DESIGNING LEARNING ASSESSMENTS

SCHOOLS 2030

HANDBOOK 1

Core Concepts in Assessment

Oxford MeasurEd

Contents

Figures

Summary

Assessment plays a key role in Schools2030. This handbook outlines three key areas of core importance to assessment in Schools2030. These provide a key foundation for the handbooks on developing Academic and Non-Academic Assessments.

1 Why Measure Learning?

Assessments are one way to understand learning attainment due to education.

Defining why it is you want to assess learning is crucial for ensuring that the assessment will be useful.

There is valid opposition to the notion of assessment as a tool in improving education. These concerns should be considered when designing assessments.

2 How to Measure Learning?

The process of measuring learning has various stages from defining constructs to analysing results.

Before we can measure learning we need to define what we mean by Learning.

Assessments will never perfectly measure learning in a domain, we therefore use selected parts of a domain to infer about the rest.

3 Core Concepts

Standardisation: (i) all students face at least some of the same tasks and, (ii) tests administered and scored in the same way.

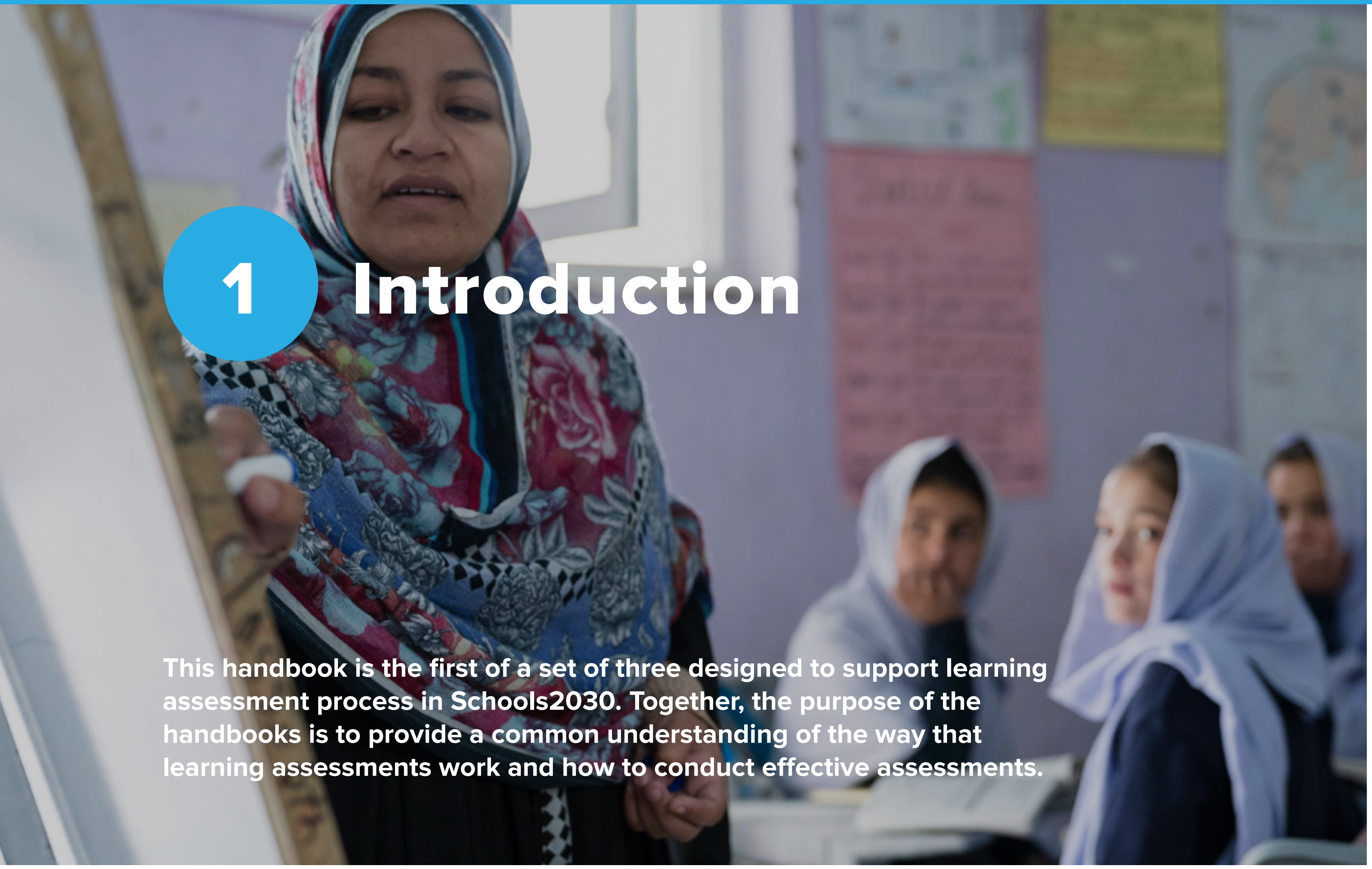
Reliability: Consistency in the hypothetical scenario that a student takes the same test several times.

