

# Development of Reporting Scales for Reading and Mathematics

A report describing the process for  
building the UIS Reporting Scales

April 2018

The ACER Centre for Global Education Monitoring supports the monitoring of educational outcomes worldwide, holding the view that the systematic and strategic collection of data on education outcomes, and factors related to those outcomes, is required to inform high quality policy aimed at improving educational progress for all learners.



United Nations  
Educational, Scientific and  
Cultural Organization



UNESCO  
INSTITUTE  
FOR  
STATISTICS



Australian Government  
Department of Foreign Affairs and Trade



## Development of Reporting Scales for Reading and Mathematics: A report describing the process for building the UIS reporting Scales

©The Australian Council *for* Educational Research Ltd (ABN: 19 004 398 145). 2018.

This publication has been funded by the Australian Government through the Department of Foreign Affairs and Trade, and The Australian Council *for* Educational Research Ltd. The views expressed in this publication are the author's alone and are not necessarily the views of the Australian Government.

This report is licensed under the Creative Commons [Attribution-noncommercial-noderivatives 4.0 International Licence](https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)<<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>>. Unless otherwise stated, you may only share and use this report for non-commercial purposes, and you may not make and distribute any derivatives of it.

**Please give attribution to:** © The Australian Council *for* Educational Research Ltd (ABN: 19 004 398 145). 2018.

We also request that you observe and retain any copyright or related notices that may accompany this material as part of the attribution. This is also a requirement of the Creative Commons Licences.

### Further Information

For further information about the copyright in this website, please contact:

Legal and Commercial Manager

Corporate Services

The Australian Council for Educational Research Ltd

19 Prospect Hill Road Camberwell

CAMBERWELL VIC 3124

Phone: +61 3 9277 5788 or email: [david.noga@acer.org](mailto:david.noga@acer.org)

ISBN: 978-1-74286-509-6

# Contents

Abbreviations .....	4
List of Figures .....	5
List of Tables.....	6
Acknowledgements .....	7
1 Background .....	8
1.1 Understanding the approach .....	9
1.2 Mapping a learning domain .....	9
1.3 Adopting a vertical scale .....	13
Vertical scales used by other countries.....	15
1.4 Development process .....	17
2 Conceptual frameworks .....	18
2.1 Conceptual framework for reading .....	18
2.1.1 Comprehension .....	19
2.1.2 Constrained skills .....	20
2.2 Conceptual framework for mathematics .....	23
3 Identifying conceptual demand .....	26
3.1 Data sources .....	26
3.2 Analysis of item conceptual demands .....	28
3.3 Central observations .....	28
3.3.1 Broad ideas about growth .....	28
3.3.2 Conceptualisation of the scale .....	29
3.3.3 Language considerations .....	30
4 Constructing a single scale for each domain .....	31
4.1 Pairwise comparison .....	31
4.2 Method .....	31
4.3 Design.....	32
4.3.1 Formation of item pairs.....	32
4.3.2 Supporting reliability in rater judgements .....	34
4.3.3 Rating process.....	34
4.3.4 Outcomes .....	35
4.3.5 Considerations for creating the reporting scales .....	44
4.4 Validation of item difficulty estimates and mapping of growth .....	45
4.4.1 Qualitative validation of the pairwise item ordering .....	45

4.4.2	Comparison of the outcomes from the pairwise comparison study with other empirical sources on item difficulty .....	46
4.4.3	Empirical validation study in collaboration with the Korea Institute for Curriculum and Evaluation (KICE) .....	56
5	Drafting reporting scale descriptions .....	58
5.1	Reading .....	59
5.2	Mathematics .....	61
5.2.1	Content .....	60
5.2.2	Complexity .....	62
5.2.3	Context .....	63
5.2.4	Competencies .....	63
5.3	Illustrative examples .....	63
5.4	Progression elements .....	64
6	The Learning Progression Explorer .....	65
7	Noted limitations .....	67
8	Conclusion and next steps .....	69
	Appendices .....	70
	<b>Appendix 1:</b> Framework for foundational concepts, knowledge, skills and applications for reading .....	70
	<b>Appendix 2:</b> Framework for foundational concepts, knowledge, skills and applications for mathematics .....	77
	<b>Appendix 3:</b> Task examples designed to illustrate the levels of the mathematics scale .....	81
	<b>Appendix 2:</b> Task examples designed to illustrate the levels of the reading scale .....	84
	References .....	85

# Abbreviations

<b>ACER</b>	Australian Council for Educational Research
<b>ACER-GEM</b>	ACER Centre for Global Education Monitoring
<b>ASER</b>	Annual Status of Education Report
<b>EGMA</b>	Early Grade Mathematics Assessment
<b>EGRA</b>	Early Grade Reading Assessment
<b>GAML</b>	Global Alliance to Monitor Learning
<b>IEA</b>	International Association for the Evaluation of Educational Achievement
<b>IRT</b>	Item Response Theory
<b>OECD</b>	Organisation for Economic Co-operation and Development
<b>PASEC</b>	Programme d'Analyse des Systèmes Educatifs des Pays de la CONFEMEN
<b>PILNA</b>	Pacific Islands Literacy and Numeracy Assessment
<b>PIRLS</b>	Progress in International Reading Literacy Study
<b>PISA</b>	Programme for International Learner Assessment
<b>SACMEQ</b>	Southern and Eastern Africa Consortium for Monitoring Educational Quality
<b>SDG</b>	Sustainable Development Goal
<b>TIMSS</b>	Trends in International Mathematics and Science Study
<b>UIS</b>	UNESCO Institute for Statistics
<b>UNESCO</b>	United Nations Educational, Scientific and Cultural Organisation

## List of Figures

<b>Figure 1</b>	Example learning metric for mathematics (from Turner, 2014b) .....	11
<b>Figure 2</b>	Q-Q plot of residuals for mathematics for the full set of comparisons and a trimmed set (2.5, 97.5%) .....	36
<b>Figure 3</b>	Q-Q plot of residuals for reading for the full set of comparisons and a trimmed set (2.5, 97.5%) .....	37
<b>Figure 4</b>	Estimated difficulty of mathematics items grouped by source assessment program.....	40
<b>Figure 5</b>	Estimated difficulty of reading items grouped by source assessment program .....	40
<b>Figure 6</b>	Relative difficulty of mathematics items from selected source assessment programs.....	41
<b>Figure 7</b>	Relative difficulty of reading items from selected source assessment programs .....	42
<b>Figure 8</b>	Observed trends between pairwise and published item difficulties .....	48
<b>Figure 9</b>	The association between the pairwise and published item difficulties for the TIMSS items only.....	49
<b>Figure 10</b>	TIMSS item M052173.....	50
<b>Figure 11</b>	TIMSS item M052362.....	50
<b>Figure 12</b>	TIMSS item M042152.....	51
<b>Figure 13</b>	TIMSS item M052429.....	51
<b>Figure 14</b>	Plot of item difficulty estimates for a selection of reading items from the MTEG Afghanistan: pairwise estimates versus empirical estimates from the source assessment program .....	52
<b>Figure 15</b>	MTEG item R0003T06P The Hole .....	53
<b>Figure 16</b>	MTEG item R0015T03P School Friends.....	54
<b>Figure 17</b>	MTEG item R0002T01P Party .....	55
<b>Figure 18</b>	MTEG item R0019T01P A Brother’s Note .....	56

## List of Tables

<b>Table 1</b>	PISA and TIMSS mathematical content categories.....	25
<b>Table 2</b>	Source of items used in the conceptual demand analysis .....	27
<b>Table 3</b>	Grouping of items for pairs formation in mathematics .....	33
<b>Table 4</b>	Grouping of items for pairs formation in reading .....	33
<b>Table 5</b>	Agreement with true ratings for each rater.....	35
<b>Table 6</b>	The proportion of mathematics items included in the pairwise comparison.....	39
<b>Table 7</b>	The proportion of reading items included in the pairwise comparison.....	39
<b>Table 8</b>	Summarised findings for association between the pairwise and published item difficulties .....	47

# Acknowledgements

The 'Development of Reporting Scales' document was authored by the Australian Council for Educational Research Centre for Global Education Monitoring (ACER-GEM), in collaboration with the UNESCO Institute for Statistics (UIS).

The process described in this document draws on the extensive experience of ACER in developing learning metrics. Sincere recognition is made of the work of the very many people whose efforts have contributed to the development, conduct and reporting of assessments over the years, on whose accumulated experience and expertise this present document is based.

Major contributions to this document were made by Ross Turner, Ray Adams, Ursula Schwantner, Dan Cloney, Claire Scoular, Prue Anderson, Alex Daraganov, Jennifer Jackson, Sandra Knowles, Gayl O'Connor, Pam Munro-Smith, Stavroula Zoumboulis, and Pauline Rogers.



# I Background

This report describes the development of the UNESCO Institute for Statistics reporting scales (UIS-RS) for reading and mathematics, created with the aim of enabling countries to examine and report the outcomes of their assessment activities using a common methodology.

This work is being undertaken by the Centre for Global Education Monitoring at the Australian Council for Educational Research (ACER-GEM) who are technical partners to the UIS.

ACER initially conceived the UIS-RS as *learning metrics* as a way of rationalising work done across a wide range of different assessment projects. However, in the early stages of development it became clear that the scales could also provide the international development community with tools to monitor learning progress across multiple locations in the context of the United Nations Education 2030 agenda. The reporting scales therefore also became part of the broader work undertaken through a collaboration between ACER-GEM and the UIS as part of the Global Alliance to Monitor Learning (GAML). GAML is an initiative to support national strategies for measuring learning and enable international reporting. Led by the UIS, GAML brings together UN Member States, international technical expertise, and a full range of implementation partners – donors, civil society, UN agencies, and the private sector – to improve learning assessment globally. To ensure the quality and timely delivery of GAML expected outputs, GAML relies on the technical work from thematic Task Forces. This innovative alliance enables stronger links to be forged among all stakeholders, to create collaborative solutions to the challenges of monitoring learning worldwide. In particular, the reporting scales can be used to support monitoring of the United Nations Sustainable Development Goal Number Four: Quality Education (SDG 4). Specifically, to support indicator 4.1.1 of target 4.1:

*Indicator 4.1.1: Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.*

*Target 4.1: By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes.*

The substantive descriptions in the UIS-RS will provide a backbone for interpreting the words ‘reading’ and ‘mathematics’ in Indicator 4.1.1. Almost two-thirds of all developing countries have sought to measure education quality by implementing national assessments or participating in regional or international learning assessment initiatives (Best et al., 2013). However, these assessments vary in approach, method, reliability, validity and comparability. Despite the high level of participation in learning assessments, clearly defined reporting scales and intra- as well as inter-assessment comparability remain limited. This presents particular challenges for measuring progress against the global development goals for learning outcomes of learners.

The learning goals and targets will only have meaning and utility if they are underpinned by empirically derived common scales that accommodate results from a range of different assessment programs. A reporting scale provides a means to assess the

emerging competencies of learners and to explore cognitive growth and trends in growth over time. The development of the UIS-RS allows policy makers, education practitioners and education investors to not only quantify and compare learner proficiency, but also describe it in a meaningful way.

The objective is to develop empirically derived common reporting scales in mathematics and reading that will support national governments to effectively measure and monitor learning outcomes for policy purposes. The UIS-RS describe and quantify learning progressions for reading and mathematics that span learning that typically takes place during primary and secondary schooling. Common reporting scales can be made freely available for all countries to use, in order to align assessment outcomes across diverse contexts. Accommodating results from a range of different assessments of learning outcomes will allow for high-quality data to be yielded that are nationally and internationally consistent. This would lead to a strong focus on improving data use and policy interface, and emphasising peer-to-peer capacity support and learning opportunities. Development of the UIS-RS provides an opportunity to use high-quality methods derived from decades of educational measurement experience to create tools that can support improvement in learning assessment and in education systems around the world.

## **1.1 Understanding the approach**

The UIS-RS can be understood as *learning metrics* and are indicative of a dimension of educational progression (see Turner, 2014b). For example, a developmental scale of reading or mathematical proficiency would be considered a learning metric. Learning metrics comprise two main elements: measures of proficiency located along a scale, and proficiency descriptions associated with locations on the scale. A learning metric is based on the idea that learning is something that builds over time and progresses continuously. The metric is depicted as a line with numerical gradations that quantify how much of the measured variable (e.g. progression in learning to read) is present. It assumes that achievement at a given level of proficiency incorporates the knowledge, skills and understanding in each of the levels below it.

## **1.2 Mapping a learning domain**

Learning metrics are designed to present a potentially infinite range of ability in a domain. This range focuses on knowledge, skills and understanding relevant to the domain as they develop rather than on specific grade- or age-appropriate curriculum objectives. In the present context, an appropriate metric should encapsulate growth from the very early stages of educational development through to middle secondary school. Further, observations of learners can be gathered in a wide range of countries, including some that have severe resource challenges or are in a rapid state of development, even where learners are not engaged in school education. In order to examine and report assessment outcomes consistently, a planned approach to defining performance levels and to associating learners with those levels, is needed. This is both a technical and practical matter of interpreting what it means to be at a level, and has significant consequences for reporting national and international assessment results.

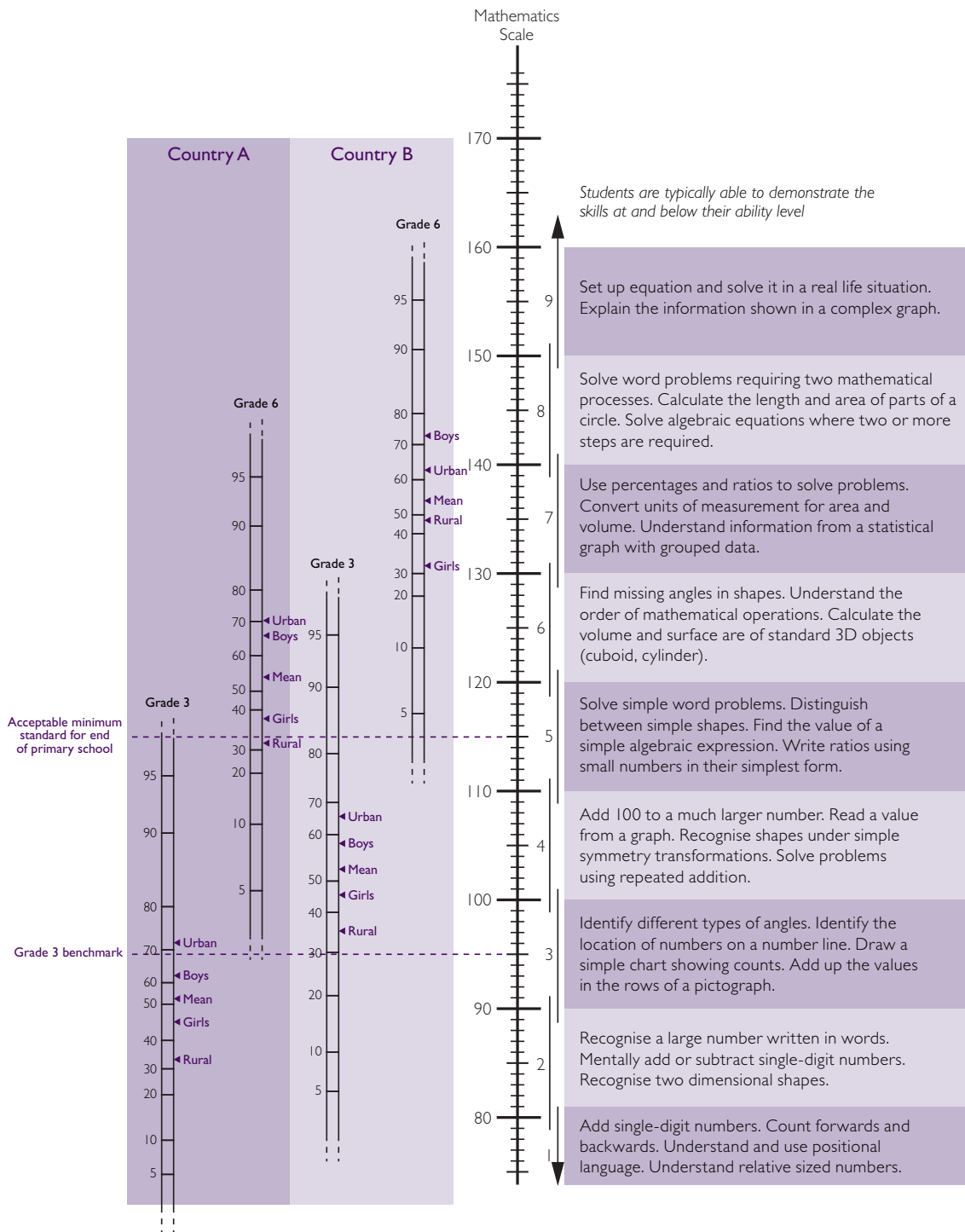
The approach to learning metrics adopted here draws on a rich methodological history that dates back to the work of Benjamin Wright and his collaborators at the University of Chicago in the 1960s. The metrics essentially serve as a roadmap of development through learning in a particular domain. The goal of these methods is to report learner performance not just as numerical scores, but also in terms of content, by describing what learners who achieve a given level on a scale typically know and can do (Masters & Forster, 1996). Typically, when considering a domain, an understanding will emerge of how it materialises at various levels of sophistication. Learning metrics allow us, not only to measure a learner's location on a continuum, but also to identify the difficulty of the measures used to assess ability. This understanding can be tested and informed by empirical data through the application of an Item Response Theory (IRT) measurement model.

The process of developing a learning metric is based on the assumptions of a unidimensional construct (i.e. the domain or skill), and existence of a mathematical function that describes the relationship between item characteristics and learner ability. Item response theory provides this mathematical model (or relationship). The models are probabilistic in nature, estimating the probability of learner success on a specific item given the relative positions of the learner and the item on the scale (Wu & Adams, 2006). To estimate these probabilities, the model uses the interaction between a learner and assessment item to determine the relative chances of success for every instance that a learner encounters an item. An IRT model will estimate the probability of success for every one of these encounters and then use the probabilities to predict scores and overall response patterns for each learner to all the items and to each item for all the learners.

In applying IRT models to assessment data, it is possible to examine how the interaction of learners and items can yield a set of scores that describe locations on the metrics. An output from the IRT analysis is the Wright map (see Wilson, 2005) which provides a visual representation of the item and learner estimates on a single scale, using logits as the scaling unit (an arbitrary unit used to enable locations of the two variables on the same metric). The map displays as two vertical histograms, with the distribution of learner ability on the left, and the distribution of item difficulty on the right. The least difficult items and lowest ability learners are at the bottom of the map, graduating up to the most difficult items and highest ability learners at the top of the map. Through referencing corresponding learner and item locations on the Wright map, the learner is estimated as having a 50% probability of answering the similarly located items correctly. What this means in terms of learner proficiency is that learners whose ability estimate places them at a certain point on the metric would most likely be able to successfully complete items at or below that location, and increasingly more likely to complete items located at progressively lower points on the scale, but would be less likely to be able to complete tasks above that point, and increasingly less likely to complete tasks located at progressively higher points on the scale. The learning metric provides a way of operationalising the fundamental properties of measurement; that there is an order, that the increasing order demands greater skill, and that those skills below that achieved have been mastered (Wright & Masters, 1982). The unidimensionality of the modelled construct, and the relative locations of items and learners along a learning metric, provides crucial information in understanding a domain

and its structure. That is, assessment data interpreted as a Wright map provides a representation of the domain as a progression or scale.

Locations along the learning metric can be described either by numerical scores or substantively (i.e. in terms of learner skills, understanding and competencies) (see Figure 1, from Turner, 2014b).



**Figure 1** Example learning metric for mathematics (from Turner, 2014b)

When the locations are described numerically, they are referred to as *proficiency scores*, which serve to quantify different performance standards for the metric. For example, a score of 115 is a proficiency score. When locations are described substantively, they are referred to as *proficiency descriptions*. A proficiency description from the mathematics learning domain might be 'Learners with a score of 115 (say) on the scale can solve simple word problems, distinguish between simple shapes, find the value of a simple algebraic expression and use numbers to write ratios in their simplest form.' It is not practical to develop a proficiency description for each proficiency score on the numerical scale so proficiency descriptions are usually developed to cover particular segments of the scale. These segments are called levels. The proficiency descriptions for a particular level can then be understood as describing the knowledge and skills of children who attained proficiency scores that are within that particular segment of the scale. For example, again in the case of mathematics, 'learners at Level 5 can solve simple word problems, distinguish between simple shapes...'

Sometimes, a location on the scale is set as a *benchmark*, which is a point on the scale against which comparisons can be made. For example, we might say that a score of 115 (the proficiency score described above) is 'a benchmark for acceptable performance after the completion of primary schooling'. An *indicator* is a quantitative expression that is used to describe the quality, the effectiveness, the equity or the trends of a particular aspect of the education system. It does so through mathematical statements about metrics, proficiency scores and benchmarks. For example, 'the proportion of learners who have achieved a score of at least 115 in mathematics' is an indicator. Further, given the proficiency description of this score, an associated indicator is: 'the proportion of learners who can solve simple word problems, distinguish between simple shapes, find the value of a simple algebraic expression, and use numbers to write ratios in their simplest form'.

Learning metrics have a long history of use. An early published example of a domain laid out and illustrated with items and their characteristics is found in Wright and Stone (1979). Learning metrics have become a common part of the reporting of assessment outcomes in a range of national and international assessment programs. In Australia, many assessment projects have adopted this approach, which include the Test of Reading Comprehension (TORCH) project that originated in Western Australia in 1982 (Mossenson, Hill & Masters, 1987), and the Basic Skills Testing Program in New South Wales (Masters et al., 1990). Internationally, the approach is used to report outcomes of the National Assessment of Educational Progress (NAEP) in the US. Learning metrics are used to report outcomes of the International Association for the Evaluation of Educational Achievement's (IEA) studies such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). The Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) project uses the approach in the reporting of outcomes across reading, mathematical and scientific literacy, as well as problem solving, digital reading, and financial literacy learning domains (see, for example, Turner, 2002). The learning metric language and approach is also currently being adopted for a new South-East Asian regional assessment program (the South-East Asian Primary Learning Metric – SEA-PLM)).

The specific methodologies used for development of reporting scales may vary across different studies, but they all stem from the same learning metrics principles as outlined above.

### **1.3 Adopting a vertical scale**

The methods outlined in the previous section can be used to develop scales for SDG 4 reporting. To do so, a simplifying assumption is needed, to the effect that the purpose of the scales is not to explain all of the reading and mathematics domains, but rather to describe a common scale that is sufficient to represent the construct across school years, and across different cultural contexts. That is, the scale balances representativeness of the construct, many contexts, and many ability levels. We assume unidimensionality as an approximation in order to achieve scales that are suitable for the purpose of SDG 4 reporting. The higher the level at which individuals or groups are measured against the learning metric, the greater their learning progress, and the better are their outcomes against indicators set to operationalise SDG 4 targets.

There are three arguments to support the conceptualisation of the UIS-RS as a vertical scale:

- vertical scales are fit for the purpose of SDG reporting
- vertical scales allow the capture of the diverse variation in abilities observed worldwide
- vertical scales align conceptually with the goal of UIS-RS to provide a common resource to enhance consistency in reporting of assessment results.

Vertical scales are being used, with varying degrees of confidence, to support decisions by policymakers, educators and students that can only be achieved through their use, due to the common metrics, to allow the learning progress of all students to be monitored by reference to the same scale and tracked over time.

The process for following an IRT approach to vertical scales involves vertical equating or linking (Slinde & Linn, 1977) in which common test items are administered on the tests for different year levels. The results are scaled so that scale scores are comparable across different grade levels. Vertical scaling and linking of test scores has been most successful when test design and item selection within and across grade levels are managed well so sufficient overlap of items in adjacent test levels enable stable links (Ferrara, Johnson, & Chen, 2005).

Indicator 4.1.1 will need to be applied in a sufficiently standardised form to permit reasonable cross-country consistency. The approach presented here sets out to develop a vertical scale that spans a large range of proficiency levels rather than developing three scales, one for each of three benchmarks. The adopted approach is consistent with the SDGs in that they refer to a single progression of a domain, with minimum proficiency determined. A vertical scale provides a consistent approach to reporting across countries and educational levels. In addition to being able to identify what students are estimated to be able to do at points along the progression, outcomes can be reported consistently across grade levels, subgroups within country, across

countries, and across time, on the same scale. Most importantly, a learning growth trajectory is defined.

