G = \frac{\sqrt{n\,(n-1)}}{n-2}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{\frac{3}{2}}}\right]$

where *n* is the sample size, y_i is the score for the i^{th} student and $\bar{y}$ is the mean. The skewness is calculated using the overall design weight and the standard error is calculated using one of the two formulae above, depending on whether or not the skewness is for a test score.

[4] Percent vulnerable. The formula for the percent vulnerable is: $P = \frac{\sum_{i=1}^{n} D_i}{n}$, where *n* is the sample size and D_i is the dichotomous score (0 or 1) for the i^{th} student. The percent vulnerable is calculated using the overall design weight and the standard error is calculated using one of the two formulae above, depending on whether or not the percent variable is for a test score.

Socioeconomic gradients

Socioeconomic gradients are estimated with an OLS regression model which includes SES and SES^2:

$$Y_i = \beta_0 + \beta_1 SES_i + \beta_2 SES_i^2 + r_i$$

where Y_i is the i^{th} student's reading score and SES_i is the SES for the i^{th} student. The parameters, r_i, are the student-level residuals; that is, the deviation of students' scores from the regression line. Four statistics are estimated with this equation:

[5] Gradient level: The intercept, β_0, is the expected score for a student with an SES score of zero. The SES variable can be 'centered' on any particular value, but usually it is convenient to centre it on the country average, or in the case of the PISA analyses, on the OECD average.

[6] Gradient slope. The coefficient, β_1, is the slope of the socioeconomic gradient. The gradient hypothesis is: $H_0: \beta_1 = 0;\ H_1: \beta_1 \neq 0$. The statistical significance of β_1, depends on the magnitude of the standard error and is assessed with a t-test with n-1 degrees of freedom.

[7] Diminishing returns. The coefficient, β_2, is the slope for SES-squared. When β_2 is negative, there are weaker effects on outcomes associated with SES at higher levels of SES. The hypothesis of diminishing returns is: $H_0: \beta_2 = 0;\ H_1: \beta_2 \neq 0$. The statistical significance of β_2, depends on the magnitude of the standard error and is assessed with a t-test with n-1 degrees of freedom.

[8] Gradient strength. The strength of the gradient, called R^2, is the proportion of variance in the outcome measure explained by SES. It is the difference between the variance of Y_i and the variance of the residuals, r_i, expressed as a fraction of the variance of Y_i.

To obtain the standard error for these statistics one must estimate the regression analysis 80 times (one for each BRR) when the outcome is a variable that is not a test score. When the outcome is a test score, the regression analysis needs to be estimated 800 times, one for each BRR and each plausible value. As noted earlier, one requires either an 80 by 1 or an 80 by 10 matrix for each of the statistics.

Inclusion indices

[9] Vertical inclusion index. Vertical inclusion is the proportion of variation in the outcome score that is *within* schools:

$$V = \frac{\sigma^2}{\tau}$$

The estimate of vertical inclusion is obtained by fittng a 'null' hierarchical linear regression model (HLM) to the outcome data:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim NID\ (0, \sigma^2)$$

$$\beta_{0j} = \gamma_{00} + U_{0j} \qquad U_{oj} \sim NID(0, \tau)$$

In HLM, the variance of the Level 1 error terms, which is the variance *within* schools, is called σ^2. The variance of the Level 2 error terms, which is the variance *between* schools, is called τ.

$$Var(\varepsilon_{ij}) = \sigma^2 \qquad Var(U_{0j}) = \tau$$

[10] horizontal Inclusion Index. Horizontal inclusion is the proportion of variation in SES that is *within* schools:

$$H = \frac{\sigma^2}{\tau}$$

It is estimated by estimating a null model with SES as the outcome variable.

Multilevel gradient model

The basic multilevel gradient model has two levels:

(1) $Y_{ij} = \beta_{oj} + \beta_{1j}SES_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim NID\ (0,\ \sigma^2)$

(2) $\beta_{0j} = \gamma_{00} + U_{0j} \qquad U_{0j} \sim NID\ (0, \tau_0)$

$\beta_{1j} = \gamma_{10} + U_{1j} \qquad U_{0j} \sim NID\ (0, \tau_1)$

where Y_{ij} is the outcome score for the i^{th} student in the j^{th} school, SES_{ij} is the SES score for the i^{th} student in the j^{th} school and ε_{ij} are the residuals, which are assumed to have a normal distribution with a mean of zero and a variance of σ^2. The levels of the gradients are β_{oj}, which are the expected scores of students whose SES is zero. Thus, they are referred to as SES-adjusted means.

[11] Within-school gradient. The within-school gradient slopes are β_{1j}. They are modelled at Level 2 as an average slope, γ_{10}, plus a residual from the average slope, U_{0j}.

[12] Converging gradient index. The variance of the SES-adjusted means, the β_{oj}'s, is the variance of U_{0j}, which is τ_0. Similarly, the variance of the within-school gradients, the β_{1j}'s, is the variance of U_{1j}, which is τ_1.

HLM provides an estimate of the covariance of the U_{0j} and U_{1j}. The correlation of the SES-adjusted means and the slopes, which is referred to here as the Converging Gradient Index, is given by:

$$CGI = \frac{Cov(U_0, U_1)}{\tau_0 \tau_1}$$

[13] Between-school gradient. The between-school gradient, which is a regression of the school means of the outcomes on school mean SES can also be estimated with a two-level model:

(1) $Y_{ij} = \beta_{oj} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim NID\ (0,\ \sigma^2)$

(2) $\beta_{0j} = \gamma_{00} + \gamma_{01}\overline{SES}_{.j} + U_{0j} \qquad U_{0j} \sim NID\ (0, \tau_0)$

The between-school slope is γ_{01}.

[14] Mean SES composition effect. The school composition effect for mean SES is estimated by adding the mean SES of the schools to the equation for the adjusted school means at Level 2:

(1) $Y_{ij} = \beta_{oj} + \beta_{1j}SES_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim NID\ (0,\ \sigma^2)$

(2) $\beta_{0j} = \gamma_{00} + \gamma_{01}\overline{SES}_j + U_{0j} \qquad U_{0j} \sim NID\ (0, \tau_0\ )$

$\beta_{1j} = \gamma_{10} + U_{1j} \qquad U_{0j} \sim NID\ (0, \tau_1)$

The estimate of γ_{01} is the school composition effect for mean SES.

[15] Variable intake composition effect. The school composition effect for intake variability is estimated in the same way as the mean SES composition effect: one adds the standard deviation of SES to the first Level 2 equation. The school composition effect for intake variability is γ_{01}.