Validity: Evidence and theory support the interpretations of test scores for the purpose for which they are being used.

Fairness: No student is favoured over other students in demonstrating what they know or understand.

Key terms

TERM	DEFINITION
Assessment purpose	The intended results, impact and decisions that should be provided and assisted by the assessment.
Construct	The concepts or topics for study that a learning assessment measures.
Target domain/construct	The skills, knowledge and values about which assessment users are primarily interested. The learning assessment should be able to draw inference about these constructs.
Assessed domain/ construct	The skills, knowledge and values measured by the assessment. This is a subset of the target construct.
Standardisation	The process of ensuring that everyone who takes the assessment does so in the same conditions so that the results have the same meaning for all test-takers.
Validity	The degree to which evidence and theory support the interpretations of test scores involved in proposed uses of tests.
Construct Irrelevant Variance (CIV)	Variation in test scores caused by factors other than the differences in attainment of the target construct.
Construct Under-Representation (CUR)	Elements of the target construct not being given due prominence within the final test score.
Reliability	Is the consistency of an assessment score in the hypothetical scenario that a student takes the same test several times?
Intra-rater reliability	The degree to which scorers scores are consistent through time so that they reward the same score to answers of equal quality.
Inter-rater reliability	The degree to which different scorers award the same scores to answers of equal quality.
Test fairness	The ability of a test to ensure that no student is favoured over other students in demonstrating what they know or understand.
Differential Item Functioning (DIF)	Items causing unfairness between groups because students with the same attainment from different groups have a different probability of answering the question correctly.



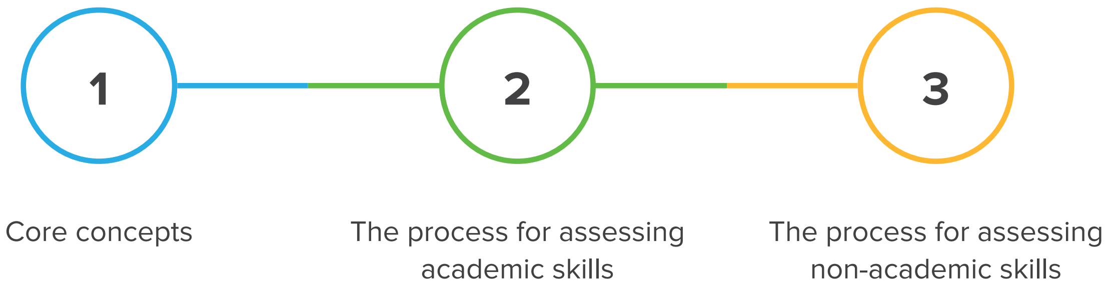
1 Introduction

This handbook is the first of a set of three designed to support learning assessment process in Schools2030. Together, the purpose of the handbooks is to provide a common understanding of the way that learning assessments work and how to conduct effective assessments.

Much of the content will be familiar to many, but the intention is that it can fill gaps or clarify ideas and processes for those who need it and provide grounds for common understanding across the programme. The handbooks are meant as a support and a guide, rather than an academic course.

The intended users are primarily the National Learning Partners as they develop and adapt learning assessment tools, but all or parts of the handbooks can be shared with other Schools2030 partners and stakeholders to help explain elements of the assessment process.

The three handbooks are:


1
2
3
Core concepts
The process for assessing academic skills
The process for assessing non-academic skills

The core concepts handbook describes the core ideas that need to be considered throughout the process. It is fairly theoretical in nature, but it is probably the most important of the three for those who are not familiar with its contents.

The academic skills handbook walks through the process for developing assessments of academic skills. The non-academic skills handbook walks through the process for developing assessments for non-academic skills.

We intend for the three handbooks to be used together. The core concepts handbook should be the starting place for people who have limited familiarity with the concepts of validity, reliability and fairness. Others may want to read it quickly to make sure that any disagreement or confusion with these ideas can be clarified together at the outset.

The academic skills handbook is the main book to refer to while planning the design and implementation of an assessment of academic skills. Similarly, the non-academic skills handbook performs the same purpose for non-academic skills. They can then be referred back to at each stage of the process. They are not comprehensive in detailing every consideration at every step, but they provide a basis for ensuring that important steps and considerations are not missed.

At numerous points along the assessment process, validity, reliability and fairness will need to be discussed and considered. At these points it may be helpful to go back to the core concepts book to provide definitions to shape those discussions. Similarly, there may be parts of the academic and non-academic skills handbooks that prompt the reader to look back to the core concepts handbook.