Within this approach, the focus is on continued student learning across each domain, not just in relation to minimum proficiency standards. As CTB-McGraw Hill identify in their TerraNova test battery, a vertical scale 'can be viewed as a developmental continuum ... scale scores are units of a single, equal-interval scale applied across all levels of [the test] regardless of grade or time of testing' (CTB-McGraw Hill, 2001, p.322). There are three benchmarks for SDG 4.1.1: (a) grades 2/3; (b) the end of primary; and (c) the end of lower secondary. The grade-based benchmarks are fuzzy in practice, due to the variations between countries in the age of school entry, and therefore the age of children when they reach the benchmark points. Globally, there will be at least two to four years of variation in student ages when children have completed two or three years of schooling with children typically entering school between five and seven years of age (UNESCO, 2012). In relation to UIS-RS reporting, it is likely that a wide range of performances will be observed across countries and a scale suitable for the end of primary or the end of secondary would have to span a very wide range to be of value to countries. Further, countries will have different interpretations of grade levels. A vertical scale allows continuity and growth in the domain to be considered separately from grade levels; instead the growth is understood as changes in magnitude of student outcomes.

Consider an alternative approach where there would be three scales, one for each of Grade 2/3, end of primary, and end of secondary. The starting point for students varies across scales, which leads to different interpretations in growth magnitudes and a result brings into question the validity of the scales. Ages may vary widely at each reporting point and abilities may vary widely within-country and between-country. There may be a large amount of overlap between the lower end of the distribution of higher performing countries and the upper end of the distribution of lower performing countries. For example, a report commissioned by the Grattan Institute (Goss & Chisholm, 2016) identifies that students from different SES groups in Australia can differ by 2.5 years by Grade 9, resulting in a wide range of performance from Grade 9 students. This gives an indication of the kind of variance that may occur in different educational settings.

An additional issue with having separate scales is that they do not allow for across-grade comparison and for establishing equivalence between the benchmarks. If a scenario occurs where subgroups of students do not reach the Grade 2/3 minimum proficiency but in subsequent years do meet the end of primary minimum proficiency, there is no easy way to compare non-linked scales. An additional difficulty is that grade levels are in nature ordinal not interval. Grades are not necessarily spaced at equal distances apart, making it difficult to interpret equal growth across grade levels along a vertical scale (Lissitz & Huynh, 2003). In contrast, vertical scales represent an underlying construct that does not depend on grade levels and hence supports an interval approach where the construct is represented with equally spaced levels (Briggs, 2013).

The approach of having a single scale presents as much more efficient and may mitigate some of these issues. Developing a single vertical scale is a widely used

approach across many national, regional and international assessment programs. These scales are used within assessment programs to make inferences about subgroups of populations that lead to supporting decisions about policy, funding, and curriculum among others.

### **Vertical scales used by other countries**

In Australia, NAPLAN uses a single common scale that spans Years 3, 5, 7, and 9 that allows student progress to be monitored on the same domain throughout schooling years. This also allows for the achievement of the most proficient students to be monitored at the same time as the least proficient, who have potentially still to reach the agreed national minimum standard. The NAPLAN tests are designed so that results between students in different year levels and students taking the test in different years can be compared. As the vertical scale for each domain is standardised, the scales for all the domains are very similar in length. Each scale has the same number of proficiency bands and the same nine cut-points on the transformed scales. The performance distribution at each year level could span approximately six bands. A student whose proficiency fell within a specific band would be expected to score at least 50% correct on a test made up of items that fell into that band (NAPLAN, 2017).

The Pacific Islands Literacy and Numeracy Assessment (PILNA) is an example of a regional assessment that uses vertical scaling. The single uniform metric, as it is referred to in this instance, is used to report student outcomes across the Pacific region. The use of a consistent vertical scale for the region allows national results to be compared to average achievement across the region (PILNA, 2017). Two vertical scales were developed, one for literacy and one for numeracy, each identifying eight levels of proficiencies with PILNA score ranges highlighted for each level. In addition, grade level benchmarks are identified on the scale, which provides regional benchmarks against which countries can compare their students' performance. PILNA makes strong recommendations to regional and national education leaders to adopt the use of a regional uniform metric as a way to track progress and trends in student learning outcomes.

In the United States, the No Child Left Behind Act of 2001 (NCLB) brought with it the requirement to track cohort growth and achievement gaps across grade levels. The Act has increased the need for states to administer annual reading and mathematics assessments from Grades 3 through 8 and determine whether students within schools are achieving adequate yearly progress. Huynh and Schneider (2005) state that vertical scales may facilitate school accountability and NCLB obligations in some situations. Many standardised tests in the US had already adopted vertical scales such as the Iowa Tests of Basic Skills (Hieronymus & Hoover, 1986), TerraNova (CTB-McGraw Hill, 1997), the Stanford Achievement Test (Harcourt Educational Measurement, 2004); and many state assessment programs are adopting vertical scales for reading and mathematics including Mississippi (Tomkowicz & Schaeffer, 2002) and Colorado (Colorado Department of Education, 2003).



For many years, the National Assessment of Educational Progress (NAEP) used vertical scales for score reporting and achievement level setting. NAEP reported student outcomes in Grades 4, 8 and 12 using vertical, across-grade scales in reading and mathematics for many testing cycles. However, in 1991 the decision was made to move away from vertical scales to within-grade scales, in part because the NAEP frameworks specify different content at grade levels (Huyhn & Schneider, 2005). Despite this, NAEP continues to report reading and mathematics on the across-grade scales to ensure continuity in trend reports, evidencing their value for this purpose. Much of the discussion about cross-grade scales in NAEP relates to validity. If a well-constructed cross-grade single scale is developed then interpretations can be made about the NAEP scale points in relation to one year's growth, or comparisons about score equivalencies across grades. However, the argument in the existing literature is that if the scale is not well-constructed, or representative of a single construct, then those aforementioned interpretations are invalid. Further, Patz (2004) indicates that there are increasingly sophisticated psychometric techniques that allow for the validity of the NAEP cross-grade unitary scale to be improved. As Thissen (2012, p. 12) suggests:

*if responses to items, or a series of items, indicated progress through a sequence that had been established to be a learning progression, then that would establish a basis for across-grade score comparability. If items representing learning progressions made up a sufficiently large proportion of the assessment, scores could be interpreted to represent positions in those sequences, and could, hypothetically, be comparable for students in grades 4 and 8.*

Some limitations of a vertical scale approach are acknowledged. There are criticisms of all measurement models. Psychometric misfit is not limited to vertical scales in the same sense that all statistical estimators are subject to errors. As Kreiner and Cristensen (2014) highlight 'statistical models never fit perfectly; and tests of fit in large sample studies such as PISA will always ultimately yield evidence against the model.' Large scale assessments adopt a similar approach to vertical scaling despite noting limitations in psychometric robustness. For example, an analysis by Adams, Bereznier and Jakubowski (2010) identified strong evidence of Differential Item Functioning (DIF) between countries, suggesting the items may be measuring different abilities for different countries. However, once those items indicating DIF were removed, there was little impact on country ranking presenting invariance between countries where only a few countries systematically went up or down in their ranking position. This is an example that demonstrates that while statistical models never fit perfectly, especially when dealing with large sample sizes, the assessment is still fit for purpose, in this case for country ranking. In the same sense, vertical scaling is still considered fit for purpose in respect to SDG reporting of indicator 4.1.1 in that the intention is to report proportions of minimum proficiency, not individual student ability. Further, the practice of vertical scaling clearly has its merits, and the limitations in validity will only be improved to the extent that criteria for creating and testing vertical scales can be refined. The UIS-RS presented in this report will go through a validation process as part of the next phase of work.

## 1.4 Development process

This report outlines Phase I in the development of the UIS-RS. The purpose of this phase was to develop a set of draft reading and mathematics reporting scales from the earliest available developmental levels to at least early secondary school. Phase 1 was undertaken without the collection of new data from learners – that is, it drew upon pre-existing performance data from a variety of assessment programs.

The steps of the development process are presented across the first four sections of this report as follows:

- **Section 1: Conceptual frameworks** – an outline of the theoretical frameworks and conceptual understandings of reading and mathematics as domains to inform the UIS-RS
- **Section 2: Identifying conceptual demand** – the process of qualitatively analysing the conceptual demand of items in a wide range of reading and mathematics assessment programs conducted by domain-based education specialists
- **Section 3: Constructing a single scale** – the technical approaches taken to derive empirical estimates of the difficulty of assessment items within and across assessments and subsequent validation and synthesis with additional assessment programs
- **Section 4: Scale drafting** – synthesising the previous steps with further input from domain-based specialists to write descriptions of the scales and devise levels

The fifth section of the report discusses the intended online presentation of the reporting scales. The sixth section notes limitations with the approach to building common reporting scales and how these will be addressed. The conclusion to this report discusses the outcomes resulting from the work undertaken to date, along with the details of Phase II of development and implementation of the reporting scales.

## 2 Conceptual frameworks

Reading and mathematics were selected as the two domains in which learner progress would provide a reliable cross-country indicator of the quality of education. These domains are universally important and there is sufficient technology to measure them. Work on the frameworks commenced with establishing a broad conceptual understanding of reading and mathematics, based on a synthesis of literature, and how these domains are typically organised in existing curricula and assessments. The labels 'reading' and 'mathematics' were adopted to signal that the broadest possible range of levels of reading comprehension and mathematical competence were incorporated into the development work, as well as to indicate they relate to standard areas of school curricula. However, the UIS-RS are not curriculum bound, rather, they adopt a 'literacy' orientation (see Turner 2014a). Education researchers and practitioners describe a learning domain in terms of literacy to emphasise the fact that the domain has dimensions that extend beyond any particular curriculum or syllabus. In a learning domain with a literacy orientation, the focus is on applying the domain's facts, skills and procedures to support creativity and inventiveness, to solve novel problems and to deal with the kinds of challenges that life presents both inside and outside the classroom.

The conceptual frameworks were used throughout the process of developing the UIS-RS to guide the item review work. The frameworks were formulated to take into consideration existing research, curriculum and assessment programs.

### 2.1 Conceptual framework for reading

The ability to read and understand text is fundamental for an individual's education, personal enrichment, and participation in society. Reading requires a broad variety of perceptual, linguistic, and cognitive skills to extract meaning from visually presented material, most commonly written text on paper, in books, and increasingly on the screens of digital devices. Because reading and understanding text involves so many skills that develop during childhood, usually at the same time that children receive formal instruction in schools, proficient reading takes years to develop and involves both understanding a language (comprehension) and understanding the symbolic representation of that language as written text (constrained skills). The conventional sequences of learning to read reflect the effects of maturation and instruction. They vary widely across children, languages, and contexts, but there are patterns and milestones in reading development that are important as instructional objectives and assessment outcomes, and they reveal successive accomplishments in reading speed, accuracy and understanding (Paris, 2011; National Reading Panel, 2000).

Because proficient reading requires the development of a variety of skills over many years, it is a central focus in the first few years of formal schooling for children around the world (Clay, 1979). Literacy curricula, classroom assessments, and teacher instruction are all designed to foster the development and integration of reading skills into a smooth, rapid, and accurate process of constructing meaning from texts (Adams, 1990). Although there are many different curricula and instructional methods used to teach reading, they all reflect similarities in the underlying developmental

progressions from simple to complex skills applied to familiar then progressively more complex texts. For example, the literacy curriculum frameworks in Australia (Custance, Hamilton, & Payne, 2013) and New Zealand (NZ Ministry of Education, 2010) are built on developmental continua of skills applied to increasingly challenging texts that are captured in a series of learning progressions across primary and secondary grades. The English Language Arts Standards for the Common Core State Standards (2017) in the United States are also based on learning progressions.

The table in Appendix 1 presents a framework developed by reading curriculum and assessment researchers from ACER. The framework groups reading concepts and knowledge into broad domains and provides some commentary to clarify scope and typical learning trajectories within each area, and highlights some considerations about the content. Within the domain of reading, four strands were identified: Retrieving, Interpreting and Reflecting (all related to comprehension) and Constrained skills.

### **2.1.1 Comprehension**

Comprehension is one part of learning how to read. Comprehension skills are unconstrained in their development, and that means that they include a large variety of skills that continue to develop before, during, and after formal schooling. Language comprehension skills, including vocabulary, develop initially as part of listening comprehension and are transferred to reading when language becomes associated with symbolic representations in print. When fast, accurate reading fluency is achieved, more cognitive resources and attention can be applied to understanding texts of increasing complexity and subtlety that are read independently (Stanovich, 2000). Reading comprehension involves cognitive syntheses of ideas in text with previous knowledge and ideas, sometimes called schemata, to achieve an understanding of the text and the situation described or implied by the text (Anderson & Pearson, 1984; Kintsch, 1998). Reading comprehension depends on memory skills, vocabulary, conceptual development, and a variety of cognitive and metacognitive skills to help monitor and regulate understanding.

The reading conceptual framework was substantially modelled on OECD's Programme for International Student Assessment (PISA; see OECD, 2016) and IEA's Progress in International Reading Literacy Study (PIRLS). These conceptual frameworks have broad international acceptance as well as extensive, freely available documentation about the detail of the constructs. These frameworks were developed by international reading specialists and many different countries now participate in the PISA and PIRLS assessment programs to report reading achievement. Based on these frameworks, the reading domain was defined as consisting of three strands: Retrieving, Interpreting, and Reflecting. Beginning readers demonstrate simple versions of these skills in reading simple texts. As their reading skills develop, they demonstrate more sophisticated skills applied to more complex texts. Eventually, advanced readers can locate information, interpret and reflect on texts with an extensive vocabulary of less familiar words, complex and subtle ideas, and unusual and varied styles and structures.

## Retrieving

Retrieving information refers to identifying and locating small, discrete, and explicit pieces of stated information in a text. Skimming and scanning text to find the relevant information are typical reading skills applied to retrieve information. These searches of text are not necessarily linear or thorough or accurate so they may differ from usual reading practices that focus on constructing deeper meanings from text. However, retrieval skills develop from simple to complex over time and become faster and more accurate with practice.

## Interpreting

Interpreting text includes a variety of skills used to construct meaning from the explicit and implicit meanings of words in order to understand the meaning of the text as a coherent whole. Interpreting may focus on understanding sections of the text, links across the text or the overall meaning of the text. Learners need to construct meaning, such as using background knowledge, and to interpret contextual clues and implied meaning. Interpreting includes skills used when retelling, explaining, summarising, synthesising, identifying main ideas and details, making comparisons, drawing conclusions, generalising and analysing information, classifying, categorising and making predictions within the context of the text with the evidence the writer has provided to support interpretations about its meaning. Thus, interpreting text includes construction of text and situation models in Kintsch's (1998) model as well as integrating the constructed meanings.

## Reflecting

Reflecting on the form or content of the text refers to critical analysis of the text and taking different perspectives on the text and its meaning. It is similar to the role of 'text critic' emphasised in the four resources model proposed by Freebody (1992). The analysis and critique of the text may be applied to any aspect of the text's construction (e.g. format, tone, genre or style), or it may be applied to the content of the ideas. Reflecting typically requires bringing substantial background knowledge to the text either in terms of technical knowledge, such as knowledge of genre structures and literary devices, familiarity with an external frame of reference, or criteria that are used to reflect on and evaluate the content. Reflections on text involve metacognitive analyses of text features, purposes and structures and can require complex skills.

### 2.1.2 Constrained skills

The other part of learning to read is mastering the decoding skills required to convert written symbols into words in a particular language. These skills are constrained in scope and time of acquisition because they involve smaller numbers of discrete things to be learned (e.g. the letters of the alphabet) and they are usually mastered completely in a few years (Paris, 2005). Constrained skills are fundamental and need to be fully mastered in order to become a proficient reader. It is important to recognise that comprehending words and language begins in infancy and continues throughout adulthood whereas decoding skills are generally learned during early reading

instruction. Thus, the learning progressions for decoding skills are more constrained in scope and duration than learning progressions for comprehension skills even though they overlap during the early years of learning to read.

The constrained skills strand concerns the development of fluency, as well as basic concepts of print. Words must be accurately, and quickly, recognised, before they can be processed for meaning. Fluency also includes the use of appropriate prosody, or proper phrasing and expression, in reading aloud which suggests some recognition of meaning. However, while fluency is necessary for comprehension it is not a proxy for comprehension. Direct evidence of comprehension demonstrated by students responding to questions about texts they have read is the best way to measure comprehension.

The UIS reading scale only describes the development of constrained skills up to Level 7. By this level, readers have sufficient fluency to support a small range of comprehension skills, applied to a range of simple, everyday texts, as described in the retrieving, interpreting and reflecting strands. Level 7 is roughly equivalent to the PIRLS low benchmark.

Fluency skills, including attention to prosody, definitely continue to develop well beyond Level 7 of the UIS reading scale, but fluency skills are no longer described above this level. Comprehension also requires knowledge of the vocabulary, morphology, grammar and syntax of the language, background knowledge of the content and knowledge of text structures. From Level 7, the evidence of reading proficiency, is described in terms of the comprehension skills that are demonstrated. If students can demonstrate the relevant comprehension skills, then their fluency skills, are sufficient for that task in that context, assuming the task is a valid measure of reading.

### Language groups

There are large differences in the orthographic complexity of languages, the consequent time it typically takes students to become sufficiently fluent and the relative importance of different features in the orthography in learning to become fluent. For example, the names and sounds for small symbol sets of 20 to 40 letters in Latin-derived scripts are typically learned by the end of the first year of school (Seymour, 2005, Share, 2008). Large symbol sets of the Indian alpha-syllabaries with 200 to 500 symbols are not fully mastered even by grades 3 and 4 (Nag & Snowling, 2013).

The constrained skills strand identifies four broad groups of languages:

- simple, transparent, alphabetic and alpha-syllabic languages (e.g. Spanish, Indonesian, Korean),
- complex, transparent, alpha-syllabic languages with large symbol sets (e.g. Khmer, Lao, Kannada)
- opaque, alphabetic and alpha-syllabic languages (e.g. English, Dutch)
- character-based languages (e.g. Chinese)

Progress in the simple, transparent alphabetic and alpha-syllabic languages is taken as the reference point with caveats for the remaining three language groups. The caveats acknowledge that learning all of the symbol to sound relationships and placement rules in a large symbol set or character-based language, or all of the rules concerning ambiguities and exceptions in an opaque language, will continue over several levels and that texts at each level are consequently restricted to words that can be decoded, or recognised, given what has been taught.

The intention is for countries to identify which group best describes their language. There is also scope for countries that have languages with large symbol sets or characters to adjust the number of letters/symbols or characters to be recognised over each of the levels. Descriptions of key, sub-syllabic phonological awareness skills that are relevant to a particular language can also be added, if required.

Some languages may straddle more than one group, such as Japanese which combines simple, transparent Kana and Chinese characters (Kanji). While Kana decoding is mastered quickly, by level 4, Chinese character learning will continue across the levels, slowing overall progress. The texts that students are able to read at any point, will be limited to the Chinese characters they have learned.

### **Comparing progress across languages**

In order to demonstrate the comprehension skills described at Level 6 and above, readers, in any language, must have sufficient fluency to support comprehension in these contexts. Regardless of the time it takes, mastery of symbol-sound knowledge and placement rules (apart from rare instances in large symbol sets) is essential. Rapid, accurate word recognition is required in all languages.

It is suggested that Level 6 is a minimum benchmark where evidence of comprehension of a simple, connected paragraph of text can be collected. The grade level at which it is reasonable to set this expectation depends on the orthographic complexity of the language and the context in which the language is taught including whether this is typically a first, second or third language, and the state of the education system. For example, Level 6 should be achieved in less than a year for a transparent language in a well-resourced context, where it is the first language of most students, but it may take four or five years to achieve this level for a language with a large symbol set, limited resources and many students with other first languages.

Levels 1 to 5 of the UIS scale provide opportunities for countries with complex or opaque orthographies to identify their progress in the mastery of the constrained skills at earlier stages, especially in contexts where Level 6 may take years for most students to reach.

### **Constrained skill sub-strands**

The UIS scale descriptions of these skills are limited to a few key features likely to apply across languages. Countries are invited to customise the descriptions of the constrained skills across Levels 1 to 7 by adding details that are relevant for particular languages. Some languages, such as character-based languages, may also justify

extending the levels above 7 to identify features of the language or characters that are only introduced after students have developed reading proficiency skills described at higher levels.

The constrained skills strand identifies what seems to be some universal indicators of progress in terms of the main components of the constrained skills strand:

- basic concepts of print (e.g., directionality of print, punctuation);
- phonological awareness (e.g., hearing and identifying distinct sounds in words);
- symbol-sound knowledge (e.g., linking written symbols to spoken sounds); and
- fluency (e.g., reading text aloud accurately with appropriate speed – prosody may be included).

Concepts of print refers to basic knowledge about how print works in a particular language. This includes differentiating print from drawings, or other symbols, recognising the difference between a letter and a word, knowing where to start reading and which direction to read written texts. It also includes basic knowledge about how books or texts are constructed and the role and purpose of a title or heading, captions and illustrations. Elements of punctuation such as knowing the name and purpose of full stops, question marks, exclamation marks, or other punctuation marks, in terms of their impact on the meaning of the text are also components of concepts of print.

Phonological awareness is the ability to identify sounds in spoken language. This skill begins with discriminations among larger chunks of sound, such as being able to differentiate words in a spoken sentence and, later, distinguish syllables within words. Phonemic awareness is an aspect of phonological awareness that is exclusively focused on phonemes and the aural identification of all the phonemes in words. Phonemic awareness skills may also include manipulation of phonemes in words such as replacing a letter to change the word.

Symbol-sound knowledge concerns the letter-sound relationships in the language. It includes recognition of all the letters of the alphabet in a range of fonts, including diacritics, tonal markers and upper case and lower case, where relevant, and knowledge of the sounds of all of the letters and letter combinations with unique sounds. In English, irregular letter-sound relationships also have to be learned.

Fluency is the ability to read aloud with sufficient speed and a very high level of accuracy in order to support comprehension of the text. Reading over 100 words with at least 98 per cent accuracy is generally considered sufficient to support comprehension; accuracy levels that fall below 95 per cent are likely to limit comprehension. Fluency always includes speed and accuracy; some measures also include prosody. Prosody refers to grouping words in meaningful clusters, reading at an appropriately varied pace and volume, with expression, and observing and using punctuation to support meaning.

## **2.2 Conceptual framework for mathematics**

Mathematics has no generally accepted definition. For the purposes of the UIS-RS, it is defined as a hierarchically structured set of inter-related concepts and procedures that have been devised by humans over millennia to make sense of the world and to



facilitate active engagement with that world. Fundamental mathematical concepts have been identified and articulated, and tools and procedures devised to describe and use those concepts in an increasingly diverse range of contexts.

Learning mathematics is conceived as a matrix of connected, developmental and cognitive processes through which individuals construct meaning from their experience of their world, and develop and use those mathematical concepts and related procedural knowledge. Mathematical competence grows with development of conceptual understanding, acquisition of procedural knowledge and skills, and building a range of mathematical competencies that are critical to the activation and use of mathematical understandings, knowledge and skills in different contexts.

A typical learning sequence would see learners building and displaying understanding of concepts through manipulating physical objects, recognising patterns that enable the building of mathematical rules that summarise the underlying phenomena, and learning procedures and skills needed to use the understandings as they develop. The ability to see patterns and commonalities across very different contexts, and to develop increasingly abstract conceptions of those patterns is a key feature of mathematical competence, and using mathematics across an increasing range of different contexts is a prime indicator of progress in mathematical learning.

The conceptual framework around which the mathematics reporting scale has been built seeks to identify mathematics competencies and related skills. It also seeks to describe them in an order that matches the ways in which learning typically occurs and that facilitates a logically ordered process of building mathematical knowledge progressively to deeper and broader levels as learning advances. These mathematical competencies include communication, interpretation and construction of mathematical representations, strategic thinking, mathematising (which includes formulating situations in mathematical terms, and interpreting mathematical results in terms of the situations to which they relate), reasoning and argument, and using symbols, operations and formal language. The underlying concept for a literacy orientation in mathematics is to forge the connections between the conceptual and procedural knowledge that constitute the basis of the mathematics domain and the real-life situations in which mathematical knowledge can be used. The mathematical competencies mentioned above are fundamental to activating those connections.

Particular mathematical concepts and elements of mathematical knowledge can often be seen in apparently different contexts, which potentially gives rise to different ways of organising the knowledge domain. For example, a fundamental concept such as 'magnitude' can be expressed through counting and aspects of number. It can also be expressed through spatial phenomena, where qualities like length, area, and volume each express variation in magnitude in different ways. Magnitude is also relevant to considerations of aggregations of data and statistical analysis, and to probabilistic situations where variation in the frequency and likelihood of events is observed.