It is intended that the handbooks can operate as reference materials so that they can be consulted as and when they are needed.

Academic *versus* Non-Academic

There are many different ways to define and sub-define different areas of competency that learners develop. For the purpose of Schools2030 the term Academic is used to refer to areas of knowledge and skills that directly map onto the curriculum, particularly literacy and numeracy. Non-Academic then refers to a range of other socio-emotional knowledge, skills attitudes or values (see Section 4.2 for more details)

1.1 Ways of working together on learning assessments in Schools2030 Structure and Purpose of Assessments

Assessment plays a key role across the three stages of the Schools2030 cycle. During the assess phase, assessment results are used to feed into the human centred design process to help teachers design their classroom-based solutions. During the innovate phase, teachers use a suite of formative assessment approaches to track and iterate their solutions. At the showcase phase, end of year assessment results are used to demonstrate the impact of classroom solutions.

Figure 1 shows how different forms of assessment map onto the cycle that each classroom in Schools2030 will go through each year. For the assess and showcase phases we refer to **teacher implemented tools** as those developed at the national level for teachers to use. These are the tools which this series of handbooks refers to. During the innovate phase, teachers will be supported to develop their own set of **teacher generated tools**. Supporting teachers in this process is covered in a separate handbook.

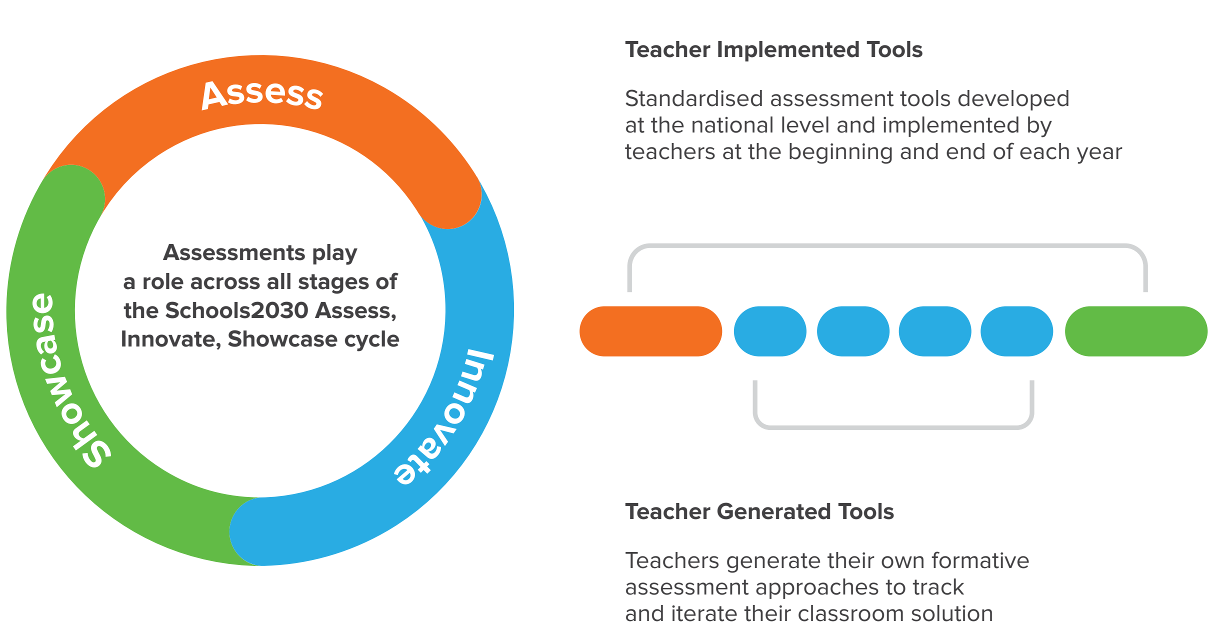

Assess
Innovate
Showcase
Assessments play a role across all stages of the Schools2030 Assess, Innovate, Showcase cycle
Teacher Implemented Tools
Standardised assessment tools developed at the national level and implemented by teachers at the beginning and end of each year
Teacher Generated Tools
Teachers generate their own formative assessment approaches to track and iterate their classroom solution

Figure 1 Assessment in Schools2030

Partnership for developing assessments

Schools2030 is a programme focused on dynamic partnership between actors at the global, national and local level. This partnership is at the heart of how assessments will be developed. It is a core belief of Schools2030 that for assessments to achieve their purpose they need to represent the realities and contexts of those who will be assessed, and those who will use the results of the assessments.

This means that teachers and schools have a key role to play, both in setting priorities for what is to be measured, as well as in delivering and analysing the results of assessment.

To ensure that assessments will be contextually relevant they will be developed specifically for each country. Based on the domains chosen in each country, an assessment partner will develop a set of tools to be used at the beginning and end of each school year (teacher implemented tools), and will work with teachers to develop approaches to formative assessment during each year (Teacher generated approaches).

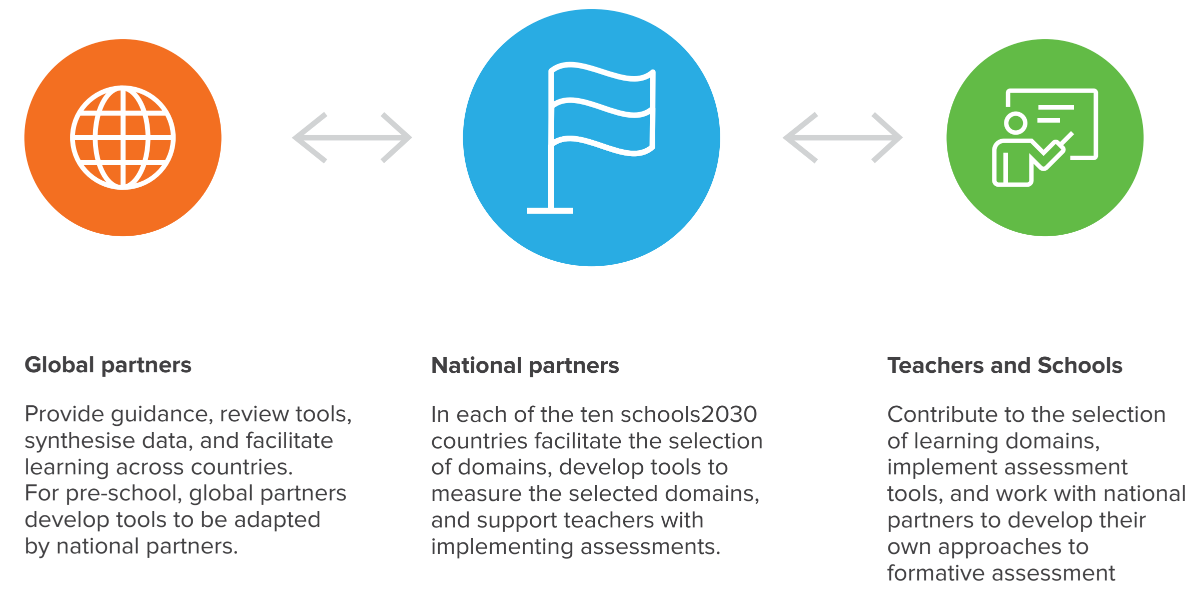

Global partners
Provide guidance, review tools, synthesise data, and facilitate learning across countries. For pre-school, global partners develop tools to be adapted by national partners.
National partners
In each of the ten schools2030 countries facilitate the selection of domains, develop tools to measure the selected domains, and support teachers with implementing assessments.
Teachers and Schools
Contribute to the selection of learning domains, implement assessment tools, and work with national partners to develop their own approaches to formative assessment

Figure 2 Actors involved in Assessment for Schools 2030

1.2 This handbook

This handbook is concerned with the core concepts in learning assessment, particularly validity, reliability and fairness. Although it outlines the assessment process, it does not engage with the details. Instead, the aim is to provide descriptions of the core concepts as clearly as possible. These concepts are so central to assessment that it is difficult to engage effectively with learning assessments without some understanding of them. The content of the rest of this handbook is described below.

SECTION	CONTENT
Why do we measure learning outcomes?	• The various purposes of learning assessments. • Opposition to learning outcomes and why we need to be careful as we design and use learning assessments.
How do we measure learning outcomes?	• The learning assessment process, in general. • The idea of sampling from the sub-domains within a construct and from the possible questions that could be used in a test. • What is standardisation in education testing?
Core concepts	• Constructs. • Validity. • Reliability. • Fairness.

2 Why do we measure learning outcomes?

Summary

- Learning attainment is the value provided by education, so it is important to understand how much and what learning is taking place. Therefore, learning assessments are necessary.
- The particular ***purpose*** of any specific assessment is pivotal for a range of decisions that need to be made in the assessment process so it must be decided and articulated at the outset.
- Valid opposition and concerns about assessment have been raised. These do not undermine the benefit and importance of learning assessments, but they demonstrate the need for care, particularly in areas raised by critics.

2.1 Assessment purposes

Learning assessment is necessary because the relationship between instruction (what a learner hears, reads, does and experiences) and learning (the acquisition of knowledge, skills and attitudes) is not straight forward. It cannot be assumed that a learner who has undergone some instruction will have learnt everything that was intended, otherwise we could simply track instruction. Further, the value of education is not in the instruction but in the knowledge, skills and attitudes that equip the learner to prosper in all areas of life. As we need to assess whether learning has happened, or is happening and we cannot assume a relationship with instruction, learning has to be measured directly.