The table presented in Appendix 2 presents a mathematics framework that was developed as the first step in the process presented here. The framework groups mathematical concepts and knowledge into broad categories and into subcategories, provides commentary intended to clarify scope and typical learning trajectories within each knowledge area, and suggests the kinds of questions that could be asked to

provide indicators of levels of understanding. The intention was to capture concepts, skills, processes and applications that span developmental stages.

In devising a mathematics reporting scale, therefore, it was useful to structure the work around strands of mathematical knowledge that reflect common ways of organising the domain. The mathematics conceptual framework takes into account material in alternative frameworks that underpins existing assessment programs, to form a description of early growth in the development of mathematical concepts, knowledge and skills. Table 1 shows the PISA and TIMSS mathematical literacy content categories.

**TABLE 1** PISA and TIMSS mathematical content categories

PISA mathematical literacy content categories (OECD, 2015)	TIMSS Grade 8 mathematical content domains (TIMSS, 2015)
<ul style="list-style-type: none"> <li>• Quantity</li> <li>• Uncertainty and data</li> <li>• Change and relationships</li> <li>• Space and shape</li> </ul>	<ul style="list-style-type: none"> <li>• Number</li> <li>• Algebra</li> <li>• Geometry</li> <li>• Data and chance</li> </ul>

An elaborated set of fundamental mathematical capabilities, or competencies, has also been described by Turner, Blum and Niss (2015). As widely known and used assessment frameworks, a similar structure for defining the domains was adopted for the UIS-RS. Despite slight differences in the language, the substance of domain definition is similar across frameworks therefore, the definitions were synthesised to describe mathematical proficiency in the reporting scale. This work led to the identification of three strands within the mathematics domain.

- The Number and Algebra strand describes the conceptual understanding of quantity and number relationships. It involves the understanding of and ability to use mathematical expressions, notations and arithmetic operations.
- The Measurement and Geometry strand describes the knowledge, understanding and skills needed to work with measurable variables; to manipulate geometric shapes and objects and describe their properties; and to locate and explore objects and navigate in two- and three-dimensional space.
- The Data and Probability strand describes the knowledge, understanding and skills needed to record, retrieve, interpret and use data; calculate and use statistics to represent and explore data; design and evaluate surveys, and sampling methods; describe chance events, and quantify and analyse the outcomes from probability experiments.

## 3 Identifying conceptual demand

Once frameworks from which to define and structure reading and mathematics were drafted (see section 1 above), items from existing international, regional and national assessment programs were analysed. This allowed the cognitive and learning demands of these items to be identified that were used to operationalise the two domains in the different assessment programs. This was an initial scoping exercise using the judgment of domain-based specialists, which was intended to lay the foundation for conceptualising the scales of growing proficiency in each domain.

### 3.1 Data sources

To analyse the conceptual demand of existing assessment items and to build a map of growing proficiency, assessment programs were selected with the view to cover learning from foundation to middle secondary, and to represent a range of item difficulties, knowledge, skills and abilities, and in different contexts (e.g. high, middle and low-income countries).

If items were not already accessible, permission was sought for this information. The owners of those items are gratefully acknowledged for their willingness to permit the use of material for this purpose. A total of 512 reading items and 533 mathematics items were sourced from the assessment programs listed in Table 2.

**TABLE 2** Source of items used in the conceptual demand analysis

Program Name	Full Name	Target Country	Age Group or School Year Level	Reading	Maths
ASER	Annual Status of Education Report	India	5–16 years	✓	✓
EGMA	Early Grade Mathematics Assessment	Developing countries	Grades 1–3		✓
EGRA	Early Grade Reading Assessment	Developing countries	Grades 1–3	✓	
LLANS	Longitudinal Literacy and Numeracy Study	Australia	Prep, Grades 1 and 2	✓	✓
MTEG	Monitoring Trends in Educational Growth	Afghanistan	Grade 6	✓	✓
OLAY NORTHERN TERRITORY	Online Literacy Northern Territory	Northern Territory (Australia)	6 years (Grade 1)	✓	✓
PILNA	Pacific Islands Literacy and Numeracy Assessment	Pacific Islands	Grade 4 or 6	✓	✓
PIRLS	Progress in International Reading Literacy Study	About 50 countries	Grade 4	✓	
PISA	Programme for International Student Assessment	About 70 countries	15 years	✓	✓
SACMEQ	The Southern and Eastern Africa Consortium for Monitoring Educational Quality	15 African countries	Grade 6	✓	✓
SISTA	Solomon Islands Standardised Tests of Achievement	Solomon Islands	Year 4 and Year 6	✓	✓
TIMSS	Trends in International Mathematics and Science Study	About 50 countries	Grade 4 and Grade 8		✓
UWEZO – KENYA	‘Capability’ (trans.)	Kenya	6-16 years	✓	
UWEZO – TANZANIA	‘We have the power’ (trans.)	Tanzania	7-16 years	✓	
UWEZO – UGANDA	‘We have the power’ (trans.)	Uganda	6-16 years	✓	✓

## **3.2 Analysis of item conceptual demands**

The items for each assessment program were analysed by reading and mathematics curriculum and assessment researchers from ACER who have worked extensively within one of the respective domain areas. These specialists were tasked with identifying the cognitive and learning demands of each item within an assessment. Where available, the ordering of items on existing empirical scales was considered. This analysis enabled the specialists to describe the skills addressed by each item and to locate where the items would typically be situated in a curriculum or teaching program that progressively builds on earlier skills.

The specialists identified where an item might be located on the scale, based on the identified skills and item difficulty. Next, they identified the difficulty of the task, for example a very simple arithmetic task would fall at a low place on the scale as the content is introduced early and the task demands are simple. A multi-step problem might also include a very simple arithmetic calculation but the difficulty of the task is considerably greater because of the demands of reading the problem, identifying the steps and completing them all correctly. Similarly, identifying the difficulty of a reading task considers the complexity of the text. Specialists must have reasonable expectations about the level of skill learners require to access the text as well as the difficulty of the reading task. It is possible to ask difficult, abstract questions of an easy text, which makes it more difficult to answer than an easy question about the same text. In the same way, it is possible to answer a very easy, superficial question about a complex text which makes the task easy, but is less revealing of the extent of comprehension. Specialists needed to balance all of these considerations when making judgements about the cognitive and learning demands of items so that they can be placed in a conceptually relevant order.

## **3.3 Central observations**

The analysis of item demands for each assessment provided important material to inform the subsequent development of the mathematics and reading scales. In particular it helped to provide broad ideas about growth, and conceptualisation of the scale.

### **3.3.1 Broad ideas about growth**

The conceptual demand analysis also led to the broad ideas for growth in each domain. In the case of reading, the items indicated a progression from low to highly skilled, critical reading. The progression of comprehension skills is broadly similar across languages and moves from the initial realisation that words and texts convey meaning, to the capacity to interpret short written texts presenting familiar ideas, and later to the capacity to reflect critically on written texts with layers of subtle meaning with unfamiliar ideas and rich uses of vocabulary. In alphabetic languages, the skills required to turn written symbols efficiently into spoken words progress from recognising sounds within words and linking sounds to letters and letter clusters to recognising words and being able to segment words into sounds and blend sounds in words. Fluency develops from initial slow and deliberate decoding to being able to read aloud quickly with a very

high level of accuracy. The finer detail of the sequence in which these skills develop, and the time it takes to master them, vary among languages and educational contexts. Critical reasoning skills are required to interpret, reflect on and evaluate the content and form of the texts. The more difficult items mainly call on higher order thinking skills that require learners to question the text and the author's intentions, find evidence to justify different perspectives and interpretations, identify logical flaws and critically evaluate most aspects of the text.

Similarly, the mathematical item analysis led to broad ideas for growth in competence. In particular, growth in development of conceptual understanding, acquisition of procedural knowledge and skills, and building a range of mathematical competencies that are critical to the activation of mathematical understandings, knowledge and skills in different contexts. A typical learning sequence would see learners:

- building and displaying understanding of concepts through manipulating physical objects
- recognising patterns that enable the identification and building of mathematical patterns and rules that summarise the underlying phenomena
- learning procedures and skills needed to use the understandings as they develop.

### **3.3.2 Conceptualisation of the scale**

One issue raised during item analysis for the reading scale was the uncertainty of whether constrained and unconstrained skills (Paris, 2005) could be placed on the same scale. 'Constrained skills' of reading is a term used here to refer to those skills that underpin reading development that are learned quickly and mastered entirely. Concepts of print, phonemic awareness and reading fluency, among others, are included in this category. 'Unconstrained skills' are those that continue to develop and have no clear ceiling, in particular vocabulary and comprehension. There has been a tendency to create one scale containing both constrained and unconstrained skills. However, Paris (2005) suggests that this approach is problematic because the constrained skills are not equally distributed on a scale and can be unstable over time. A review of the items during the process outlined in this section, indicated that there is a larger quantity of items that focus on constrained skills in international assessment programs. This could be attributed to the belief that mastery of constrained skills is predictive of later reading competencies. A common reporting scale for reading should include a sufficient range from foundational concepts to focusing on whether children can understand what they read or have read to them. Therefore, a decision was made to include both constrained and unconstrained skills on the single reporting scale for reading. A reading scale that includes unconstrained skills will promote a shift in focus for assessment programs to include more unconstrained skills and will provide valuable information at both system-wide and school-based levels.

Further analysis during the next phase of this program of work<sup>1</sup> will test the feasibility of including constrained and unconstrained skills on the single common scale.

### 3.3.3 Language considerations

Examining the conceptual demand also raised the implications of some cross-country and language considerations for a single common scale. Not all constrained skills on a common scale would be applicable or appropriate to every language, the differences in language structure makes comparison of languages on a common scale problematic.

These problems can be largely overcome in relation to comprehension by ensuring that there are appropriate consultation and translation processes in place. In relation to sound and word structure, these problems are less easily resolved. The Early Grade Reading Assessment Toolkit argues against the comparison of the EGRA assessment across languages because of 'differences in language structure and rate of acquisition' (RTI international, 2005). Even within languages it can be difficult to produce a precise common scale for constrained skills of reading. This is because there is not always a clear progression along the various constrained skills that underpin reading, so it is difficult to be specific about which skills are learned before others. For example, identifying capital letters may be learned before, after, or at a similar time to recognising initial phonemes.

The progression of constrained skills also depends on the order in which they are taught, which will differ according to language and curriculum priorities. To negotiate this issue, it may be plausible to have a single unconstrained skills scale and separate constrained skills scales for each writing system (syllabic, alphabetic etc.).

These complex issues showed that the relative difficulty of items may be interpreted in different ways, even by specialists in the relevant domains. The next step in the process sought to establish a convergence of expert judgements about item difficulty between tests, by conducting a pairwise comparison analysis. In addition, this next phase focuses on ordering all items on a common scale within a domain, through a comparison of item demands and proficiency descriptions across assessments.

---

1 See Discussion paper on 'Equating existing assessments and validating the UIS reporting scales' (July 2017)

## 4 Constructing a single scale for each domain

While many assessment programs conduct psychometric analyses and report parameters including item difficulties, there are no equating studies that would map these parameters onto a common scale to indicate a comparison of learning progression across different assessment programs. Further, some assessment programs, for example RTI's EGMA and EGRA, do not report psychometric properties of their items. To fill this gap, and to generate data that would permit comparison of the difficulty of the different item sets used, an additional methodology was employed – a 'pairwise comparison'. The purpose of this comparison was to generate a set of difficulty estimates across the entire item set used in the initial steps of development of the reporting scales for reading and mathematics, respectively. A pairwise comparison of items enables the different assessment programs from which those items were sourced to be approximately aligned, providing insight into the underlying learning progression represented by the items.

### 4.1 Pairwise comparison

Pairwise comparison is an effective ranking method for sets with a large number of items, and where it is easier to compare two items to one another, than to rank items across a whole set (Vista & Adams, 2015). The pairwise comparison of item difficulties was designed to obtain and map difficulty estimates of the items from the different assessments on a common scale. By combining many comparisons and many test development specialists who rate the items ('raters'), a numeric scale can be constructed and estimates of the difficulties of items on that scale can be obtained.

### 4.2 Method

The pairwise comparison involved a team of raters who were tasked with comparing pairs of test items and judging their relative difficulties. The raters comprised 12 mathematics learning domain specialists and 12 reading learning domain specialists, each of whom had long previous experience as classroom teachers and also experience as raters for a range of assessments. Separate analyses were conducted for a set of mathematics items and for a set of reading items, both in English. Each rater was presented with approximately 100 item pairs and asked to judge which of a pair of items was more difficult. The results of the pairwise comparison provide a map of the relative difficulty of the items available from different assessments, and a robust estimate of relative difficulties of all items analysed. The data sources for the items were the same as those presented in Table 2, Section 3.1.

ACER ConQuest (Adams, Wu & Wilson, 2012) was used to analyse the pairwise comparison data. The model implemented in ACER ConQuest is based on the BTL model of Bradley and Terry (1952) and Luce (1959). The implementation is described and illustrated in Adams (2010) and in Vista and Adams (2015).



## 4.3 Design

The raters were presented with pairs of items selected from those available in the relevant domain, and asked simply to judge which one of the pair would be more difficult for learners. The pairs were selected and assigned to the raters in a linked design such that relative difficulty across the entire set for each domain could be estimated.

### 4.3.1 Formation of item pairs

For the 533 mathematics items selected for inclusion in the study, almost 142 000 item pairs were formed<sup>2</sup>, and for the 512 reading items, almost 131 000 pairs were formed. Making this many pairwise judgements would not be efficient and it is likely some of the items are very different in their level of difficulty and provide little statistical information about the relative difficulty of items. For example, items from PISA target 15-year-olds and items from EGRA target early elementary reading skills and each would be expected to always be rated the same way by all raters so that the easier item is never selected as the more difficult item.<sup>3</sup> Therefore, a strategy was developed to select an optimal sub-set of item pairs for raters to judge. This strategy ensured that only pairs were compared that were closer in apparent difficulty in order to gain statistically useful information. Comparisons between pairs that were plausibly very different in difficulty were not compared because such comparisons would provide little statistically useful information.

For each domain, all items were combined into one list and were sorted into four groups representing estimated increasing difficulty based on raters' prior knowledge of the different assessments' target populations and published item difficulties (where available) within each assessment. The pairs of items for comparison were made as follows. Within the four groups of estimated relative difficulty the items were sorted into blocks. Within these blocks all possible, unique, pairings of the items were created. To assist in monitoring the relative performance of different raters, in particular to check whether any raters exhibited significantly different standards in the comparisons made, a common set of 100 pairs of items was assigned to all raters in each domain (every second pair within a designated range). These pairs were referred to as the 'reliability set'.

---

2 actually  ${}^{533}C_2$  or  $\frac{533}{5312}$

3 For example, see Luce's Choice Axiom, which (among other things) states that, if item *a* is never selected over item *b* in a set of items, item *a* can be removed from the set without affecting the pairwise probabilities within the set (Luce, 1959).

**TABLE 3** Grouping of items for pairs formation in mathematics

Item block	Item 1	Item 2
1	1–65	2–75
2	66–135	67–140
3	136–200	137–210
4	201–269	202–275
5	270–334	269–344
6	335–398	336–409
7	399–463	400–473
8	464–532	465–533
9	36–50	76–140
10	51–120	141–210
11	121–185	211–275
12	186–254	276–344
13	255–319	345–409
14	320–383	410–473
15	384–398	474–533
16	449–463	474–533
17 *	180–378	379–181

\* 'Reliability' set where items 180-378, and 379-181 were listed in 2 columns and every second item pair retained for comparison by all raters. This results in 100 pairs common to all raters.

**TABLE 4** Grouping of items for pairs formation in reading

Item block	Item 1	Item 2
1	1–65	2–75
2	66–133	67–140
3	134–198	135–207
4	199–264	200–273
5	265–328	266–339
6	329–387	330–398
7	388–447	389–457
8	448–511	449–512
9	36–50	76–140
10	51–118	141–208
11	119–183	209–273
12	184–249	274–339
13	250–313	340–398
14	314–372	399–457
15	373–447	458–512
16 *	180–378	379–181

\* 'Reliability' set where items 180–378, and 379–181 were listed in 2 columns and every second item pair retained for comparison by all raters. This results in 100 pairs common to all raters.

The pairs were presented to each judge in a random order and pairs were selected to minimise overlap of ratings between raters (to maximise the number of unique pairs rated). The 'reliability set' was presented to all raters. For each of the raters, each item appeared 6–18 times with increasing frequency of appearance towards the middle of the list due to the design of item pairs.

### 4.3.2 Supporting reliability in rater judgements

Raters were trained to undertake the pairwise comparisons using a sample of the item set that varied in estimated item difficulty range. In addition, review sessions were held periodically throughout the 'live' comparison work at which raters could share issues and concerns, and establish a common understanding of strategies to be used to form comparative judgements. Some issues were raised and discussed during the training and in subsequent review sessions:

- After some discussion around sample mathematics items, it was decided that judgements should be made on the assumption that a learner who answers the questions has the required assumed knowledge, and that this assumption should be applied no matter the nature of the items. It was decided that year level of typical curriculum coverage was a better predictor of relative difficulty, rather than the item content. Factors that influenced rater judgement as to which item was more difficult included complexity (e.g. reading load, number of steps, and occurrence of sequential processes), the nature of reasoning required, any requirement for spatial thinking (which may tend to create higher demand). Some of the factors that made items easier included the existence of supportive cues (e.g. grid lines when comparing line lengths).
- The reading rater group discussed the features of items that made them more or less difficult: the number of matches of content, direct versus synonymous matches, familiarity, complexity of text, amount of content to be read, the presence of idiom, ease of interpretation of the question.

### 4.3.3 Rating process

All assessment items were scanned and converted to digital form and added to a database. A workflow was established to assign items to the raters, and for the raters to record their judgements for each item pair reviewed. Items from sequential number intervals marked as 'item 1' in Table 3 and Table 4 appeared on the left side of the screen for collecting results of comparisons, and 'item 2' on the right side of the screen. The rater selected which item of each pair was judged as more difficult, the degree of difficulty in forming that judgment (on a three-point scale of easy, moderate, hard) as presented in the middle of the screen. The software recorded the responses as well as any comments or explanations regarding the rater judgements.

Taking into account the rate of judgment (initially estimated at 70 judgements per hour, but considerably faster in practice), the algorithm used to allocate item pairs to the raters incorporated about a quarter of the possible pairs (36 000 of the 142 000 possible comparisons for mathematics, and similarly for reading). The raters completed their judgements in five days over 28 working hours, at a rate of approximately 105 judgements per hour. A total of 29 478 comparisons were made for the reading items; and 34 368 comparisons were made for the mathematical items. This included the 100 pairs that were offered for comparison to all 12 raters (1200 comparisons in total in each domain) to monitor reliability of decisions.

### 4.3.4 Outcomes

The result of the pairwise comparison study provided a scale with the assessment items located along it, representing their estimated difficulty. The results of the study were explored by first considering the quality of the results – the reliability of rater responses and exploring any anomalous results.

#### Reliability of pairwise ratings

Two approaches were used to assess the reliability of the estimations of item difficulties. The first was a review of the reliability sets – the comparisons completed by all raters. The second was an analysis of the residuals of the estimated pairwise predictions.

To analyse the reliability sets, the level of agreement among the raters was measured. The reliability sets contained 100 comparisons for each domain that were presented to all raters. The *true* value for any single comparison is not known and so a criterion was established. It was decided that the underlying true response was represented when 8 or more of the 12 raters agreed on the result of a comparison. In mathematics, 88 comparisons were used to assess rater reliability; in reading, 87 comparisons were used. Cohen’s Kappa (Cohen, 1960) was calculated for each rater in comparison to a ‘perfect’ pseudo rater who always selected the true underlying value.

Overall, there is strong agreement between the raters and the criterion rating for each comparison. As shown in Table 5, no raters scored below  $\kappa = 0.7$  for mathematics, and nine raters scored at least 0.75 (most even higher) for reading. However, three of the raters for reading had weaker scores, between 0.42 and 0.64. These results are considered further in the following paragraphs.

**TABLE 5** Agreement with true ratings for each rater

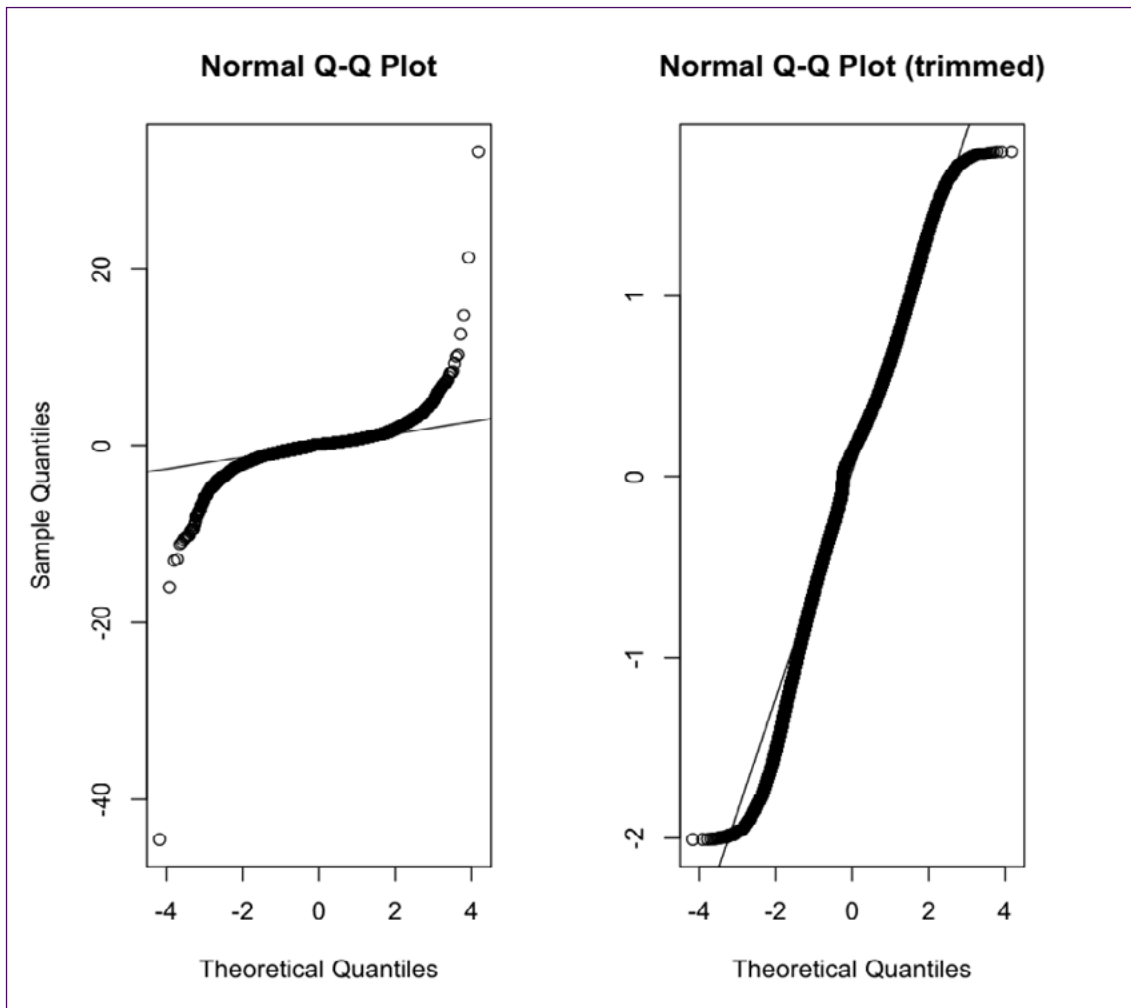
	Rater											
Kappa	1	2	3	4	5	6	7	8	9	10	11	12
<b>Mathematics</b>	0.80	0.75	0.80	0.80	0.84	0.89	0.80	0.70	0.82	0.77	0.86	0.93
<b>Reading</b>	0.93	0.86	0.98	0.91	0.93	0.91	0.75	0.63	0.64	0.42	0.88	0.93

The quality of the ratings was also considered by modelling the residual term for the estimated score for each pairwise comparison (that is, the difference between the observed and predicted estimate, when the model was used to estimate the pairwise comparison). If the data were compatible with the Bradley, Terry and Luce (BTL) model the residuals ( $r$ ) given by:

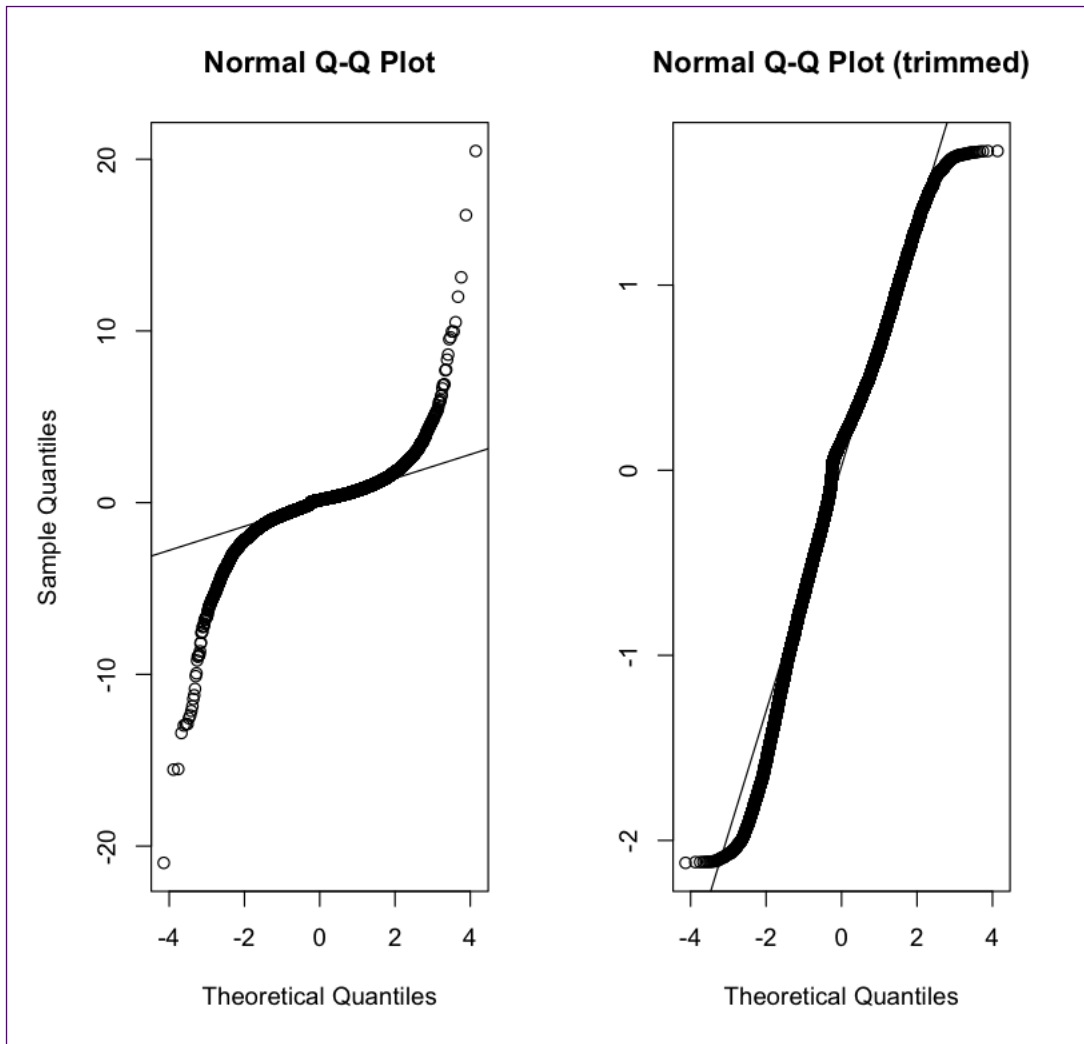
$$r = [(o - e) / \sqrt{(e(1 - e))}]$$

where  $o$  is the observed outcome of a comparison,  $e$  is its expectation under the model should approximate a standard normal distribution.