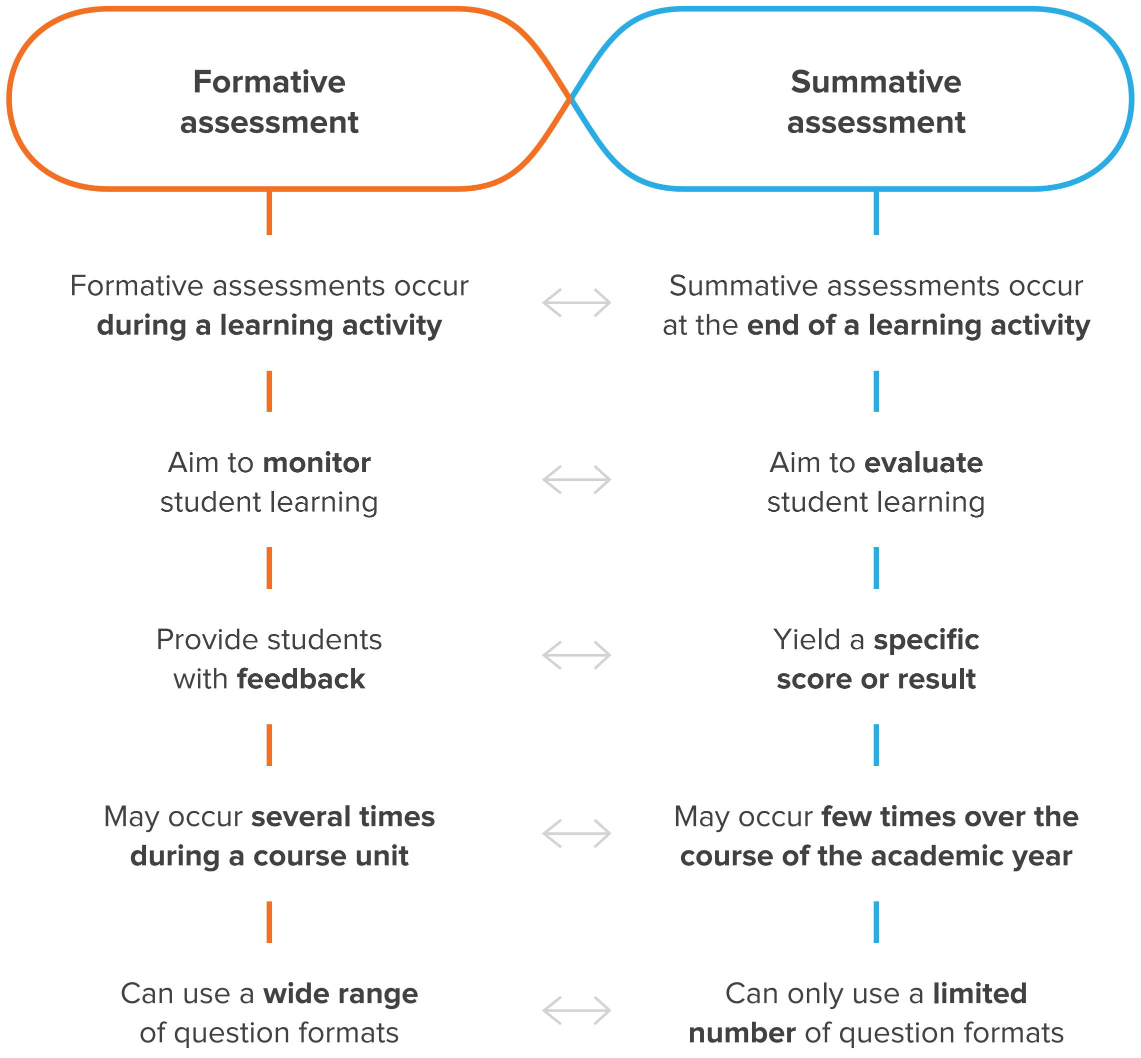
The particular ***purpose*** of any specific assessment is pivotal for a range of decisions that need to be made in the assessment process so it must be decided and articulated at the outset.

There are many possible purposes for measuring learning outcomes. Broadly, these can be placed in two categories: formative and summative. These have different meanings in different contexts (e.g. in evaluations), but for learning assessments we refer to assessments that are designed to improve the teaching and learning of a group of students by providing useful information to the teacher as formative assessment. This is sometimes referred to as assessment for learning. Summative assessment measures what learning has taken place.

Assessments can be divided into other categories, such as high stakes or low stakes or sample-based or census-based. These describe important elements of the assessment design.