A normal QQ-plot of these residuals shows very heavy tails, indicating that there were more unusual comparison outcomes than would be statistically expected. Trimming the top and bottom 2.5 per cent of cases yielded a far more acceptable distribution. The distribution of residuals and trimmed residuals for the mathematics domain is shown in Figure 2, and for reading in Figure 3. In the trimmed plot, it is clear that the ordered quantiles of the empirical distribution more closely follow the theoretical standard normal distribution.



**Figure 2** Q-Q plot of residuals for mathematics for the full set of comparisons and a trimmed set (2.5, 97.5%)



**Figure 3** Q-Q plot of residuals for reading for the full set of comparisons and a trimmed set (2.5, 97.5%)

An analysis of the residuals in the trimmed 2.5 and 97.5 percentiles was undertaken. Exploring the cases that are excluded when trimmed can provide evidence of whether any particular rating or assessment disturbs the overall outcomes, and if this agrees with the analysis of the reliability blocks, there is evidence that further refinement of the comparisons included in the analysis is warranted. This was approached in two ways for each of reading and mathematics. The first approach was to identify if any rater was over-represented in the comparisons trimmed. The second approach was to identify if any individual assessment program was over-represented in the comparisons that were trimmed.

Within mathematics, if we trimmed five per cent of cases from 34 368 comparisons we would expect an equal number of cases to be excluded from each rater (12 raters would mean that there was an expectation that approximately 143 observations would be trimmed for each rater, assuming a constant difficulty for each rating). A Chi square test of goodness of fit ( $1 \times 12$  contingency table with expected cell counts

of ~143) indicated significant cell divergence from the expected proportion,  $\chi^2(11) = 92.98, p < 0.001$ . A review of standardised residuals by rater indicated that rater ten had the largest (positive) deviance from the expected value. A binomial test of the count of trimmed comparisons for mathematics rater ten was significant (probability of success is greater than 0.083),  $p = 6.7e^{-8}$ .<sup>4</sup> Similarly, for reading, trimming five per cent of cases would lead to the expectation that approximately 122 comparisons per rater would be excluded. A Chi square test of goodness of fit also indicated significant divergence,  $\chi^2(11) = 232.04, p < 0.001$ . A review of standardised residuals by rater indicated that reading raters eight and ten had the largest (positive) deviance from the expected value. A binomial test of the count of trimmed comparisons for raters eight and ten were both significant:  $p = 1.6^{-10}$  and  $p = 5.6e^{-5}$  respectively.

The analysis of residuals within raters shows strong agreement with the analysis of the reliability sets: the majority of raters were consistent and their ratings were strongly aligned with a 'perfect' rater. In the interest of minimising the contribution of the poorest performing raters, and accounting for the heavy tails of the distribution of residuals, all subsequent analysis is conducted using a trimmed dataset where the comparisons with residuals in the lower- and upper-bound 2.5 percentiles are excluded.

Similarly, if we trimmed five per cent of cases from 34 368 mathematics comparisons we would expect a proportional number of cases to be excluded from each source assessment set. The proportion of items from each assessment in each of the comparisons establishes the expected proportion of items that should be cut for each assessment when the values are trimmed. A Chi square test of goodness of fit (1 x 11 contingency table with expectations as above) indicated significant cell divergence from the expected proportion  $\chi^2(10) = 41.36, p < 0.001$ . A review of the standardised residuals indicated that TIMSS has the largest (positive) deviance from the expected value of each of the source item sets. A binomial test of the count of excluded cases for TIMSS is significant (probability of success is greater than 0.05),  $p = 9.6e^{-4}$ .<sup>5</sup>

In reading, if we trimmed five per cent of cases from 29 478 comparisons, we would expect a proportional number of cases to be excluded from each assessment.

The proportion of items from each assessment in each of the comparisons establishes the expected proportion of items that should be cut for each assessment when the values are trimmed. A Chi square test of goodness of fit (1 x 13 contingency table with expectations as above) indicated significant cell divergence from the expected proportion  $\chi^2(12) = 40.10, p < 0.001$ . A review of the standardised residuals indicated that LLANS has the largest (positive) deviance from the expected value of each of the source item sets. A binomial test of the count of excluded cases for LLANS is not significant (probability of success is greater than 0.05),  $p < 0.05$ .<sup>6</sup>

---

4 All significance tests in this paragraph are reported using a p value with a simple Bonferroni adjustment,  $0.05/12 = 0.0042$ .

5 This value was significant even with a simple Bonferroni adjustment,  $0.05/11 = 0.0045$ .

6 This value was significant even with a simple Bonferroni adjustment,  $0.05/13 = 0.0038$ .

**TABLE 6** The proportion of mathematics items included in the pairwise comparison

Assessment	Proportion of items
ASER	1
EGMA	5
LLANS	7
MTEG Afghanistan	17
Northern Territory OLAY	10
PILNA	22
PISA	5
SACMEQ	2
SISTA	25
TIMSS	5
UWEZO Uganda	1

**TABLE 7** The proportion of reading items included in the pairwise comparison

Assessment	Proportion of items
ASER	1
EGMA	10
LLANS	28
MTEG Afghanistan	10
Northern Territory OLAY	12
PILNA	9
PISA	12
PIRLS	5
SACMEQ	1
SISTA	5
UWEZO Kenya	3
UWEZO Tanzania	2
UWEZO	2

### Results of the pairwise ratings

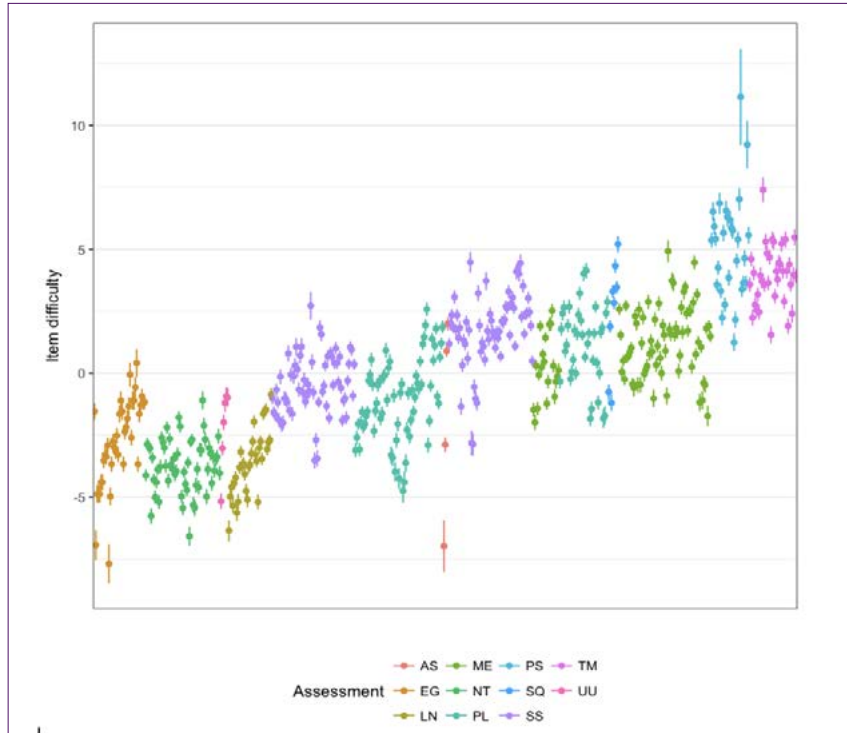
The BTL analysis provided difficulty estimates and standard errors for those estimates for each of the 512 reading items on a single metric, and similarly for the 533 mathematics items. The two scales showed good reliability. The mathematics scale reliability was 0.992 and the reading scale reliability was 0.994<sup>7</sup>. The analysis can be visualised by plotting the estimated locations of the item parameters, which is presented in Figure 4 (mathematics) and Figure 5 (reading)<sup>8</sup>. In these figures, the location of each item on the scale is plotted and the items are grouped (by colour) for each source assessment program.

7 In this case, reliability is reported as the person separation reliability (Wright and Stone, 1979): the ratio of the true variance of the latent variable to the observed variance in the parameter estimates. For a discussion of reliability in IRT and its relationship to reliability in classical test theory, see Adams (2005).

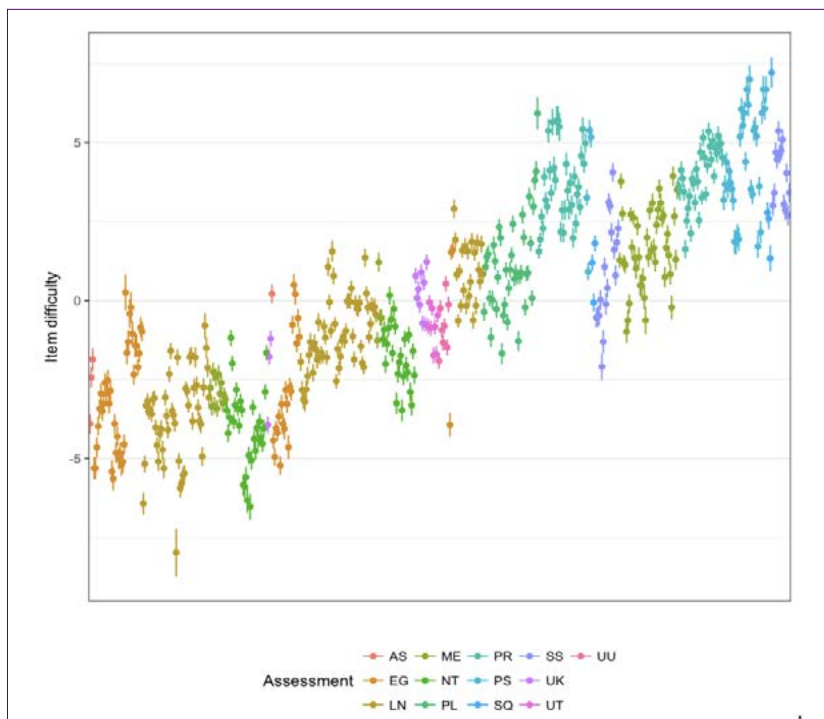
8 Labels used for item sets in figures (see Table 1 for full assessment names):

AS – ASER; EG – EGMA (mathematics) or EGRA (reading); LN – LLANS; ME – MTEG; NT - OLAY Northern Territory; PL – PILNA; PS – PISA; SQ – SACMEQ; SS – SISTA; TM – TIMSS; UU, UK, UT - Uwezo Uganda, Kenya, or Tanzania.



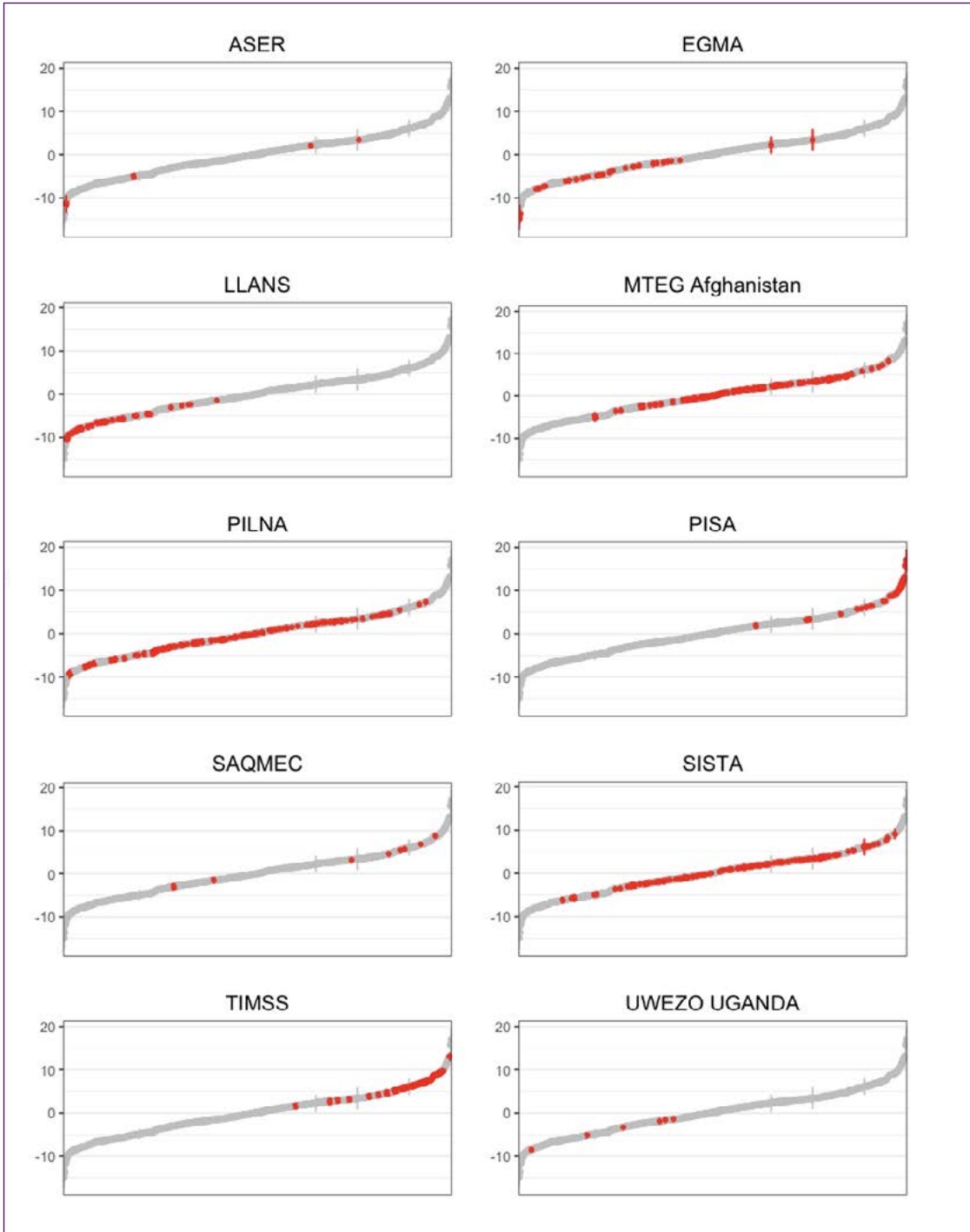


**Figure 4** Estimated difficulty of mathematics items grouped by source assessment program

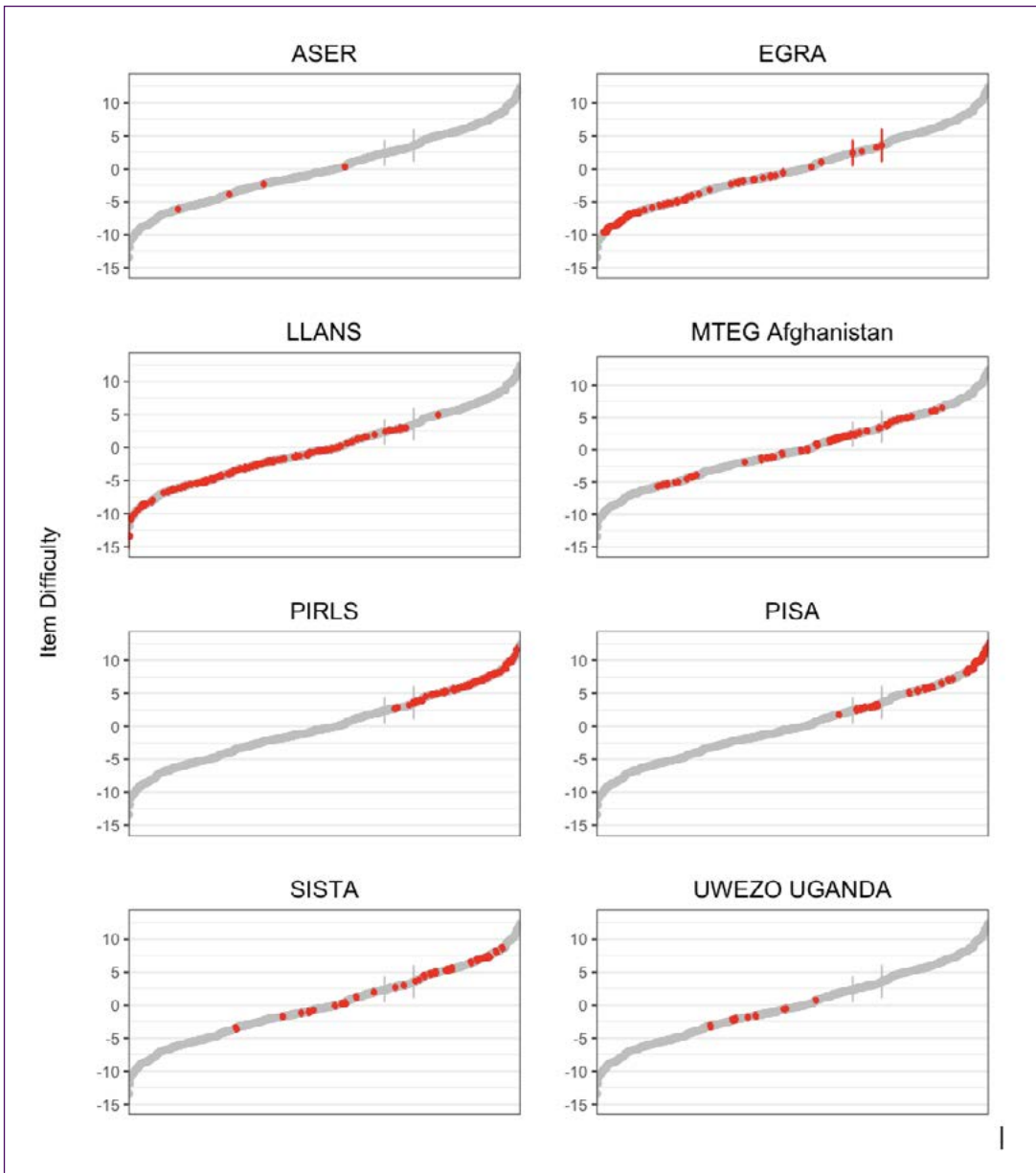


**Figure 5** Estimated difficulty of reading items grouped by source assessment program

The items can also be ordered by difficulty and the items from specific assessment programs highlighted to show the relative position of items. Figure 6 and Figure 7 illustrate this for a selection of assessment programs.



**Figure 6** Relative difficulty of mathematics items from selected source assessment programs



**Figure 7** Relative difficulty of reading items from selected source assessment programs

Both Figure 6 and Figure 7 show spans of item difficulty levels across the different assessment programs, related largely to the intended target audience. Together they operationalise mathematical and reading growth across a wide learning range.

While items from most assessments cluster in locations along the scale that are consistent with expectation (based on the *a priori* assumptions of relative difficulty used to group the items – see section 3.1.2 Formation of Item Pairs), PIRLS seems to be located too far up the continuum. PIRLS targets Grade 4 students while PISA, for example, targets students aged 15 years. Close analysis, however shows that the PIRLS items included have high student demand because of both the length of the passage and the requirement for written responses, often having multiple

reasons required for a correct response. This results in the PIRLS items being estimated as more difficult than ASER, LLANS, EGRA and UWEZO, all of which have comprehension questions orally delivered and orally responded to, which is easier than completing multiple choice and significantly easier than having to write an answer. An expert review of the PIRLS items relative to items that require multiple choice or written response agreed that the PIRLS items were more difficult than the SISTA and MTEG items. The review also noted that the description of PIRLS targeting Grade 4 and SISTA targeting Grade 4 and 6 implies a false equivalence of difficulty as the PIRLS items target students from higher-income countries and with potentially higher levels of average achievement.

After careful consideration, the raters agreed that the PIRLS items were correctly located for several reasons.

- The PIRLS items were correctly located above the items requiring oral responses (e.g. ASER, LLANS, EGRA and UWEZO).
- The PIRLS items were correctly co-located and located above some of the items from assessments that target older-aged children (e.g. SISTA, MTEG, UWEZO) because the PIRLS items are more difficult and the texts more complex (e.g. the text *Fly Eagle Fly* is objectively difficult relative to those in the assessments targeted at older-aged students, and the constructed response items in PIRLS often require students to provide multiple reasons to justify their responses).
- The PIRLS items were co-located or located just below the PISA items because many of the PISA items included in this study were multiple choice or the texts included were not from the extended/more difficult units (e.g. Miser and his Gold is an objectively lower-demand text from within the PISA assessment).

This analysis of relative difficulty between assessments demonstrates the need to consider not just the item difficulty, but also the text complexity, item format, and core competency being primarily targeted. That is, an item that asks a younger-aged student to interpret a complex text with competing information can be rightfully considered more difficult than a multiple-choice item that requires only the location of information in an assessment that targets older-aged children.

Another interesting finding is that some of the items targeted at younger children, that appear higher on the scale than perhaps might be expected, tend to have quite broad confidence intervals. For example, the EGMA and EGRA items that are higher on the mathematics and reading scales are placed with disproportionately low confidence relative to other items. This is plausibly related to idiosyncratic characteristics of the specific items – perhaps being perceived by some raters as more difficult items than they in fact are. For example, some of the items in EGRA require the respondent to do a simple task 100 times (for example alphabet knowledge). For some raters, the required attentional capabilities of such an item led to them to estimate that a relatively easy task repeated many times could be more difficult than a short, but more academically demanding task. The low level of agreement in these pairwise judgements leads to disproportionately high standard errors being estimated. This is a rare phenomenon in this analysis – with only two items (less than 0.1% of all items included) displaying this pattern.

### 4.3.5 Considerations for creating the reporting scales

In addition to its primary purpose of estimating the relative difficulty of items, the pairwise comparison raised some broader considerations to be taken into account in creating the draft reporting scales. Some of these considerations constituted resolution of issues that had been raised in the initial conceptual framework development. Other considerations that emerged constituted new issues, which would require further exploration in subsequent steps of the scale development process. Examples are provided below.

One new issue that emerged was a lack of clarity about the impact of different test formats on item difficulty. Raters from the pairwise study struggled to confidently compare the difficulty of different test formats, such as online versus interview style assessment. For example, in reading, although they were aware that students were able to listen to a picture story book online as often as they needed (time permitting), they were unsure whether this would make the questions easier than having a picture book read to them. They were also unsure whether one way of assessing a skill would be more difficult than another, for example, whether identifying the matching of an initial phoneme from a list of three would be more difficult than generating a word with the same initial phoneme as another. This resulted in items that were essentially testing the same skill spread over a greater range of the scale than would be expected from a conceptual perspective.

One issue that was partially resolved through insight from the pairwise comparison was the suitability of having items calling on both constrained and unconstrained reading skills on the same scale. Although it was not always clear where reading comprehension, listening comprehension and components of reading skills coincided, there did seem to be some identifiable developmental stages that concur with other evidence. This suggests that there is little correlation between 'word for word' accuracy in reading (a constrained skill) and comprehension of the text (an unconstrained skill). This finding coincides with the Longitudinal Literacy and Numeracy Study (LLANS) data and suggests that a focus on word for word accuracy is not a good measure of reading ability (see Khoo & Meiers, 2006).

The pairwise comparison study was not able to provide much insight into the issue of language differences in developing the reading scale. Many of the assessments used have been written for different writing scripts and/or translated into a variety of languages, but there was limited data available to compare with any findings from the pairwise comparison. The available data (SISTA and MTEG Afghanistan assessments) showed that cultural exposure does have an effect. For example, SISTA data reveals that any item requiring writing was significantly more difficult for the students. The raters for the pairwise study were only required to compare the difficulty of items according to their own judgment and were not expected to consider language differences.

For mathematics, the pairwise comparison revealed that some areas of mathematical content are not as well represented by the items as others. For example, for 'arithmetic operations', there is continuous coverage across the scale but for 'line graphs', there is uneven coverage across the scale. Such discontinuities could reflect a lack of items in the pairwise set addressing the intermediary steps in conceptual or skill

development for a particular domain progression. Alternatively, such gaps could be a genuine reflection of a 'leap' in conceptual or skill development on the scale for that domain. The pairwise comparison focused more on primary than secondary school assessments. This results in a limited number of items representing the higher end of the scale and may result in difficulties to describe the domain at its most complex levels.

#### **4.4 Validation of item difficulty estimates and mapping of growth**

The outcomes from the pairwise comparison study made it evident that additional information and resources were required to assist in identifying and describing the existing gaps in the scales. Subsequent validation activities were three-fold:

- qualitative validation of the pairwise item ordering
- comparison of the outcomes from the pairwise comparison study with other empirical sources on item difficulty
- comparison of the outcomes from the pairwise comparison study with outcomes from the Korea Institute for Curriculum and Evaluation validation study.

##### **4.4.1 Qualitative validation of the pairwise item ordering**

The research teams for mathematics and reading reviewed the relative locations of all items to look for consistency of the resulting ordered descriptions with typical curriculum sequences. The learning domain specialists also inspected items with relatively large differences between estimated and empirical difficulties identified in the comparison.

For each of the source assessments included, it was possible to look at the level of agreement or association between the estimated item difficulties and published item difficulties. In each case, where empirical data differed from the estimates, the research team was able to identify possible causes. In some cases, a particular feature of an item could be identified that helped to explain a difficulty estimate from the pairwise study that seemed anomalous. For example, a mathematics item may have proved to be more empirically difficult than expected because factors may not have been properly taken into account by the raters such as: the impact of the reading load, the unfamiliarity of the context used to frame the item, the absence of scaffolding that might normally have been present when undertaking such a task in a classroom context.

Professional judgements were made as to the way in which the cognitive processing required would relate to growth in the reading or mathematics construct, and therefore align the item with others that made similar demands. The cognitive growth sequence implied by the increasing difficulty estimates of the items was scrutinised more generally to look for consistencies and inconsistencies with expectations of the research team. That analysis supported the development of generalised descriptions of the knowledge, understanding and skills shown at different levels of the learning progression.

#### 4.4.2 Comparison of the outcomes from the pairwise comparison study with other empirical sources on item difficulty

A quantitative validation task was also undertaken to compare the difficulty estimates generated by the pairwise comparison with item difficulty estimates available from other empirical sources. The relative difficulty of items sourced, and the consistency between estimated and empirical difficulties, were examined both within and between assessments. This was undertaken by reviewing technical documentation relating to the assessment program, or contacting the owners of the data where the information was not publically available.

There was data available for 13 assessments (seven for mathematics, and six for reading). The associations between the published difficulties and pairwise difficulties were assessed using a mixed-effects regression model with group-mean centering. This accounted for the hierarchical data structure where the item difficulties are nested within assessments. The groups were defined as the interaction between domain and assessment (e.g., one group was Maths PISA and another was Reading PISA). Each group was also standardised (i.e., as part of the group-mean centering), as some assessments reported difficulty as scale-scores (e.g., PISA in the range of ~500), others as logits (e.g., LLANS), and others reported the proportion of children who answered items correctly (e.g., SISTA). The model predicted the estimated pairwise difficulty, with a level 1 effect for the published difficulty and level 2 effect allowing the parameters to vary between groups (e.g., a random intercepts and random slopes model). This allows the extraction of predicted values for the level 1 coefficient for each group which represents the association between pairwise difficulties and published difficulties. It is worth noting that because of the use of group mean centring, the random intercept term is redundant because all intercepts are zero (and this assists in generating interpretable slope coefficients that are analogous to correlations rather than the deviation from the group mean). The model is described in Equation 1, where the pairwise difficulty ( $Y_{ij}$ ) of item  $i$  in assessment program  $j$  is given by a random intercept ( $\beta_0$ ) that equals zero, a random slope ( $\beta_1$ ) given by the average difficulty of all items plus some deviation for each assessment program ( $u_{1j}$ ) times the published difficulty (group mean centred) ( $x_{it}$ ), plus some residual term ( $\epsilon_{ij}$ ).

**Equation 1** Mixed effects model predicting pairwise difficulty for each assessment program

$$Y_{ij} = \beta_0 + \beta_1 x_{it} + \epsilon_{ij}$$

$$\beta_0 = \gamma_{00} + u_{0j} = 0$$

$$\beta_1 = \gamma_{10} + u_{1j}$$

Table 8 summarises the findings. The standardised coefficients can be interpreted as the strength of association between the pairwise and published difficulty estimates: as the published estimates increase from zero to one, the pairwise estimates increase by the value in the *slope* column. Because the only predictor in the model is the published

difficulty parameter and both the predictor and outcome are standardised, the slope values can be interpreted in the same way that a correlations coefficient is: bounded by 1, and 1. It is worth noting that these values are also affected by measurement error in both the published and pairwise difficulty parameters. Because neither measure of difficulty is perfectly reliable, even under perfect conditions, an estimate of a true underlying correlation of 1 would be biased downwards. The expected correlation,  $e(r_{xy})$ , between two random variables is given by the true underlying correlation,  $\rho_{xy}$ , times the square root of the product of the reliabilities of the two random variables,  $\sqrt{r_{xx}r_{yy}}$  (Muchinsky, 1996). In this case, where the reliabilities are typically quite high (>0.9) between the assessments, the estimated correlation for two perfectly correlated variables would not be expected to be greater than 0.9. In this context, correlations in the range 0.50 – 0.84 between the pairwise difficulties and published item difficulties are acceptably high.

**Equation 2** Expected correlation given attenuation due to measurement error

$$e(r_{xy}) = \rho_{xy}\sqrt{r_{xx}r_{yy}}$$

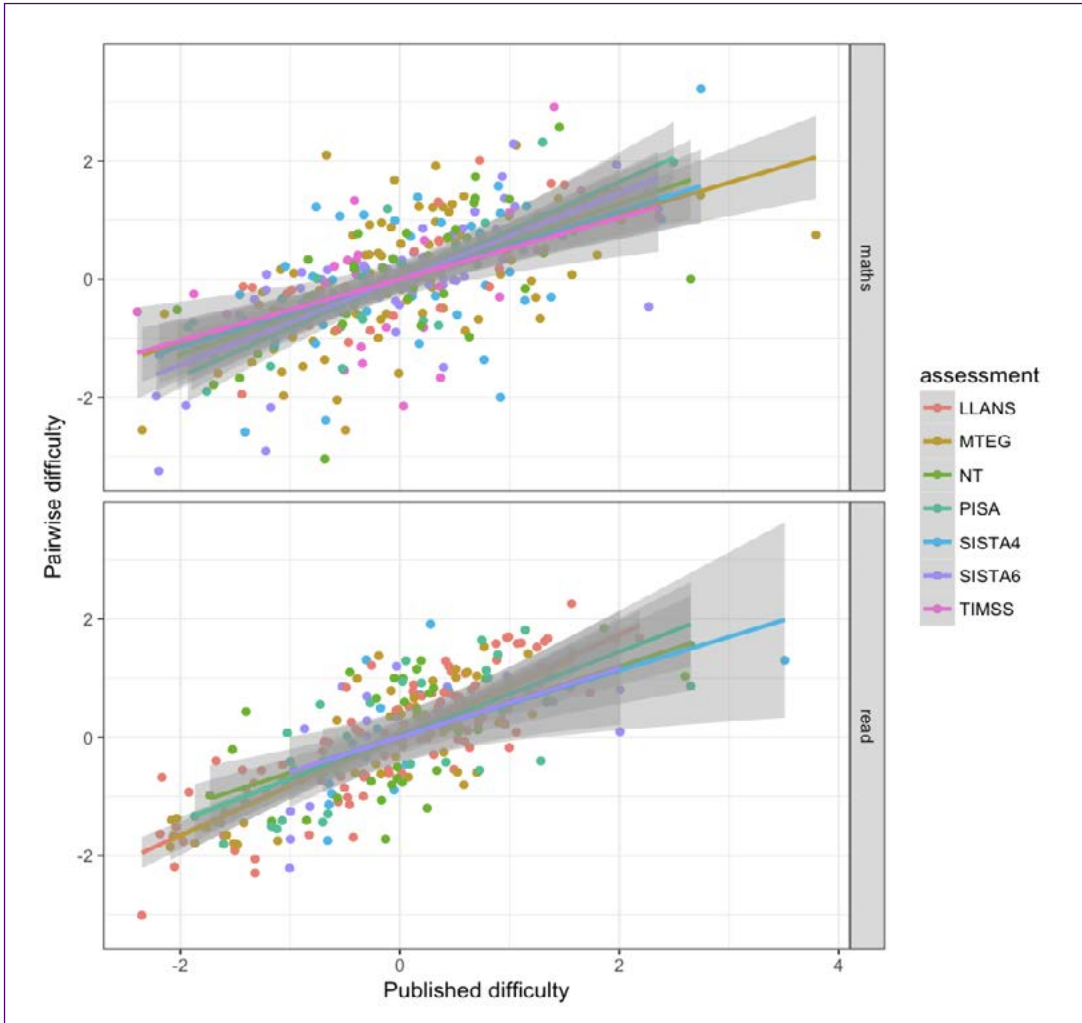
Given that the difficulty estimates in both the pairwise and published data are made with uncertainty – measurement error – values of 0.5 and greater represent good agreement.

**TABLE 8** Summarised findings for association between the pairwise and published item difficulties

Group	Slope ( $\beta$ )
maths:LLANS	0.72
maths:MTEG	0.54
maths:NT	0.62
maths:PISA	0.79
maths:SISTA4	0.56
maths:SISTA6	0.71
maths:TIMSS	0.50
read:LLANS	0.84
read:MTEG	0.82
read:NT	0.58
read:PISA	0.69
read:SISTA4	0.53
read:SISTA6	0.55

These results are visualised in Figure 8 where the consistent and positive trend can be observed. The confidence intervals are relatively wide as there are typically only tens of item difficulties within each assessment.

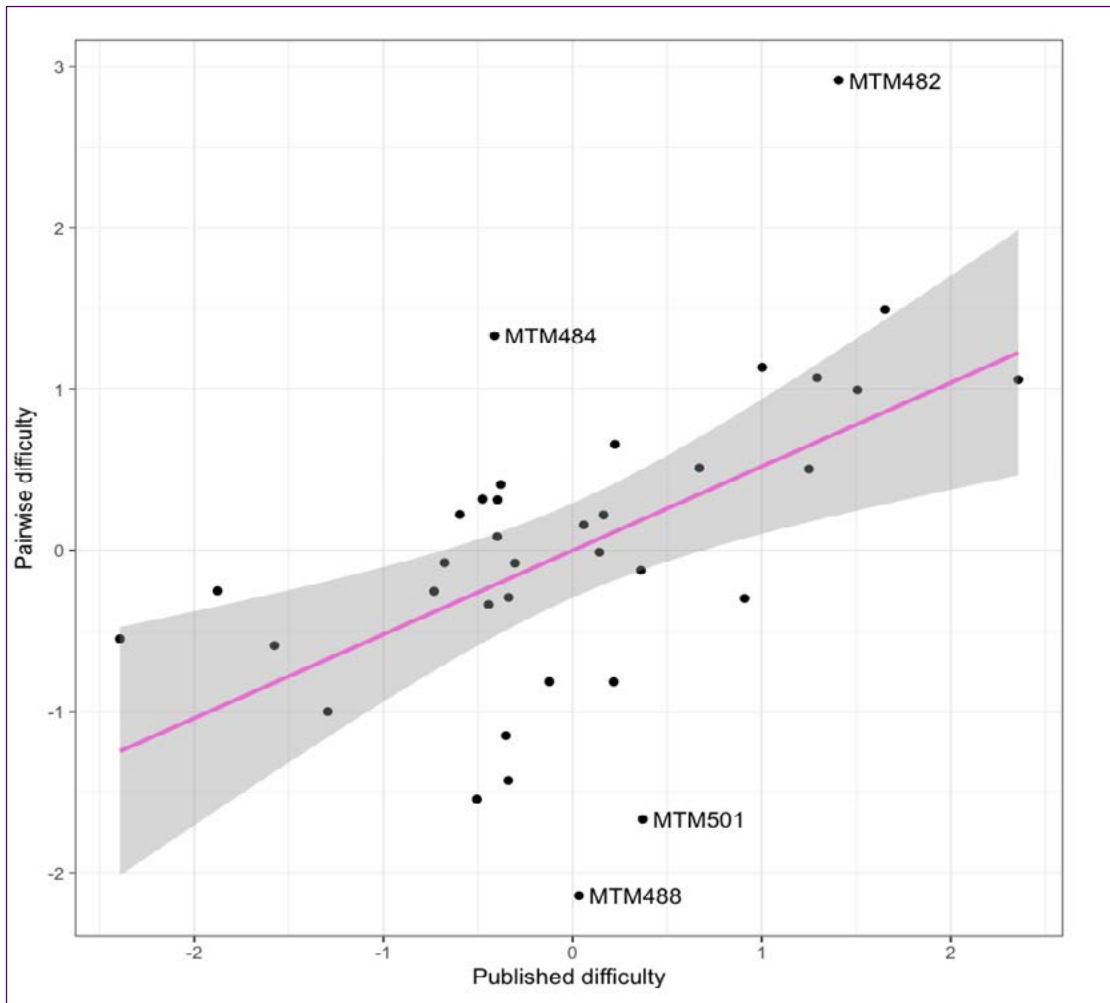




**Figure 8** Observed trends between pairwise and published item difficulties

### Mathematics

For mathematics items, the estimates for the TIMSS items were given careful scrutiny because of the finding above showing that the assessment was over-represented in the tails of the distribution of residuals for the pairwise comparisons. This indicated that item comparisons that included TIMSS items were more difficult for raters to judge consistently. Figure 9 is an illustration of the association between the published and pairwise difficulty estimates for TIMSS only. The items associated with largest residuals are labelled. The labels are from the pairwise study and relate to the following TIMSS items: M052362 = MTM484, M052173 = MTM482, M042152 = MTM501, M052429 = MTM488.



**Figure 9** The association between the pairwise and published item difficulties for the TIMSS items only

The items identified as having large residuals were carefully looked at after the fact. Technical language and formal mathematical terminology were identified as common features that may have made the items difficult to rate. The two items labelled above the line in Figure 9 (the items are shown in Figure 10 and Figure 11) were expected by raters to be more difficult than they in fact were. The item in Figure 10 contains apparently technical content (quadratic expressions) that may have led raters to overestimate item difficulty. The item shown in Figure 11 demands a level of formal geometric understanding that looks difficult, but may in fact be more straight-forward than expected. In both cases, content knowledge expected of students at the relevant age, and some straight forward reasoning in relation to the stimulus provided, would lead students to the correct answer without too much difficulty. The two items labelled below the line in Figure 9 (the items are shown in Figure 12 and Figure 13) were empirically more difficult than was predicted by the raters. The item in Figure 12 looks simple – there is minimal text, and the task simply involves selecting from four options. It is possible that raters underestimate the cognitive demand stemming from the visual reasoning required. The item in Figure 13 involves reasoning about probability

at a level that is commonplace in Australian curriculum (with which the raters were familiar) but may not appear so prominently in the curriculum of some other TIMSS participant countries.

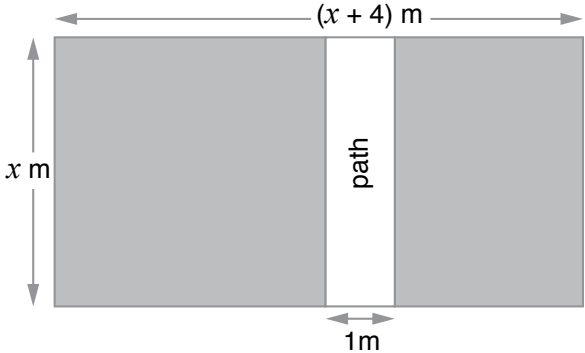
This is a diagram of a rectangular garden.

The white area is a rectangular path that is 1 meter wide.

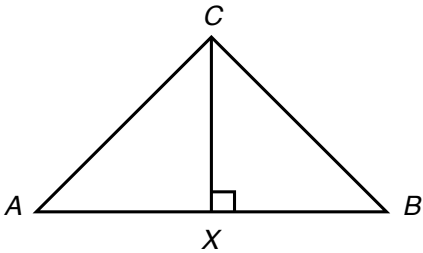
Which expression shows the area of the shaded portion of the garden in  $m^2$

**A**  $x^2 + 3x$   
**B**  $x^2 + 4x$   
**C**  $x^2 + 4x - 1$   
**D**  $x^2 + 3x - 1$

MTM482                      Key: **A**



**Figure 10** TIMSS item M052173



In this triangle:

$AC = BC$

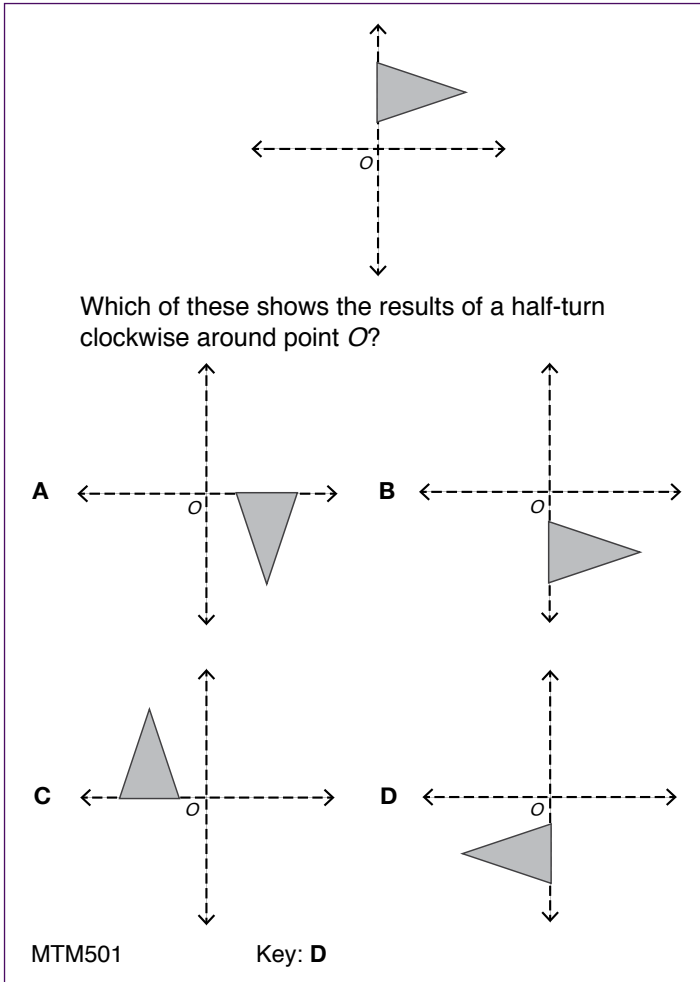
$AB$  is twice as long as  $CX$ .

What is the size of the angle  $B$ ?

Answer: \_\_\_\_\_°

MTM484                      Answer: **45**

**Figure 11** TIMSS item M052362



**Figure 12** TIMSS item M042152

There are 10 marbles in a bag: 5 red, and 5 blue.

Sue draws a marble from the bag at random. The marble is red.

Sue puts the marble back into the bag.

What is the probability that the next marble she draws at random is red?

**A**  $\frac{1}{2}$       **B**  $\frac{4}{10}$

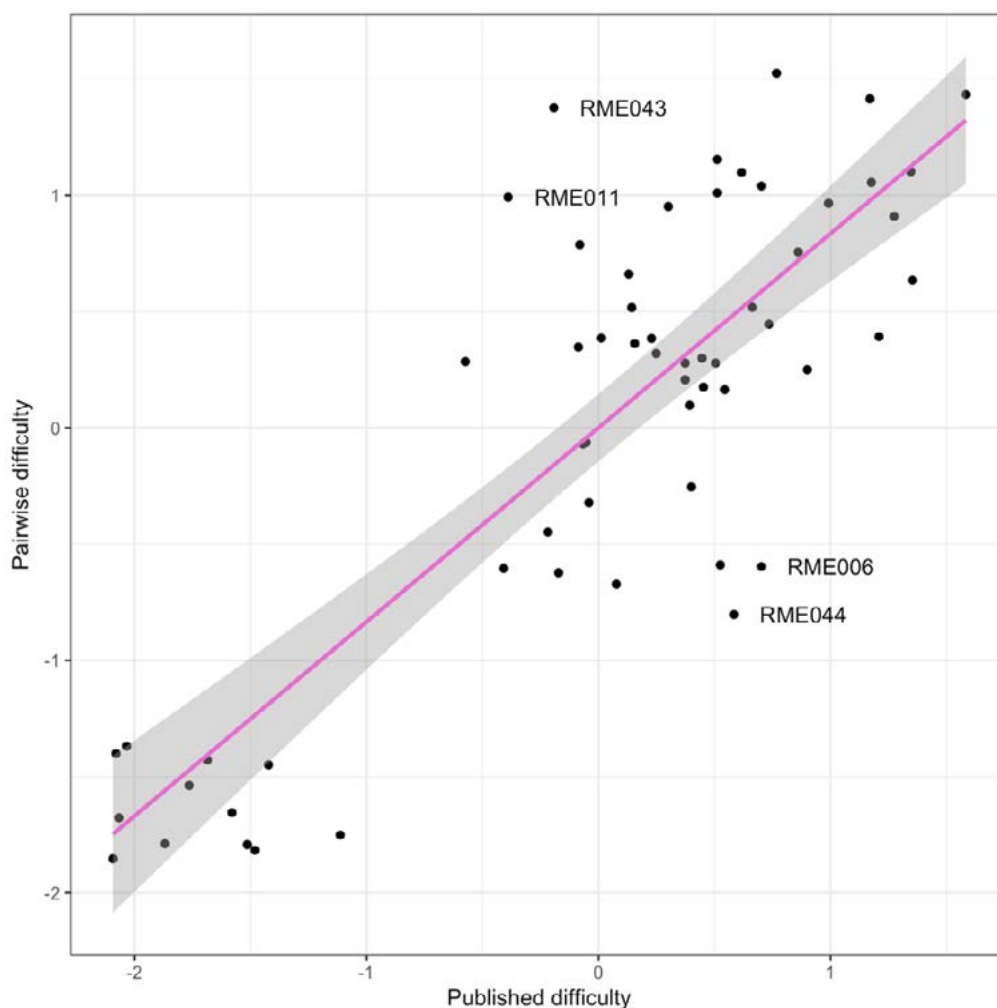
**C**  $\frac{1}{5}$       **D**  $\frac{1}{10}$

MTM488 Key: **A**

**Figure 13** TIMSS item M052429

## Reading

Inspection of reading items from the MTEG Afghanistan shows similar strong relationship between the expectation of raters and the empirical data. The correlation between the two sets of items difficulty estimates in this case was strong ( $r = 0.82$ ). Figure 14 shows a plot of the two sets of item difficulty estimates for these items. Four items are highlighted – two (lying above the trend line) that were judged to be more difficult than was shown in the source assessment data, and two (lying below the line) that were judged to be less difficult than actually observed. Both of the items judged to be more difficult than empirical estimates involved interpreting explicit text, in different forms of written communication between two people. Both items judged to be less difficult than empirical estimates required learners to interpret and reflect on information given.