High stakes assessments are those that have significant impact for the student who takes the assessment. Examinations are a good example because the results determine which options are available to the student in future. Low stakes assessments are the opposite – the outcome makes little or no difference to the student who undertakes them. Rather, the outcome is reported at the population level to inform decisions. While these categories may be clear to test designers and administrators, they also need to be made clear to test takers as it is easy for low stakes assessment to impose unnecessary pressure on the test takers.

Sample-based assessments are administered to a carefully selected sample of students and the results are used to draw inference about the whole population of interest using statistical methods. By contrast, census-based assessments are administered to all students within the population of interest.



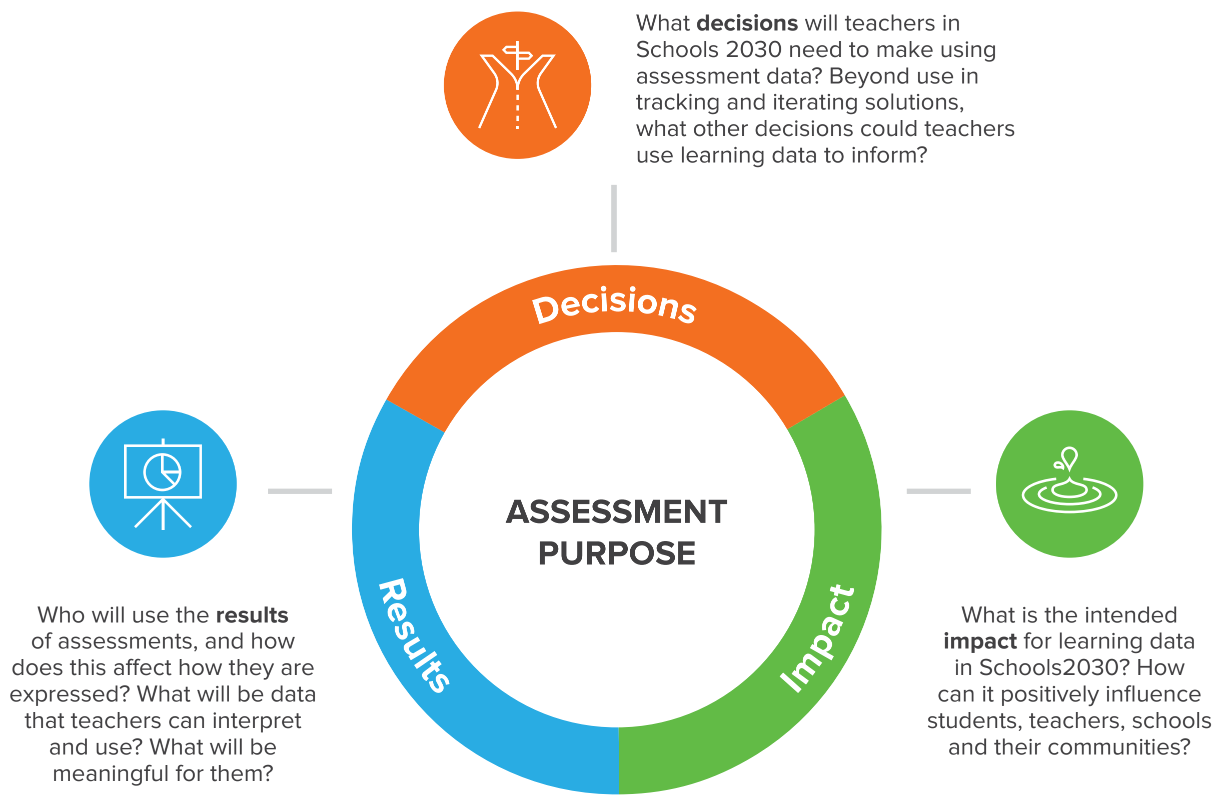
Formative assessment		**Summative assessment**
Formative assessments occur **during a learning activity**	⟷	Summative assessments occur at the **end of a learning activity**
Aim to **monitor** student learning	⟷	Aim to **evaluate** student learning
Provide students with **feedback**	⟷	Yield a **specific score or result**
May occur **several times during a course unit**	⟷	May occur **few times over the course of the academic year**
Can use a **wide range** of question formats	⟷	Can only use a **limited number** of question formats

Figure 3 Formative v. Summative Assessments

Defining Assessment Purpose for Schools2030

We can think of assessment purposes on three levels: result, impact and decision, all of which need to be considered by assessment designers and users.[1] There is a 'result' level – the assessment provides a judgement or description of learning outcomes. This could be expressed in terms of performance bands, pass or fail, percentile ranking, etc. The 'impact' level is concerned with the social or educational consequences of the learning assessment. These could be motivating learners, focusing attention on weak areas within the curriculum or raising political awareness of a problem or challenge. Finally, the 'decision' level of purposes are the decisions, processes and actions that are supported by the learning assessment. This may be programme-level monitoring and decision making.

[1] - Newton 2010 the multiple purposes of assessment



What **decisions** will teachers in Schools 2030 need to make using assessment data? Beyond use in tracking and iterating solutions, what other decisions could teachers use learning data to inform?

Who will use the **results** of assessments, and how does this affect how they are expressed? What will be data that teachers can interpret and use? What will be meaningful for them?

What is the intended **impact** for learning data in Schools2030? How can it positively influence students, teachers, schools and their communities?

Figure 4 Defining Purpose for Assessment

The central purpose of assessments in Schools2030 is to support teachers with designing, iterating and showcasing classroom-based solutions. When designing assessments this central purpose can be clearly, and contextually defined using the three levels of assessment purpose outlined above. Considering these areas will help understanding of how assessments can be designed to produce data which is most useful for teachers.

2.2 Opposition to learning assessment and the need for caution

There has been a fair degree of opposition to learning assessments from a range of education actors and stakeholders. Three common complaints are the incompleteness of learning assessments, the effects on teachers and teaching and the stress imposed on learners. We will look at each a little more below, but the problems that people identify are real.

In our view the existence of these problems does not undermine the importance and value of learning assessments but demonstrates that assessments can do harm as well as good. Rather than denying that these are risks or problems that occur, we need to recognise the risks, do everything possible to mitigate them and cease assessment processes when the risks become too great to justify potential benefits.

Learning assessments fall short of measuring all of the objectives of education and therefore can shift attention away from outcomes that are important but which are not assessed. It is difficult, or even impossible, to reach consensus about what education should achieve for learners nor is it possible to assess all intended outcomes equally reliably. Whereas we may aim to measure everything that is important, we measure what is measurable. Therefore, there is a risk that outcomes that are more difficult to measure are side-lined in favour of measurable outcomes. This problem needs to be born in mind by assessment users. The measurement of learning outcomes provides useful information, but it ought to be considered and used alongside information about other important outcomes that may be gathered using qualitative methods.

Related to this, learning assessments can have undue influence over what is taught. Rather than simply measuring learning outcomes resulting from teaching practices, assessments can shape what teachers do, both in terms of the content and the way that it is taught. Teachers can narrow their focus and exclusively teach the contents of the test. They may also adjust the way that they teach to more closely resemble the format of the test (e.g., using lots of multiple-choice questions in preparation for a multiple-choice assessment).

These adjustments can also undermine the ability for test users to draw correct inferences and make good decisions. Some of the assumptions that underpin the design of the assessment (such as the sampling of the target construct to select the assessed construct, see above) can be undermined if teachers focus only on the assessed construct. It should be emphasised that these adjustments from teachers are a natural response and should not be viewed as a 'fault'. Rather, they are potential consequences that need to be counteracted and prevented.

Finally, many are concerned about the impact of testing on learners, particularly in a context where they are being tested more often and in greater volume. This too, is an important consideration for test designers and administrators.

Assessments need to have a proper rationale and the benefits need to outweigh the costs. All care should be taken to minimise the emotional impact of assessments and avoid any harm.

3 How do learning assessments tools measure what children know and can do?

Summary

- The assessment process progresses in stages from construct definition, tool development, administration through scoring, aggregation, extrapolation and evaluation to decisions and impact.
- The first important question in measuring learning outcomes is 'what learning do we want to measure?'
- It is often not possible to assess the whole of a target domain or construct so it is necessary to select parts from the target domain to include in the assessed domain.