**Figure 14** Plot of item difficulty estimates for a selection of reading items from the MTEG Afghanistan: pairwise estimates versus empirical estimates from the source assessment program

## The Hole

'I can see something shiny at the bottom,' said Asa. 'Maybe it's a gold coin.'

'Don't be silly,' said Niki, peering into the hole. Her younger brother was always seeing things, creating objects out of nothing.

'Maybe it's a sword,' continued Asa. 'Maybe a king buried a gold sword in the ground many years ago, and then forgot about it.'

'Maybe it's dirt, covered in dirt, covered in more dirt,' said Niki. 'It's just a hole, probably made by a wild animal.'

'You are wrong!' exclaimed Asa. 'No animal could make a hole as big as this!'

'Well, if you are so sure this is not an animal's hole, perhaps you should climb into it.'

Asa began to turn pale. 'Erm ... No. I cannot go in the hole ... because ... I have a sore foot!'

Niki smiled; it had nothing to do with Asa's foot. A big hole could mean a big animal.

'I have an idea,' she said, picking up a stone that lay beside her. 'I will drop this into the hole. If we hear a clink, there is treasure. If we hear a thud, there is dirt. If we hear a yelp, there is an animal.'

Niki dropped the stone and they heard nothing for a moment.

Then they heard a splash.

9 Which word best describes Niki?

- A clever
- B scared
- C excited
- D greedy

**Key A**      **RME011**

**Figure 15** MTEG item R0003T06P The Hole

## School Friends

Dear Nina,

Thanks for your letter. It was great to hear about the town where you now live. I miss our chats and walking home from school together.

I am surprised that you are now hoping to leave school next year. I still enjoy school. I hope that if I study hard I can one day become a nurse and help children to be healthy. This way I can help my community and maybe people in other parts of our country. Staying at school is the only way I will have a chance to find a good job.

Please try to study hard. I think you are a very clever student. You were always much better than me at maths and spelling. In the future, you may change your mind about what you want to do in your life, so it is best to get as much education as possible.

My sister Anna says hello. She misses you reading stories to her – that's another talent you have!

Your friend

Shanti

38 What is the main purpose of Shanti's letter?

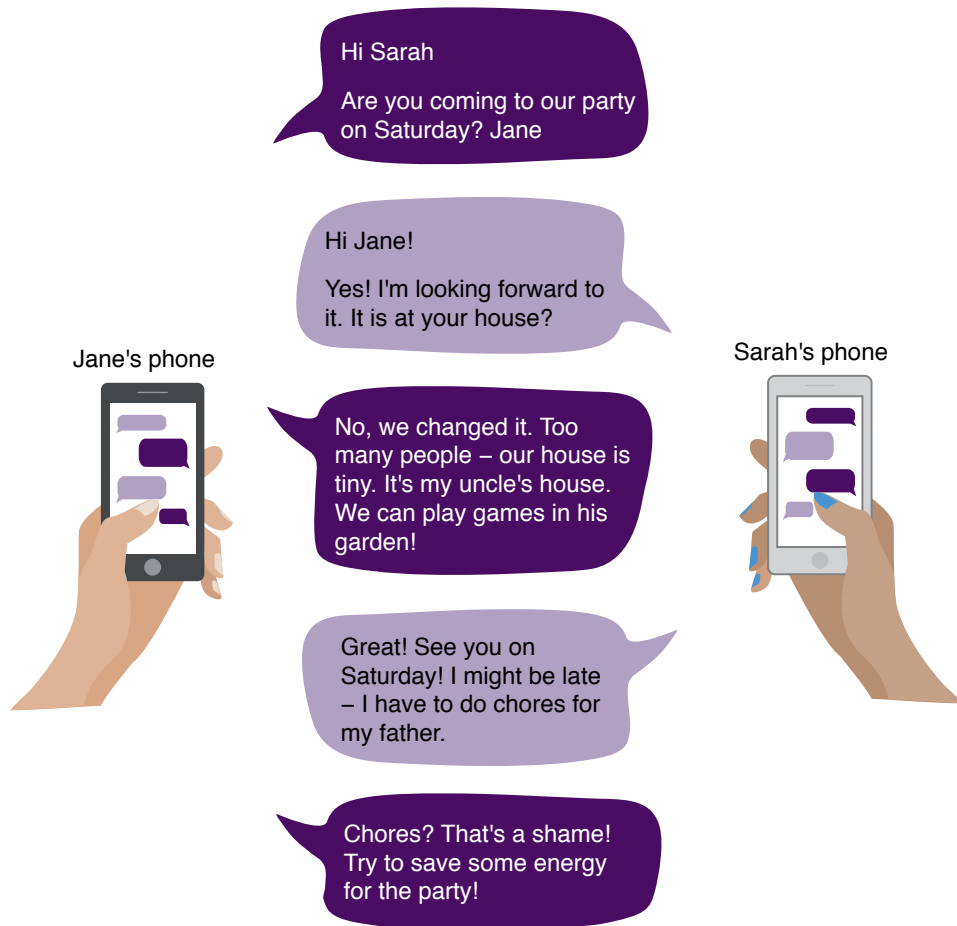
- A to make her friend laugh?
- B to complain about her new school?
- C to persuade her friend to stay at school?
- D to boast about her success?

**Key C**      **RME043**

**Figure 16** MTEG item R0015T03P School Friends

# Party

This is a conversation by text message between two friends.



5 Where is the party?

- A Sarah's house
- B Jane's house
- C Sarah's uncle's house
- D Jane's uncle's house

**Key D**      **RME006**

**Figure 17** MTEG item R0002T01P Party



## A Brother's Note

A boy left this note at home for his brother.

Hi Nathan,

I tried to ring you but my phone wasn't working. I hope you see this note this morning.

We need to meet at Uncle's shop today at 1pm to help clean up.

Can you bring food for five people?

Kye

40 Who wrote the note?

Name: \_\_\_\_\_

Code 1: Identifies Kye (Kamran in Dari; Jamil in Pashto). \*Kye wrote it \* Nathan's brother (Nader's brother in Dari; Ajmal's brother in Pashto) Code 0: Other responses. \* A boy \* brother \* Nathan

**RME044**

**Figure 18** MTEG item R0019T01P A Brother's Note

### 4.4.3 Empirical validation study in collaboration with the Korea Institute for Curriculum and Evaluation (KICE)

A selection of the items used in the pairwise comparison study was administered in a national assessment program for learners at Year 6 level of elementary school conducted in the Republic of Korea, alongside items from the national assessment program. Selection and administration of items occurred in collaboration with researchers from the KICE.

The outcomes of the validation study undertaken in collaboration with the Korea Institute for Curriculum and Evaluation (KICE) using the item ordering from the pairwise comparison study are briefly summarised in abstract form on the KICE website (Kyongah, Taeijoon, Jisun, & Mee-Jee, 2016). Korean experts' review of the reading item ordering endorsed the view that the coverage is sufficient for reading progress as intended (p.29) and that items used were broadly aligned with expectations based on the Korean experience. The study noted that some of the items used appeared to behave differently when used in the Korean national assessment program, and suggested strongly that further validation efforts should be carried out in other countries, and with other languages (p.68) – and note that this is proposed for the next phase of the scale development process. The study also indicates that a level of precision when writing level descriptions could be hard to support in an international context as they may be difficult to interpret in a different linguistic and cultural context.

In the case of mathematics, Korean experts identified a number of elements that would likely be affected by the structure and sequence of curricula in different countries (p.38). ACER-GEM researchers used the identified elements during development of the reporting scales for which information was not consistent. Korean experts judged the difficulty of items to be broadly consistent with expectations based on their domestic experience. Empirical results showed a moderately high correlation ( $r = 0.72$ ) between their item difficulty estimates and estimates from the pairwise comparison study, with only three items showing significant differences between the two estimates.

## 5 Drafting reporting scale descriptions

Draft descriptions of progression were developed for reading and mathematics by synthesising the work undertaken in the previous steps. Items identified as having similar levels of difficulty were examined to identify the underlying knowledge and skills. The scale descriptions were designed to encapsulate the kinds of conceptual knowledge and skills demanded by items, rather than to describe the specific items themselves. Items that addressed similar knowledge and skills that fell into lower or higher levels of difficulty were also examined to identify the contributing factors that seemed to make the underlying skill harder or easier for learners to demonstrate.

Generalisations about the development of an underlying skill across varying levels of progression were then made to develop the level descriptions. The set of ordered items for each domain was reviewed by subject matter specialists to identify where notable progression in learning was observed that would help to divide the scale into levels. This was an iterative process. As the items were examined, the definition of levels on the scales were refined while also trying to make the steps in the progression a similar size across the scale.

As the refinement of proficiency descriptions proceeded, particular attention was paid to the completeness and consistency of the descriptions, both as they depicted the growth aspect of a typical progression, and as they depicted consistency across the major strands of each domain. Sometimes an accomplishment was described at a particular level that implied some earlier learning had occurred, but which had not yet been made explicit – sometimes because no items had been available in the earlier development process that assessed the prior related concept or skill. Similarly, achieving consistency across the domain strands was essential, so that descriptions of achievements in a particular strand of the domain were aligned with related accomplishments in the other strands at a particular level, and that related conceptual and skill development occurring in different strands was complimentary.

To enrich the proficiency descriptions further, the Progressive Achievement Tests (PAT) for reading and mathematics (ACER, 2014) were also compared to the developing scale. The PAT tests were selected due to the availability of the item ordering and demand descriptions; the methodology used to develop their scales was similar to that used during the pairwise comparison, with the scales being refined and validated over an extended time frame. The PAT scales contain detailed descriptions over 10 levels, which span the scaled scores of all the items in each pool, and are written for a teacher audience. Each PAT item has detailed annotations that describe the kinds of skills that the item addresses and the kinds of reasoning required. Generally, there was a high degree of alignment between the pairwise and PAT scale descriptions for both reading and mathematics, especially where both scales had drawn on an extensive set of items located in the same region. Misalignments between the scales largely occurred for regions of the scale where there were only a few items. The pairwise comparison for reading and mathematics had fewer more difficult items populating the higher regions of the scale, so the PAT descriptions were used to further validate the complex conceptual descriptions at the upper end of the scale.

Data from LLANS (Meiers & Forster, 2000) were used to elaborate on the lowest levels of the scale. If discrepancies were identified, items were scrutinised by the subject matter specialists to explain the discrepancies. In some cases, upon re-examination, the skills the items were assessing were clarified. In other cases, the wording of the generalisation about these skills was modified to take into account the wider pool of items now informing the description of this region.

As the program of work moves into Phase II, continued comparisons of the descriptions of progression with additional assessment scales will be essential to further refine and validate the work.

## **5.1 Reading**

The processes outlined in this report to develop the UIS-RS have confirmed the utility of the four defined reading strands outlined in the initial conceptual framework (Retrieving, Interpreting, Reflecting, and Constrained skills). These processes have also clarified the progression with which learners develop reading proficiency. The draft UIS-RS describe progression in reading comprehension across 14 levels, beginning with Level 1. While Level 14 indicates advanced skills in reading, proficiency can develop beyond this level. Each level has a general description of the knowledge, skills and understandings that are typical for learners at that level.

In contrast, constrained skills are only described up to Level 7. At this point sufficient fluency has been achieved with reading aloud being automated, highly accurate and fast enough to allow the reader to focus on the meaning of the words, not on how to decode them. That is, from Level 7 on, problems in the comprehension of text meaning can be affected by many variables including insufficient knowledge of the vocabulary and grammar of the language, insufficient background knowledge to contextualise meaning, insufficient understanding of the structure of the text as well as possible decoding errors or insufficient decoding speed.

The UIS reading scale is not a diagnostic instrument. It is not intended to identify the myriad of possible reasons why readers might have comprehension problems. The purpose of the UIS reading scale is to clarify expectations of what it means to become a more proficient reader in terms of the kinds of comprehension skills that can be demonstrated. Readers who are reading at lower levels than desired, or expected, require further diagnostic assessments to identify possible problems.

Scale descriptions of the comprehension skills demonstrated when the text is read aloud are clearly differentiated from the skill descriptions for texts that are read independently. The listening comprehension descriptions are not continued beyond Level 7 as this is deemed to be the point at which learners have sufficient fluency so that limited constrained skills no longer interfere with their ability to comprehend the text. Listening comprehension skills are described in the lower levels of the scale because in the early stages of reading development, the learner is likely to have very limited skills in turning written symbols into words. Clearly, the learner cannot comprehend a text independently without these skills. The learner may be capable of demonstrating comprehension skills, up to quite high levels, when an illustrated text is read aloud and he or she is able to respond to aural information supported by

illustrations. Learners, whose constrained skills are not yet sufficiently developed to allow them to access the text for themselves, are still able to demonstrate their comprehension skills when texts are read aloud and respond to questions about the meaning. The early years' data examined in the development of the reading scale identified that some learners at the start of school can demonstrate a range of quite sophisticated understandings of authentic, illustrated stories that are read aloud. These texts were well beyond their constrained reading skills at the time. Limited constrained skills, rather than limited comprehension skills restricted what these learners were able to demonstrate in terms of understanding, when required to read the text by themselves. Also, the kinds of very simple, transparent texts that an early decoder might be able to read, provide very limited scope for the assessment of anything other than very simple retrieving of information. By including listening comprehension and drawing on data from tasks where learners were read authentic illustrated texts of some complexity, the research team was able to populate almost all of the strands in the lower levels of the scale with skill descriptions. The focus of the reading scale descriptions from Level 7 onwards is on comprehension of independently read texts.

The complexity of the text is a key factor in determining the comprehension challenges that might be faced by the learner in a reading task. Every aspect of a text can potentially be made more or less complex, which makes it challenging to summarise how text complexity increases as the scale progresses. The underlying element in the progression was defined as the extent and severity of the complexities in the text. At the lower levels of the scale, texts are simple in every aspect. At higher levels, there are one or two moderate complexities in an otherwise simple, familiar text. At higher levels, again, the texts have multiple complexities, some of which start to become challenging. The description of increasing text complexity is given as part of an overarching description of each level to avoid repetition in each of the reading strands. Examples of some of the more challenging vocabulary that might be encountered in texts at different levels, where meaning may need to be construed from context, are also provided to support the descriptions of text complexity.

The level descriptions in the reading reporting scale refer to 'authentic texts' and 'practice texts'. *Authentic texts*, which are read aloud and discussed, support learning to comprehend text-based meaning and motivate reading for meaning. Authentic texts:

- are designed to inform, entertain or persuade
- include details or examples
- have layers of meaning
- are modelled on a range of everyday texts used in the community
- include a range of vocabulary that moves beyond everyday conversational vocabulary
- use sentence structures that are more commonly found in written rather than spoken texts.

At lower levels, illustrations in authentic children's texts provide extensive support for comprehension along with the oral text. The need for supportive illustrations diminishes as listening comprehension skills are developed. When fluency is accurate and fast,

authentic texts can be read independently.

*Practice texts* support learning to decode and become fluent readers. They make the meaning as transparent as possible. They use many words that are easy to decode or recognise. Practice texts are particularly important in languages with complex orthographies or opaque languages where reaching fast accurate fluency may take several years. They are also important in contexts where students have limited knowledge of the language of the text, such as second language learners. Features of practice texts include:

- illustrations that strongly support the meaning
- vocabulary has many phonetically regular words and/or familiar irregular words
- texts have a highly predictable format, such as repetitive sentence structure and familiar narrative development
- the content relates to everyday knowledge or a familiar story topic
- the meaning is prominent with few details or secondary ideas
- the meaning may be secondary such as in word play texts using phonetically regular words.

Practice texts are read independently below Level 7 of the reading scale. Authentic texts are also read to learners in the listening comprehension tasks described at and below Level 7. The texts from Level 7 and above are all authentic texts.

## **5.2 Mathematics**

The draft UIS-RS identifies three defined mathematics strands adopted from the initial conceptual framework: number and algebra, measurement and geometry, and data and probability. The pairwise comparison process helped to identify the progression with which learners develop mathematics proficiency. The draft UIS-RS describes progression in mathematics across 11 levels, beginning with Level 1. Each level has general descriptions of the mathematical competencies that are typical for learners at that level. The concept of competence is based on the six mathematical competencies:

- communication
- representation
- mathematising
- using symbols, operations and formal language
- devising strategies
- reasoning and argument.

Learning progression along the scale is reflected in a growth in these competencies and the capability to activate the competencies when dealing with increasingly complex content in an increasing range of contexts, and to forge and use connections across different but related aspects of mathematics. In addition to the general description of mathematical progression at each level, there are more detailed descriptions of knowledge, skills and understandings within the three content strands for each level.

### 5.2.1 Content

The description of progression aims to build consistency around related content across the strands so that the use of particular related concepts and skills appears at similar levels in the different strands of the scale. The categories of mathematical knowledge, as identified in the framework for foundational concepts, knowledge, skills and applications for mathematics (Appendix 2), were found to fall into broader strands, further justifying the approach of three strands as organisers for the construct. For example, key aspects of 'arithmetic operations', 'fractions and decimals' and 'algebraic thinking' can be identified in the Number and Algebra strand; key aspects of 'properties of shapes' can be identified in the Measurement and Geometry strand. For 'measurement and data', some aspects were better placed to show a trajectory of learning in the Measurement and Geometry strand (e.g. length, area, time), and some aspects were better placed in the data and probability strand (e.g. Pictographs, central tendency).

The mathematics scale describes progression in development for each of these strands. Each strand was considered from Level 1 upwards, to ensure there was both a logical and substantiated flow in the description of the learning trajectory. This process was focused: each component of a strand was considered in turn, such as the stages in development of the conceptual understanding of place value. Through collegial discussion of the information derived from the earlier stages of development of the mathematics scale, the ordering of the stages of skill development were identified, and where apparent 'gaps' were identified, the intermediate steps were described and added to the strand description at the appropriate level. Gaps were usually exposed from the earlier stage of development of the scale, where individual items were used to inform the level descriptions, and particular steps in the scale were not reflected in the items used. In other instances, while the progression was logical, there was a 'jump' across levels. If a skill was described at Level 3, for instance, but then did not 'reappear' until Level 5, if there was an intermediate step that could be justified, it was added to more fully describe development across the levels.

Some aspects, particularly in the Number and Algebra strand, are inherently difficult concepts due to their abstract or formal nature, and so even at the highest level of the scale, the knowledge, understanding and skills required are not as yet expected to be fully developed. For example, for algebraic concepts, the upper levels of the scale include the ability to recognise and manipulate alternative representations of linear and quadratic functions, and to explore and identify functions and their graphs, but exponential and trigonometric functions and differential calculus is 'above' the highest level currently described (Level 11). This is consistent with the expectations for these algebraic concepts to be taught in senior high school mathematics classes.

### 5.2.2 Complexity

For the lower levels of the scale, the complexity of the mathematical processing observed is generally low, that is, the lower levels of mathematical progression permit individuals to engage with tasks that typically involve demonstration of basic knowledge, conceptual understanding and skills in a direct manner. Learners generally do not link concepts within, or between strands. That is not to say that learning

progress does not increase across the lower levels, but rather, the degree to which learners can bring a range of knowledge from the three strands is relatively low; and the solutions to problem situations are typically limited to one or two processing steps. At the higher levels, for example Level 10, mathematical progress permits learners to activate and use a wide range of formal mathematical language, knowledge and related skills *across different mathematical areas. Complexity was often found to be related to context, so that at Level 10, the learners are able to activate this wide range of knowledge and skills in a variety of contexts. At the higher levels, competencies such as reasoning have also developed substantially beyond making direct inferences. For example, at Level 10, learners use connected chains of reasoning to link the different problem elements.*

### 5.2.3 Context

As for reading, context affects the ability of learners to activate and use their mathematical knowledge, and one indicator of growing mathematical proficiency is the ability to activate an increasing range of mathematical knowledge and skills across increasingly diverse contexts. In PISA, an important aspect of mathematical literacy is engaging in problem solving, where the problem is set in a context. Working within a context affects the demands on the problem solver, usually increasing them. As mathematical learning progresses, learners are increasingly able to connect their conceptual and procedural knowledge to the contexts in which that knowledge might be useful. Conversely, they are increasingly able to interact with a context, and notice how their mathematical knowledge can be used.

### 5.2.4 Competencies

Descriptions of increasing mathematical proficiency in PISA are associated with increasing item difficulty (OECD, 2015). This observation was helpful in identifying the key distinguishing aspects of progression in these capabilities both within and across the levels of the scale, and these could then be incorporated in the more general level descriptions.

For example, within Level 8, a similar degree of reasoning skills could be identified across strands: to apply and generalise properties of numbers (commutative, associative and distributive laws); apply and generalise properties of shapes (properties of triangles and shape transformations); and evaluate the validity of simple conclusions based on given data. At Level 4, the reasoning used is less sophisticated, but nevertheless is also similar across strands. At this level, for example, reasoning is used to make comparisons across multiple events: when considering a repeating pattern (numbers or shapes), or related data elements from a table or chart.

## 5.3 Illustrative examples

Reading for meaning and solving mathematical problems are both complex cognitive processes. The nutshell summaries and brief level descriptions of the scales are drawn from definitions of the domains of reading and mathematics. The use of a variety of materials designed for learning support and for assessment purposes was critical to support the learner to understand the intended meaning of the level descriptions.



Copyright-free models of the kinds of test items have been used to illustrate the levels of the scale. Where possible, these item models reflect the scope of the test items used in a range of international tests. Illustrative examples were created for each of the strand descriptions at each of the levels in both reading and mathematics. As the learning progressions are used in a variety of contexts, additional illustrative material will be identified and added. Illustrative examples for mathematics and reading are provided in Appendix 3 and Appendix 4 respectively.

The reading learning scale also includes examples of text complexity for each of the levels. A few texts that illustrate each level are deemed to be of the maximum level of complexity that a learner working at this level of the scale would be able to read and understand the general meaning. It is acknowledged that there are likely to be some aspects of this text that are accessible to less proficient readers and some more subtle aspects that only a more proficient reader is likely to understand. The sample text complexity material is not definitive, it is provided as support to help the learner of the scales better understand the level descriptions.

## 5.4 Progression elements

The scale descriptions for reading and mathematics were further refined to identify and highlight indicators of learning progress within reading and mathematics. These indicators are referred to as 'progression elements'. The progression elements highlight key features of the descriptions at each level that progress as learning advances. The level descriptions are constructed to provide a consistent snapshot of the elements, in order to help the learner to see the progressions.

Some progression elements:

- familiarity of information
- proximity of related information
- complexity of information
- prominence of information
- relatedness of information
- complexity of reasoning

The progression elements in mathematics are competencies that are central to activation and use of particular elements of mathematical knowledge in response to challenges met by individuals. The progression elements used in the mathematics learning progression are:

- communication
- representation
- mathematising
- using symbols, operations and formal language
- devising strategies
- reasoning and argument.

## 6 The Learning Progression Explorer

The final refinement of the UIS-RS has to take into account the demands of presenting and displaying the enormous volume of complex information that is included. In order to make the scales as accessible, useful and user-friendly as possible, an online tool – the Learning Progression Explorer (LPE) – was developed. The LPE presents domain-specific content of the reporting scales in a number of ‘layers’ that reflect different degrees of detail. The LPE’s purpose is to display and present the material in a consistent and coherent format and to facilitate exploration of the information contained in the reporting scales.

The information encapsulated in the LPE for the reading and mathematics scales includes these components, structured in four layers.

### **Layer 1 General description**

A high-level statement that describes the domain in very general terms

### **Layer 2 Levels of the scale**

Descriptions of several levels of overall learning progress in the domain, showing growth from the very early learning steps typically taken at the beginning of formal education, through to more advanced accomplishment expected towards the end of formal schooling. The reading scale describes 14 levels of progression, the mathematics scale describes 11 levels.

### **Layer 3 The strands**

Descriptions at a finer resolution that describe progression across the same levels in each of the main aspects of the domain. In the domain of reading, this involves a closer focus on constrained skills, on retrieving information, interpreting the meaning of text and reflecting on meaning. In mathematics, it involves a closer focus on the major mathematical content strands of Number and Algebra, Measurement and Geometry, and Data and Probability. These are also the major aspects of mathematics that typically define school curriculum structure.

### **Layer 4 Skills illustrations**

At the most detailed layer of the LPE, the conceptual understanding and skills described at each level are illustrated with annotated tasks and questions.

Using the LPE, it is possible to navigate up or down the levels (the overall domain level descriptions, or within the strand); and across the strands within a particular level that may be of interest (to examine, for example, what the main features of learning progress in each of the different strands of the domain are at Level 5). Across each level, progression elements are highlighted. These facilitate a close focus on particular elements within the domain that change and develop as learning progresses.

A major aim of developing the UIS-RS is to build and articulate a common understanding of how learning progresses within a domain, and how learning growth can be described. Establishing such a common understanding, and presenting it using the LPE, may have a number of benefits:

- it provides a basis to interpret measures of learning progress of individual learners
- it provides a basis for meaningful comparison across time, among different learners, or among groups of learners
- it helps educators identify next steps in learning
- it provides support for reviewing and implementing curriculum in ways that take account of typical progressions in learning as a key tool in driving learning progress.

Work on the development of the LPE is ongoing, together with the revision of the reporting scales.

## 7 Noted limitations

Development of common reporting scales has been critiqued based on whether they realistically represent actual learners' growth, and whether these representations are applicable across diverse education systems and cultures. One noted limitation is the extent to which the proficiency descriptions developed actually make sense in relation to the learning growth of individual learners in the domain of interest. The program of work addresses these concerns by responding directly to a need for international reporting tools, driven by a shared commitment to the SDG 4 learning goals and targets. This commitment necessitates a joint effort to confront the conceptual limits of assessment and reporting in rigorous, innovative ways.

A second potential limitation relates to the meaningfulness of and the methods used to define discrete levels on what is, in reality, a continuous variable (the ability being measured). The approach proposed here to that matter recognises that decisions in this area are essentially arbitrary.

Another consideration is of a technical nature, based on the fact that assumptions underlying the models and analysis forms used are never strictly met. The proposed method for developing the reporting scale is one among many possible approaches, all of which have strengths as well as limitations that may place the validity of the scale at risk. The suitability of the approach used here is supported by its origins in a well-established body of assessment theory and practice, which has been applied internationally in PISA (OECD, 2015), PIRLS and TIMSS (TIMSS, 2015), and in many large-scale national assessments. These methods have proven to be effective in enabling the development of comparable international tests, and are also fit-for-purpose for empirically deriving common numerical scales that accommodate results from a range of different assessments.

The development work has proceeded on the basis that a workable and useful set of scales can be built that will provide a perspective on global growth in reading and mathematics outcomes that is currently not available. The task has been approached with the aim of filling this gap. The learning outcomes data that are used for SDG 4 monitoring and reporting are derived from various international, regional and national learning assessments, each having their own designated purposes. Hence, an added concern for reporting on SDG 4 learning outcomes is the comparability of the data across the various assessments and contexts (regions/countries). The steps undertaken in the development of the reporting scales to date, including a wide external review process, show that the learning progressions described with the scales are meaningful, suitable and applicable in a wide range of contexts. The outcomes of the development process, as described in this report, strongly support the conclusion that the scales will be suitable for the purpose of reporting learning outcomes to evaluate progress towards achieving SDG 4. The 'fitness for purpose' in this global context of SDG 4 monitoring, is concerned with the appropriateness of the data for the purpose of international reporting on learning outcomes. It aims to strike a balance between technical rigour and the practical implications of using and comparing data from a variety of existing learning assessments.

While limitations are acknowledged, the development of draft scales as described in this report provides new information that, when further refined, will support national, regional and international assessment programs and the international education community at large to measure, compare and ultimately to improve learning outcomes in mathematics and reading. The approach provides a model that could be extended to other learning domains as required.

## 8 Conclusion and next steps

The approach to develop the UIS-RS aims to balance two seemingly competing necessities: the necessity for common scales to underpin meaningful learning goals and the necessity of having a global framework for monitoring learning outcomes that recognises and can accommodate country-specific contexts and activities. While reconciling these necessities presents complex challenges, the work is driven by a shared purpose to build a workable, meaningful set of scales that are suitable for providing a global perspective on growth in reading and mathematics. Although the assumptions of equivalence underlying the UIS-RS may never be perfectly realised across diverse international contexts, the process outlined in this paper is designed to achieve the best-possible approximation of international comparability. A key element of the scale development approach is that the conceptual frameworks and substantive descriptions of learning progression draw from existing learner assessments – national and international – ensuring that the reporting scales are relevant for different countries and education systems.

In Phase II, the draft UIS-RS will be validated in different contexts (e.g. regional, national, and international). Data will be collected by administering combinations of items, which will enable the empirical determination of the relative difficulties of items across assessment programs.

The next phase of activities will involve multiple linking exercises of items from existing assessment programs against the draft UIS-RS. The start-up of activities in Phase II will see extensive consultation with the view to working with at least 15 countries across different continents. A clearly defined coordination mechanism will be established to facilitate strong cross-country peer support. In-country technical teams will be identified and through a process of cross-country consultation and collaboration, countryspecific plans for test administration will be developed.

Phase II has five outputs. The first will be a pool of calibrated items. The second will be an empirically-based update and validation of the draft UIS-RS. The third will be performance benchmarks set on the scales using an empirical standard-setting exercise. The fourth will be a mapping of performance on items from the assessment programs used in this phase onto the UIS-RS. The fifth will be the establishment of a peer-to-peer capacity support coordination mechanism across multiple country locations.

The process of building scales for reading and mathematics, as described here, has brought together conceptual and empirical work in a contested space. A key driver has been to provide tools that can be used by the international development community to monitor countries' progress in relation to the United Nations Sustainable Development Goals for education, in particular SDG 4.1. Providing a single scale for each of reading and mathematics creates a new possibility for the international community to align the outcomes of various existing assessment programs with common scales, hence to monitor national progress against the benchmarks and indicators that are determined as part of the Education 2030 agenda.

# Appendices

**Appendix 1:** Framework for foundational concepts, knowledge, skills and applications for reading

Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
Concepts of print <sup>9</sup>	Environmental print – identifying and interpreting print in the environment	Recognising print in the environment, which includes distinguishing between writing and other forms of print. This skill grows into understanding what environmental print, such as signage, is communicating.	What is this? / What is in this juice box? [image of orange juice] Where is the writing on this box? What does this sign tell you? [image of signage with and without words e.g. image for toilet sign, road sign with the name of an appropriate city]	Different countries have different kinds of environmental print. Afghanistan, for example, does not have uniform signage for most things (e.g. toilet and exit signs).
	Print conventions – recognising books and their features	The ability to handle books correctly and understand the directionality of print. This grows into an ability to distinguish capital and lowercase letters, and recognise word and sentence boundaries and the meaning of basic punctuation.	Point to the title of this book? Show me where to start reading on this page. Point to a letter/word. Point to a lower case/capital letter. What are these [inverted commas]? What are they for?	Directionality of print, and letter, word and sentence formation differs between languages. Punctuation differs between languages; in Arabic for example there is no clearly standardised use of inverted commas or brackets, and so the way they are used can overlap. There is no capitalisation in the Arabic script. In Arabic, prepositions and indefinite articles are joined onto the following word in some contexts, e.g. 'togo'. This raises the question of what constitutes a word. The conjunction 'and' may be joined onto the following word, which has implications for understanding clauses in sentences as distinct. Arabic focuses on consonants; vowels are commonly only written in for disambiguation. In Early Years testing, all vowels would need to be included.

<sup>9</sup> EGRA does not assess 'Concepts of print' because the required skills have a ceiling effect.

Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
Phonemic awareness	Identifying sounds in spoken language	Recognising rhyme and identifying the sounds in words. This grows into an ability to manipulate phonemes by segmenting, blending and swapping sounds.	<p>Do these words rhyme? [sandal/handle]</p> <p>What word has the same first sound as [house]?</p> <p>What sound is at the end of [rat]?</p> <p>What word do these sounds make? [p/-/o/-t/]</p> <p>What are all the sounds in the word [mud]?</p> <p>Listen to this word 'cat'. Swap the /c/ with /r/. What word do you have now?</p>	<p>Rhyme is common in many languages, although different cultures may have different understandings of rhyme. In Dari and Pashto, for example, rhyming couplets are common in textbooks as a way to encourage memorisation, but ideas of what constitutes rhyme may vary; for example, last letter rhyming may be sufficient even if the vowel sounds are different.</p> <p>Phonemes and phonics can be complicated to test in Dari and Pashto because letters can change shape depending on where they appear in the word and what letters they go with.</p>
Phonics	Making connections between sounds and letters	<p>The ability to differentiate between the names and sounds of letters. This grows into an understanding of letter clusters (digraphs and trigraphs) and other letter patterns.</p> <p>This becomes the basis for phonetically decoding words when reading and spelling words when writing.</p>	<p>Point to the letter (a).</p> <p>Select the letter that makes the sound /n/.</p> <p>What sound does this letter make? /t/</p> <p>What is the first letter of the word? [accompanied by image]</p>	<p>'Languages around the world may vary in the number of sounds and symbols they contain, but they are all finite sets that are mastered by literacy users.' [Paris 2005]</p> <p>Phonics is only relevant to alphabetic languages. In character-based languages, like Chinese, there is no regular relationship between symbols and sounds.</p> <p>Need to consider the impact of different writing systems:</p> <p><b>logographic</b> writing systems – Chinese (only one in the world)</p> <p><b>syllabic</b> writing systems – Japanese kana (nowhere near as simple as this, they use all three)</p> <p><b>alphabetic</b> systems with almost perfect orthography (Estonian, Spanish, Finnish, Italian, Hungarian, Turkish)<sup>10</sup></p>

<sup>10</sup> 'Regardless of language, all children who learn to read advance from being non-readers (unable to read words) to partial readers (can read some items but not others) to readers (can read all or a majority of items). In languages with transparent or 'shallow' orthographies (often called phonetically spelled languages), the progression through these levels is very rapid (just a few months of learning); in languages with more complex or 'deeper' orthographies, this process can take several years. In English, for example, completing the foundation steps requires two or more years, with a rate of gain of only a few new items per month of learning; in comparison, regular and transparent languages such as Italian, Finnish, and Greek require only about a year of instruction for students to reach a comparable level (Seymour et al., 2003).' [EGRA Toolkit]



Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
Vocabulary	Understanding oral vocabulary	Demonstrating the knowledge and use of everyday words, which expands into an ongoing range of more specialised subject-specific words.	Point to the picture of a girl. Listen to these words (donkey, carrot, moon, banana). Which of these is the correct word for the picture?	<p>EGRA does not support the use of images in early years assessment due to the problem of uniformity of images within and across countries.</p> <p>'Care must also be taken to select items that are known cross-culturally. For assessment purposes, difficulties involved in words with multiple meanings can be circumvented by using words that are concrete and visualisable.' [PISA – Reading Components Conceptual Framework]</p> <p>Cross-cultural considerations e.g. moon is better than raccoon.</p> <p>Visualizable considerations e.g. train is better than walk.</p>
Understanding written vocabulary	Understanding written vocabulary	The sight recognition of high frequency words and simple/familiar words common to everyday speech. This extends into reading more complex words and using context to understand new words.	Put the labels on the fruit and vegetables (accompanied by image). Identifies the meaning of a familiar word by making a synonymous match in a short piece of text.	<p>According to the PISA Reading Components Conceptual Framework, reading vocabulary is one of the 'component skills that show the greatest promise for cross-country comparability'. However even easily accessible concepts can vary greatly in the way they are used and understood in different languages and cultures.</p> <p>The testing of the root form of words, or prefixes and suffixes, is compromised in languages where the concept of a word is different to English. In Arabic, for example, 'I write' would be one word.</p>

Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
Reading fluency <sup>11</sup>	Recognising words automatically and reading in a phrased and fluent way	Word- by word-reading develops into the grouping of words into meaningful phrases. This develops into the appropriate use of expression (intonation and pitch) to enhance meaning (e.g. make it entertaining for the audience).	Read this text aloud. [Teachers record fluency on a running record, which assesses accuracy and self-correction.]	According to the PISA Reading Components Conceptual Framework, fluency is another of the 'component skills that show the greatest promise for cross-country comparability'.
<p>As learners become more accomplished readers they progress from reading simple texts, based on familiar topics, to reading increasingly complex texts on less familiar / unfamiliar topics. The skills described below relate to texts of varying complexity because simple and complex inferences can be made from seemingly simple texts and retrieving information from a complex text can be quite difficult.</p> <p>Listening and Reading Comprehension are separated here as they are assessed in different ways, but the same skill set applies to both. The term 'text' refers to both written and spoken texts, although genre-specific and textual feature components are arguably not as relevant to spoken texts.</p> <p style="text-align: center;"><b>Comprehending Text</b></p>				

<sup>11</sup> 'An individual reading aloud accurately and fluently is demonstrating visual word identification processes that are efficiently feeding language processing systems (e.g. working memory) to produce prosodic speech. However, such speech does not guarantee that the reader has comprehended the text. Skilled readers can read familiar texts somewhat fluently aloud without attending to the meaning. However, when oral reading fluency has been operationalised as relatively error-free reading of a simple passage aloud at a normal speaking rate, it has reliably served as a solid indicator of the integration of some basic component skills (e.g. Daane, Campbell, Grigg, Goodman & Oranje, 2005; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Alternatively, breakdowns in accuracy, rate, or both, suggest difficulties in other subcomponents.' [PISA Reading Components Conceptual Framework]

Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
Listening comprehension	Retrieve information	The ability to locate explicitly stated information. This skill begins with word matching and progresses to recognising the relevance of the information or idea in relation to information sought. This recognition may include making synonymous matches, and/or locating information that is not prominent in the text, where more than one piece of information is required, or where there is competing information.	<p>Locates directly stated information about the setting of a story.</p> <p>Locates directly stated information by making a synonymous match.</p> <p>Locates the character responsible for an action in an illustration.</p>	According to the PISA Reading Components Conceptual Framework, 'sentence comprehension and basic passage comprehension' are other 'component skills that show the greatest promise for cross-country comparability'.
	Integrate and interpret	Integrating involves connecting various pieces of information to make meaning. This includes making connections between adjacent sentences, several paragraphs or across multiple texts. Interpreting involves making meaning from something that is not stated and includes identifying underlying assumptions or implications. Readers need to integrate and interpret information to form a broad understanding or make deductions about a specific piece of text. Readers may draw on personal knowledge to make implicit connections, inferences or predictions.	<p>'I like pizza. I always put cheese and tomato on mine but never olives.'</p> <p>Question: What food don't I like?</p> <p>Answer: Olives. [Makes a simple inference]</p> <p>'The sun was shining but the trees were swaying wildly. Elham decided to leave her sunhat at home.'</p> <p>Question: Why didn't Elham wear her sunhat?</p> <p>Answer: It was a windy day. [Make a simple inference by linking across adjacent sentences]</p> <p>Identifies the main idea of the text.</p> <p>Infers the reason for a character's actions.</p> <p>Infers the likely motive of a character from an illustration.</p>	

Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
	Reflect on content and form	<p>Reflecting on content requires the use of external knowledge to make judgements about the text. This could involve articulating or defending a point of view, or evaluating the plausibility of an event or coherence of a plot.</p> <p>Reflecting on form involves a critique of the language usage, presentation features and general or genre-specific features of texts. Readers may need to consider a range of perspectives when making a judgement on content or a particular text feature.</p>	<p>Recognises a likely outcome to a situation by applying outside knowledge.</p> <p>Identifies a literary technique used to create drama.</p> <p>Identifies strategies used by the author to persuade the reader.</p> <p>Recognises the intended effect of an illustration.</p>	
Reading comprehension	Retrieve information	<p>The ability to locate explicitly stated information. At its most basic level, this skill involves word matching and progresses to recognising the relevance of the information or idea in relation to information sought. This recognition may include making synonymous matches, and/or locating information that is not prominent in the text, where more than one piece of information is required, or where there is competing information.</p>	<p>Locate directly stated information about the setting of a story. Locate directly stated information by making a synonymous match.</p> <p>Locates the character responsible for an action in an illustration.</p>	

Broad domain	Detail	Comments (growth elements)	Possible items and/or descriptors	Cross-country and language considerations
	Integrate and interpret	Integrating involves connecting various pieces of information to make meaning. This includes making connections between adjacent sentences, several paragraphs or across multiple texts. Interpreting involves making meaning from something that is not stated and includes identifying underlying assumptions or implications. Readers need to integrate and interpret information to form a broad understanding or make deductions about a specific piece of text. Readers may draw on personal knowledge to make implicit connections, inferences or predictions.	<p>'I like pizza. I always put cheese and tomato on mine but never olives.'</p> <p>Question: What food don't I like?</p> <p>Answer: Olives [Make a simple inference.]</p> <p>Identifies the main idea.</p> <p>Infers the reason for a character's actions.</p> <p>Infers the likely motive of a character from an illustration.</p>	
	Reflect on content and form	Reflecting on content requires the use of external knowledge to make judgements about the text. This could involve articulating or defending a point of view, or evaluating the plausibility of an event or coherence of a plot. Reflecting on form involves a critique of the language usage, presentation features and general or genre-specific features of texts. Readers may need to consider a range of perspectives when making a judgement on content or a particular text feature.	<p>Recognises a likely outcome to a situation by applying outside knowledge.</p> <p>Identifies a literary technique used to create drama.</p> <p>Identifies strategies used by the author to persuade the reader.</p> <p>Recognises the intended effect of an illustration.</p>	

**Appendix 2:** Framework for foundational concepts, knowledge, skills and applications for mathematics

Mathematical knowledge	Growth elements	Possible items
<p>Concept of magnitude</p>	<p>After learners have established concept of one-to-one correspondence, cardinality of sets can be used to define and distinguish collections and magnitudes. A typical growth trajectory might start with simple ideas such as greater/lesser, more/fewer, larger/smaller, longer/shorter, and progress with increasingly precise use of language, application across progressively wider areas, and increasing formalisation of concepts and their measures (length, area, count, volume, ...)</p>	<p>Which pile has the most objects? Which line is longer? Which surface is bigger? Which object is larger?</p>
<p>Positional language, relational language</p>	<p>Spatial concepts begin to be formalised through the use of positional and relational language. A typical growth trajectory would start with general positional concepts (such as near/far, above/below, in front/behind, left/right, up/down, ...) and would develop further to include more specific aspects of the objects being compared (such as similarity, congruence, equivalence), through to finer quantification of properties (e.g. counts to quantify difference)</p>	<p>Shade the object that is above (below, next to, behind, ...) the (specified) object shown Shade the objects inside/outside the shape Draw an object to the left/right of the object shown Which of these groups are equal in size? Draw a picture having the same number of objects as the picture shown</p>
<p>Classification of objects etc.</p>	<p>This is relevant beyond 'number' and 'space', but these areas provide good grounds for this area of concept development. Attributes such as colour, thickness, size, shape, quantity, etc. Attributes of two- and three-dimensional shapes (open/closed; number of edges, corners, faces)</p>	<p>Group these sets according specified criteria (quantity, shape type, etc). What attributes do these objects have in common?</p>

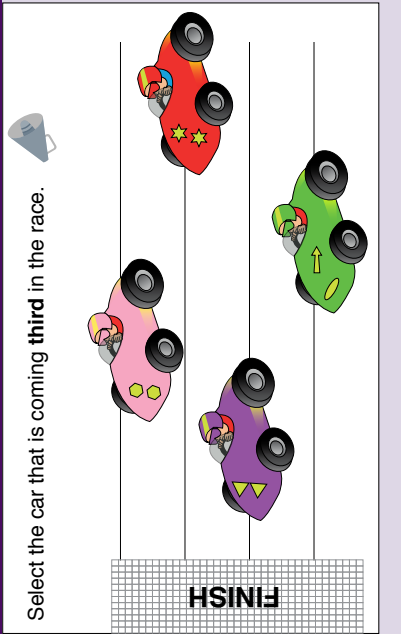
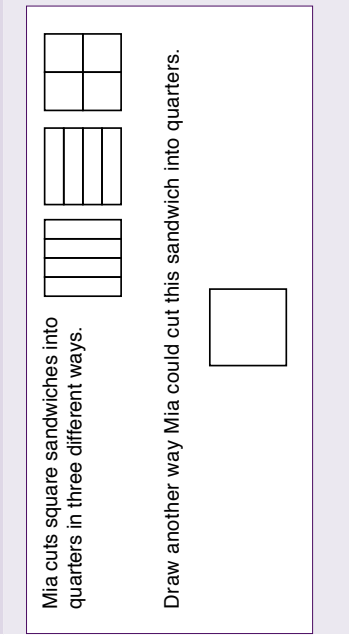
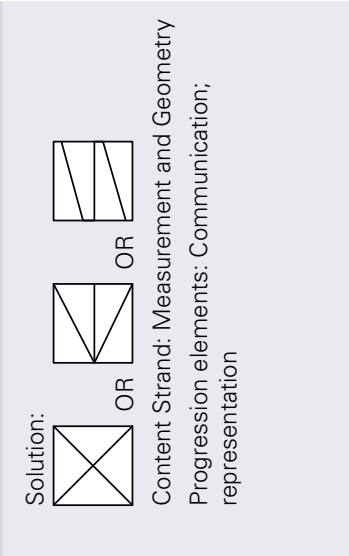
Mathematical knowledge	Growth elements	Possible items
<p>Natural numbers recognising numerals as summaries of counts counting forwards and backwards (including 'skip' counting) understanding relations such as 'more than', 'fewer than'</p> <p>Combining and dividing groups of objects</p> <p>Order, relative magnitude</p> <p>Location of numbers on a number line</p> <p>Number properties</p>	<p>A learning trajectory in the development of basic number concepts starts with the ability to use numbers as models of quantities (counts, measures ...). The meaning of 'four' (etc) is established as an abstraction, through recognising the link across counts of a wide variety of objects and phenomena. Learning the number labels and their meaning, recognising the number of objects in a small group without counting (subitising), composing and decomposing elements of a set and then of numbers as an abstraction, and progressively developing techniques to compare numbers and quantities; using different representations of numbers (e.g. a number line); counting forwards and backwards, in steps of different size; and exploring properties of numbers such as oddness/evenness; associative, commutative, distributive properties, factors, multiples, powers, prime numbers, square numbers, ... These comprise a sequence of advancing conceptual and applied knowledge of number.</p>	<p>Match the numeral with the number of objects it represents</p> <p>Count forward (or backwards) from a particular starting point (by ones, twos, fives, tens, etc)</p> <p>Identify which of these numbers represents the largest (or smallest) number</p> <p>Write these numbers in order from smallest to largest</p> <p>What number is missing from the given set of numbers?</p> <p>What number would come next if you extended the pattern shown?</p> <p>What number would represent the total number of objects shown?</p> <p>Draw three equal groups each of size 5</p> <p>Split this pile into three equal groups</p> <p>Explain what each digit represents in this number</p> <p>Draw a picture of these two numbers that shows what each digit represents</p>
<p>Arithmetic operations</p> <p>Addition</p> <p>Subtraction</p> <p>Multiplication</p> <p>Division</p> <p>Mental arithmetic strategies</p>	<p>The development of understanding of the arithmetic operations begins with observations of concrete and familiar objects, and using appropriate language to characterise those observations. Once the concepts are well established, symbols can be introduced to indicate intention, and algorithms can be introduced to teach methods of applying the operations (for example in subtraction, teaching one or both of decomposition and equal addition). Strategies for mental arithmetic are also important for reinforcing (and assessing the presence of) understanding basic number concepts and skills.</p> <p>For each operation, we need to ask very basic questions (using small numbers), through to much more advanced, and taking account of a variety of procedural strategies, in order to capture information on the stage of development observed. These would then appear at different levels of the metric, and would provide a useful way to distinguish levels of proficiency</p>	<p>You have five pears; I take three of them, how many pears do you have left?</p> <p>Two children are playing with a ball. Three more children join them. How many children are playing now?</p> <p>Draw a picture that shows <math>2 \times 3 = 3 \times 2</math></p> <p>How many groups of size 3 can be made from these 15 objects?</p> <p>How many would be in each group if you split these 12 objects into three equal groups?</p> <p>Draw a picture (no words) to show what this means: <math>2 + 3 = 5</math></p> <p>Show two ways of decomposing and re-grouping the numbers 34 and 17 to demonstrate that <math>34 + 17 = 51</math></p>

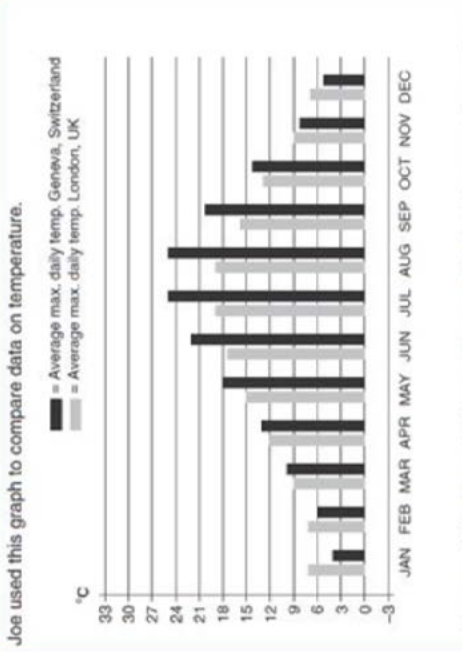
Mathematical knowledge	Growth elements	Possible items
Money arithmetic	<p>Recognise the value of denominations in local currency, classify elements, order elements according to value</p> <p>Shopping (e.g. calculating change)</p>	<p>How many 50 cent pieces can you exchange for a \$5 note?</p> <p>What change would you get when paying a specified amount for an item costing a specified amount?</p>
Fractions and decimals	<p>Fractions of a contiguous whole</p> <p>Fractions of a collection of objects</p> <p>Place value and decimal representation</p> <p>Explore related processes such as doubling and halving</p> <p>Representing fractions and decimals on a number line</p>	<p>Which of the following pictures represents a half?</p> <p>Write a number sentence to show what each of the digits in 372 represents</p> <p>Place these fractions on the number line shown</p>
Integers, integer arithmetic	<p>Elementary ideas based on representations of integers can be introduced early, with formalisation only considered as learner knowledge advances. Beginning with concrete representations of integers (e.g. temperatures above and below zero, height above and below sea level, positive and negative bank balances); and continuing to carrying out arithmetic operations on integers.</p>	<p>At midnight the temperature was 3 degrees C. By 6 am the next morning it had dropped by 10 degrees. What was the temperature at 6 am?</p>
<p>Properties of shapes</p> <p>Symmetry properties</p> <p>Number of edges, corners, faces of two- and three-dimensional figures</p> <p>Composing and decomposing shapes</p> <p>Classification and naming of shapes</p> <p>Geometric properties of plane figures</p>	<p>Awareness of different properties of shapes can start with basic ideas such as closure (open/closed), inside/outside; observing variables such as the number of sides, corners, angles (internal and external), faces; classification of shapes according to such properties; symmetry properties of different shapes; composition and decomposition of shapes; learning the names of shapes; using geometric objects to model familiar objects; using numbers and algebraic thinking to formalise properties; geometric constructions, properties of two- and three-dimensional figures</p>	<p>Which of these pairs of shapes are mirror images of each other (rotated images, dilated images)?</p> <p>Use these two shapes (rectangle, circle) to make a model of a tree.</p>



Mathematical knowledge	Growth elements	Possible items
<p>Algebraic thinking</p> <p>Recognise patterns in number sequences</p> <p>Express number relationships in symbolic form</p> <p>Recognise relations such as those involving doubling and halving</p> <p>Represent simple relations in tabular and graphical form</p>	<p>Use symbols to describe simple relations (e.g. for doubling)</p> <p>Generalisation and symbolic representation of number properties and relationships (e.g. associative, commutative, distributive properties)</p> <p>Using symbols to represent properties of shapes, numbers and relationships</p>	<p>Write three number sentences that show different ways to make the number 12.</p> <p>What number must be added to the box to make this number sentence true (<math>4+6=...+5</math>)</p> <p>Write a number sentence that expresses the perimeter of this shape.</p> <p>Write a number sentence to represent a sequence like 'think of a number, add two, multiply the result by 3, ...'</p> <p>How many of these lengths would you need to make up 6 identical squares?</p>
<p>Measurement and data</p> <p>Length (measuring, comparing)</p> <p>Area (comparing, composing and decomposing shapes)</p> <p>Volume</p> <p>Direction</p> <p>Time</p> <p>Angles</p> <p>Mass</p> <p>Pictographs as representation of frequency</p> <p>Central tendency and dispersion (most frequent value, middle value, representations and quantification of spread)</p>	<p>A typical learning trajectory for measurement would involve starting with experimentation and hands-on experience of comparing and contrasting measurable attributes of objects (length, duration, angle, ...), recognising those attributes can be measured using naive units, developing standardised units to make measurement more systematic, then exploring geometric measurement properties of objects.</p> <p>Data exploration would include interpreting and constructing a variety of data representations, beginning with simple pictograms.</p>	<p>Describe the duration of events (in seconds, minutes, hours, days, weeks, months, years)</p> <p>Tell the time (digital, analogue)</p> <p>If you followed these directions from a given location on the map, where would you end up?</p> <p>Draw two rectangles that each has a perimeter of 18 centimetres.</p>

**Appendix 3:** Task examples designed to illustrate the levels of the mathematics scale

Scale location	Illustrative example	Solution:	Commentary
Level 1	 <p>Select the car that is coming <b>third</b> in the race.</p>	<p>Solution:</p> <p>Content Strand: Number and Algebra</p> <p>Progression elements: Devising strategies, communication, representation</p>	<p>This task requires learners to understand information presented verbally in one short sentence, including the term 'third', and link that to the elements within an image as a representation of ordered objects; find a strategy (direction, how many is 'third', counting three) to identify the 'third' car and select their response.</p> <p>Learners can listen to a recording of the text and/or read the text.</p>
Level 6	 <p>Mia cuts square sandwiches into quarters in three different ways.</p> <p>Draw another way Mia could cut this sandwich into quarters.</p>	<p>Solution:</p>  <p>OR</p> <p>OR</p> <p>OR</p> <p>Content Strand: Measurement and Geometry</p> <p>Progression elements: Communication; representation</p>	<p>This task requires learners to interpret instructions, understand the given diagrams are examples of fraction representations, and to present another construction where a square can be divided into quarters, recognising fractions and their different equivalent representations (represented as parts of a whole) to find a suitable solution.</p>

Scale location	Illustrative example	Commentary
Level 11	<p>Joe used this graph to compare data on temperature.</p>  <p> <input type="checkbox"/> = Average max. daily temp. Geneva, Switzerland  <input type="checkbox"/> = Average max. daily temp. London, UK </p> <p>Joe says, 'When I was in Geneva for a week in April last year, the maximum temperature was 15 degrees each day.'</p> <p>Is this possible?</p> <p><input type="checkbox"/> Yes   <input type="checkbox"/> No</p> <p>Fill in one box and give a reason for your answer.</p> <hr/>	<p>This task requires learners to read and interpret several lines of written text, and interpret a data representation (comparative bar graph for the monthly average temperatures for two cities), including reading both a key and scale of a graph. Learners need to apply reasoning to evaluate the accuracy of a given statement in light of data about average; and construct an argument to support their choice.</p>

Solution:

'Yes' with an explanation that highlights the graph is of average temperatures and so values above average are possible on individual days.

For example, one of the following:

Yes. It's still possible for Geneva to have above average temperatures.

Yes. The graph shows average temperatures only.

Yes. The graph is for many years. Joe was only in Geneva for 1 week.

Yes. The other 3 weeks could have been cooler. [Assumes graph is of last year].

Yes. The average for April is 12 but it's only an average so the temperature can be higher.

Content Strand: Data and Probability

Progression elements: Communication; Reasoning and argument; Using symbols, operations and formal language

**Appendix 4:** Task examples designed to illustrate the levels of the reading scale

Scale Location	Illustrative example			Commentary		
Level 4	<p>Marta had a pen, Jay had a book and Dirí had a ball. Who had a book?</p> <p>Marta Jay Dirí</p>	<p>Lili and Mum went to the shops and Mirka stayed at home. Who stayed at home?</p> <p>Lili Mum Mirka</p>	<p>Vijay has a red hat, a blue coat and yellow socks. What colour is the hat?</p> <p>red blue yellow</p>	<p>Learners can match key words from the question in a sentence and locate the answer in the adjacent words. Understanding syntax helps learners to work out if the required information is likely to be before or after the matched words even if they do not know what some of the words mean.</p> <p>Questions that do not use the same words as the text require learners to demonstrate understanding of the meaning of the text. These kinds of questions are much harder and appear higher on the reading scale.</p>		
<p>Reading Strand: Retrieving Target Skill: Locate information adjacent to the matched word in a compound sentence.</p>						
Level 8	<table border="1"> <tr> <td data-bbox="655 1509 919 1816"> <p>What learners have for lunch</p> <p>Lani</p> <p>Most days my father brings a hot meal to school for me at lunchtime. He likes to cook and takes some food to my mother at work as well. My favourite dish is curry.</p> <p>Tooh</p> <p>I usually take leftovers for my lunch. Mum makes a little more for dinner in the evenings and there is some food left for my lunch next day. I don't mind eating leftovers cold.</p> </td> <td data-bbox="655 1178 919 1509"> <p>Pukz</p> <p>I like sandwiches for my lunch. Any kind of sandwiches will do, but I especially like peanut butter and jelly. I love it.</p> <p>Mum thinks I will get sick of eating the same thing but I never do.</p> <p>Mawi</p> <p>My Mum packs little snacks for me. I might have nuts, cheese, vegetables, fruit or biscuits. I never eat bread because it makes my tummy ache.</p> </td> </tr> </table>			<p>What learners have for lunch</p> <p>Lani</p> <p>Most days my father brings a hot meal to school for me at lunchtime. He likes to cook and takes some food to my mother at work as well. My favourite dish is curry.</p> <p>Tooh</p> <p>I usually take leftovers for my lunch. Mum makes a little more for dinner in the evenings and there is some food left for my lunch next day. I don't mind eating leftovers cold.</p>	<p>Pukz</p> <p>I like sandwiches for my lunch. Any kind of sandwiches will do, but I especially like peanut butter and jelly. I love it.</p> <p>Mum thinks I will get sick of eating the same thing but I never do.</p> <p>Mawi</p> <p>My Mum packs little snacks for me. I might have nuts, cheese, vegetables, fruit or biscuits. I never eat bread because it makes my tummy ache.</p>	<p>Learners need to locate the speaker which can be done with a simple word match. They can make simple links across sentences to identify the reason for each writer's food preferences. They need to understand the meaning of each text as the answers to some of these questions are implied. Learners need to realise that sandwiches are made of bread and that Mawi does not eat bread, therefore she does not take sandwiches. Learners need to link ideas in Tooh's text to realise that the left-overs Tooh eats are the dinner that his mother cooked the night before. In the text by Pukz learners need to recognise that 'favourite' and 'especially like' are synonyms. In the text by Lani learners need to link the word 'curry' at the end of the text to the hot meal that dad brings to realise that dad is bringing curry.</p>
<p>What learners have for lunch</p> <p>Lani</p> <p>Most days my father brings a hot meal to school for me at lunchtime. He likes to cook and takes some food to my mother at work as well. My favourite dish is curry.</p> <p>Tooh</p> <p>I usually take leftovers for my lunch. Mum makes a little more for dinner in the evenings and there is some food left for my lunch next day. I don't mind eating leftovers cold.</p>	<p>Pukz</p> <p>I like sandwiches for my lunch. Any kind of sandwiches will do, but I especially like peanut butter and jelly. I love it.</p> <p>Mum thinks I will get sick of eating the same thing but I never do.</p> <p>Mawi</p> <p>My Mum packs little snacks for me. I might have nuts, cheese, vegetables, fruit or biscuits. I never eat bread because it makes my tummy ache.</p>					
<p>Question: Why doesn't Mawi take sandwiches to school? Solution: She can't eat bread.</p> <p>Question: When does Tooh's mum make the food for his lunch? Solution: At dinnertime.</p> <p>Question: What is Pukz's favourite lunch? Solution: peanut butter and jelly sandwiches</p> <p>Question: How does Lani get hot curry at lunch time? Solution: Her dad brings it to school.</p> <p>Reading Strand: Interpreting Target Skill: Link pieces of related, prominent information in several adjacent sentences when there is a little competing information</p>						

Scale Location	Illustrative example	Commentary
Level 12	<p data-bbox="304 1675 343 1787"><b>Beans</b></p> <p data-bbox="379 965 475 1787">An old farmer was becoming frail. He decided to give his farm to a younger man. He had two nephews but he was not sure if he could trust either of them. He invited them to his farm and gave them a test.</p> <p data-bbox="496 958 592 1787">He gave each nephew a pot and a handful of beans. He told the nephews to plant the beans in their pot and come back in one month. He told them that he would then decide who was the most suitable man to take over his farm.</p> <p data-bbox="612 1339 639 1787">One month later the two nephews came back.</p> <p data-bbox="660 965 756 1787">The first nephew showed the farmer his pot. 'I worked very hard, Uncle. I gave my beans sun and water just like you said. Look at my plants now. They are healthy and green and are almost up to my knee.'</p> <p data-bbox="777 958 873 1787">The second nephew showed the farmer his pot. 'I don't understand, Uncle. I gave my beans sun and water just like you said, but nothing has grown. I don't deserve the farm.'</p> <p data-bbox="893 958 989 1787">The farmer reached into his pocket for the keys to his farm. 'Thank you my nephews. This little test has shown me who I can trust.' He handed the keys to his second nephew. 'The beans had already been cooked. They were never going to grow.'</p> <p data-bbox="1034 1193 1061 1816">Question: Why did the farmer give beans to his nephews?</p> <p data-bbox="1074 1256 1185 1816">to teach the nephews how to be good farmers to see which nephew could grow the biggest plants to show the nephews that farming is hard work to find out if the nephews were honest</p> <p data-bbox="1198 1267 1225 1816">Solution D: to find out if the nephews were honest</p> <p data-bbox="1238 1514 1265 1816">Reading Strand: Interpreting</p> <p data-bbox="1273 815 1329 1816">Target Skill: Make inferences, even when the evidence is slim or the clues subtle, discounting inferences that are highly plausible but not supported by the text.</p>	<p data-bbox="272 230 384 696">This question requires the learner to read closely and carefully to discard an inference that is highly plausible, but not actually supported by evidence in the text.</p> <p data-bbox="435 192 759 696">Up until the last paragraph, it is highly plausible that option B ("to see which nephew could grow the biggest plants") is correct. The key piece of information – that the beans could never have grown – appears in the last sentence of the text. Once this information is learned by the reader, the events of the story have to be reinterpreted and the plausible inference that the test is about caring for plants must be discarded in favour of the correct interpretation that the story is about honesty.</p>

## References

- ACARA (2017). *The Australian National Assessment Program Literacy and Numeracy (NAPLAN) assessment framework: NAPLAN online 2017–2018*. Retrieved from <https://www.nap.edu.au/docs/default-source/default-document-library/naplan-assessment-framework.pdf?sfvrsn=2>
- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Adams, R. J. (2005). Reliability as a measurement design effect. *Measurement, Evaluation, and Statistical Analysis*, 31(2), 162–172. <https://doi.org/10.1016/j.stueduc.2005.05.00>
- Adams, R. J. (2010). *Pairwise comparisons*. Retrieved from <https://www.acer.edu.au/files/Conquest-Notes-2-PairWiseComparisons.pdf>
- Adams, R., Berezner, A., & Jakubowski, M. (2010). *Analysis of PISA 2006 Preferred Items Ranking Using the Percent-Correct Method*. OECD Education Working Papers, No. 46. OECD Publishing. Retrieved from <https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED529643&site=eds-live&scope=site>
- Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). *ACER ConQuest: Generalised item response modelling software* [computer software]. ACER.
- Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255–291). New York: Longman.
- ACER (Australian Council for Educational Research) (2014). *Progressive achievement tests in reading: Comprehension and vocabulary*, 3<sup>rd</sup> Edition. Author: Melbourne.
- Progressive Achievement Tests (ACER, 2017). Retrieved from <https://www.acer.org/pat>
- Best, M., Knight, P., Lietz P., Lockwood, C., Nugroho, D., Tobin, M. (2013). *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries*. (Final report). London: EPPI-Centre, Social Science Research University, Institute of Education, University of London.
- Bradley, R. A., & Terry, M. E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons, *Biometrika*, 39, 324–345
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204–266.
- Clay, M. (1979). Theoretical research and instructional change: A case study. In L. Resnick & P. Weaver (Eds.), *Theory and practice of early reading* (pp. 149–171). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20 (1), 37–46.
- Colorado Department of Education (2003). *CSAP 2003 Technical Report – Part 2*. Retrieved from [http://www.cde.state-co.us/cdeassess/reports/2003/CSAP\\_Tech\\_part2.pdf](http://www.cde.state-co.us/cdeassess/reports/2003/CSAP_Tech_part2.pdf)
- Common Core State Standards (2017). *Preparing America's students for success*. Retrieved from: <http://www.corestandards.org>
- CTB-McGraw Hill (2001). *Terranova*. Monterey, CA: Author.
- Custance, B., Hamilton, R., & Payne, C. (2013). *Building bridges: Moving with the times without losing strong functional foundations*. Paper presented at the ASFLA Conference, Melbourne
- Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., and Oranje, A. (2005). *Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading* (NCES 2006-469). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.
- Ferrara, S. F., Johnson, E., & Chen, W. H. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education*, 18(1), 35–59.
- Freebody, P. (1992). A socio-cultural approach: Resourcing four roles as a literacy learner. In A. Watson & A. Badenhop (Eds.), *Prevention of reading failure* (pp. 48–60). Sydney: Ashton-Scholastic.
- Goss, P., & Chisholm, C. (2016). *Widening gaps: What NAPLAN tells us about student progress. Technical report*. Melbourne: Grattan Institute. Retrieved from <https://grattan.edu.au/wp-content/uploads/2016/03/937-Widening-gaps-technical-report.pdf>
- Harcourt Educational Measurement (2004). *Stanford Achievement Test Series, 10th Edition technical data report*. San Antonio, TX: Author.

- Hieronimus, A. N., & Hoover, H. D. (1986). *Iowa Tests of Basic Skills manual for school administrators*. Chicago: Riverside.
- Huynh, H., & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education, 18*(1), 99–113.
- Khoo, Siek Toon and Meiers, Marion (2006). Literacy in the first three years of school: A longitudinal investigation [online]. *Australian Journal of Language and Literacy 29*(3): 252–267
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University.
- Kreiner, S. & Cristensen, K. B. (2014). Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231.
- Kyongah S., Taeijoon P., Jisun C., & Mee-Jee K. (2016). *Report on the validation study of draft learning metrics in Korea*. Korea: Korea Institute for Curriculum and Evaluation. Retrieved from <http://www.kice.re.kr/board-Cnts/view.do?boardID=1500254&board-Seq=5004821&lev=0&m=0302&-searchType=null&statusYN=W&page=1&s=english>
- Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research, & Evaluation, 8*(10). Retrieved from <http://PAREonline.net/getvn.asp?v=8&n=10>
- Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis*. New York: J. Wiley
- Masters, G., & Forster, M. (1996). *Developmental assessment*. Melbourne: Australian Council for Educational Research.
- Masters, G., Lokan, J., Doig, B., Khoo, S. T., Lindsey, J., Robinson, L. and Zammit, S. (1990). *Profiles of learning: The Basic Skills Testing Program in New South Wales 1989*. Melbourne: Australian Council for Educational Research.
- Meiers, M., & Forster, M. (2000). *The ACER longitudinal literacy and numeracy study (LLANS)*. Retrieved from [http://research.acer.edu.au/monitoring\\_learning/9](http://research.acer.edu.au/monitoring_learning/9)
- Mossenson, L., Hill, P., & Masters, G. (1987). *Tests of reading comprehension [TORCH]*. Melbourne: Australian Council for Educational Research.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63–75.
- Nag, S. & Snowling, M. J. (2012). Reading in an alphasyllabary: Implications for a language universal theory of learning to read. *Scientific Studies of Reading, 16*, 404–423. doi:10.1080/10888438.2011.576352
- National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Bethesda, MD: NICHD.
- New Zealand Ministry of Education (2010). *Literacy learning progressions*. Wellington, NZ: Learning Media Inc.
- No Child Left Behind Act (2001), *Pub. L. No. 107–110, 115 Stat. 1425* (2002). Retrieved from <https://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf>
- OECD (2015). *PISA 2012 Assessment and ANALYTICAL FRAMEWORK: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. Retrieved from: <http://dx.doi.org/10.1787/9789264190511-en>
- Pacific Islands Literacy and Numeracy Assessment (2016). *2015 Pacific Islands Literacy and Numeracy Assessment (PILNA)*. Suva, Fiji: Educational Quality Assessment Program.
- Paris, S. G. (2005) Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*(2), 184–202.
- Paris, S. G. (2011). Developmental differences in early reading skills. In S. B. Neuman and D. K. Dickinson (Eds.), *The handbook of early literacy research*, (pp. 228–241). NY: Guilford.
- Patz, R. J. (2004). Comments on item response theory in NAEP and vertical scaling. Presentation at the Technical Panel Meeting to Discuss the Implementation of Within- and Cross-Grade Scaling for the NAEP 2009 Reading Assessment, Washington, DC, October 29.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogische Institut.
- RTI International (2005). *Early grade reading assessment toolkit*. Research Triangle Park, North Carolina: Author.
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–174.
- Seymour, P., H., K. (2005). Early reading development in European orthographies. In (Eds) Snowling, M.J. & Hulme, C. *The science of reading: A handbook*. Boston, MA,

- Blackwell, 296-315.
- Share, D., L. (2008). On the anglocentricities of current reading research and practice: the perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(40), 584-615.
- Slinde, J. A., & Linn, R. L. (1977). Vertically equate tests: Fact or phantom? *Journal of Educational Measurement*, 14, 23-32.
- Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford.
- Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results*. San Mateo, CA: American Institutes for Research. Retrieved from <http://files.eric.ed.gov/fulltext/ED528992.pdf>
- Tomkiewicz, J. & Schaeffer, G. (2002). *Vertical scaling for custom criterion-referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Trends in International Mathematics and Science Study (2015). *TIMSS 2015 Mathematics Framework*. Retrieved from [https://tims-sandpirls.bc.edu/timss2015/downloads/T15\\_FW\\_Chap1.pdf](https://tims-sandpirls.bc.edu/timss2015/downloads/T15_FW_Chap1.pdf)
- Turner, R. (2002). Proficiency scales construction. In R. Adams, & M. Wu (Eds.). *PISA 2000 Technical Report*. Paris: OECD Publishing.
- Turner, R. (2014a). *The 'literacy' idea*. Assessment GEMs no. 5. Melbourne: Australian Council for Educational Research. Retrieved from: <http://research.acer.edu.au/cgi/viewcontent.cgi?article=1004&context=assessgems>
- Turner, R. (2014b). *Described proficiency scale and learning metrics*. Assessment GEMs no. 4. Melbourne: Australian Council for Educational Research, ACER. Retrieved from: <http://research.acer.edu.au/assessgems/4/>
- Turner, R., Blum, W. & Niss, M. (2015). Using competencies to explain mathematical item demand: A work in progress. In Stacey, K. and Turner, R. (Eds.) *Assessing mathematical literacy: the PISA experience* (pp. 85-115). Switzerland: Springer.
- UNESCO (2012). *International standard classification of education ISCED 2011*. Montreal, Quebec: UNESCO Institute for Statistics. Retrieved from <http://uis.unesco.org/en/topic/international-standard-classification-education-isced>
- Vista, A. & Adams, R. J. (2015). *Modelling pairwise comparisons using the Bradley-Terry-Luce (BTL) model*. Retrieved from: <https://www.acer.edu.au/files/Tutorial-13-Pairwise-comparisons.pdf>
- Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, 41(2), 85-120.
- Wilson, M. (2005). *Constructing measures: An item response modelling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. MESA Press: Chicago.
- Wright, B., & Stone, M. (1979). *Best test design*. MESA Press: Chicago.
- Wu, M., & Adams, R. J. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, 18(2), 93-113.