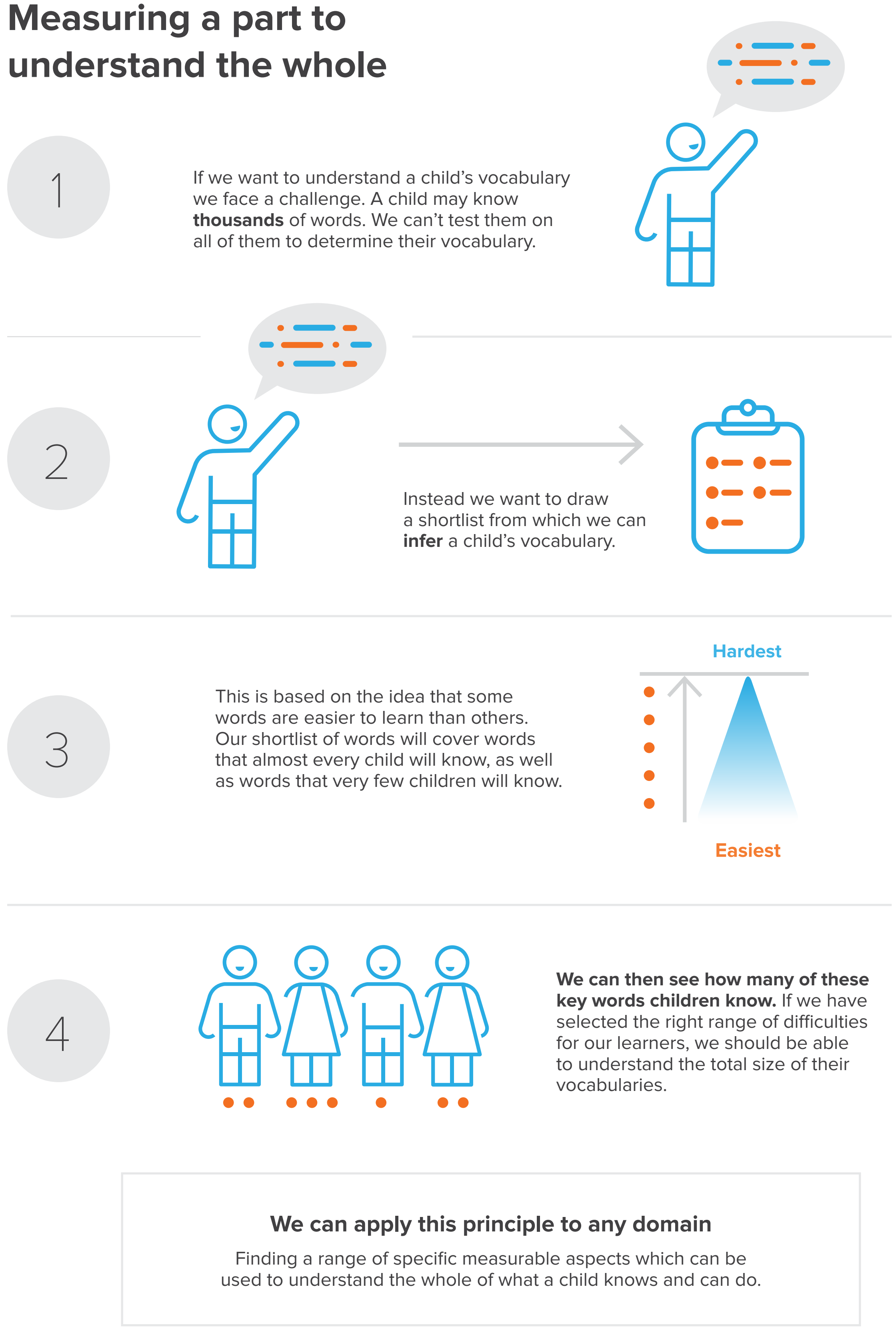
Learning outcomes are a latent trait – they are skills, knowledge, values and attitudes that are possessed by a person but that are not visible. Whereas other traits are visible or easily measured (such as weight and height), latent traits cannot be measured directly.

Instead, latent traits need to be measured by eliciting visible evidence and using that to measure the skills, knowledge, values and attitudes possessed by a person. For learning assessments, this takes the form of setting tasks or questions. These can be multiple choice questions, short answer questions or long answer questions. Alternatively, they can be oral questions, group tasks or creative tasks.

It is often not possible to assess the whole domain. Therefore, parts of the domain are assessed, and these are selected to be representative of the wider domain. This sample from within the target domain is referred to as the assessed domain. Koretz (2009) provides the example of a vocabulary test. It is not possible to test the many thousands of words a candidate may know so it is necessary to select a short list that enables inference to be drawn about the size of a person's vocabulary. This draws on a theory of learning that understands that broadly, words are learnt in some order so that it is possible to say that some words are 'easier' than others. Indeed, teachers who have taught children within a similar context would have a reasonable degree of consensus about which words are 'easier' than others. From this, words must be selected at the appropriate level for the range of candidates taking the test. The words should also enable test designers to distinguish between people with a larger or smaller vocabulary. This is referred to as discrimination, which is discussed in more detail in the academic skills handbook.

If the words are selected carefully, it is possible to make inference about a large domain (vocabulary) from a small sample of words. Similarly, for other domains, it may not be possible to assess the whole domain, but samples of it can be used to draw inference about the whole domain, based on an understanding of sequences of learning. This is equivalent to the way that we can draw samples of children and draw inferences about the population of children.

Measuring a part to understand the whole



1

If we want to understand a child's vocabulary we face a challenge. A child may know **thousands** of words. We can't test them on all of them to determine their vocabulary.

2

Instead we want to draw a shortlist from which we can **infer** a child's vocabulary.

3

This is based on the idea that some words are easier to learn than others. Our shortlist of words will cover words that almost every child will know, as well as words that very few children will know.

4

We can then see how many of these key words children know. If we have selected the right range of difficulties for our learners, we should be able to understand the total size of their vocabularies.

We can apply this principle to any domain

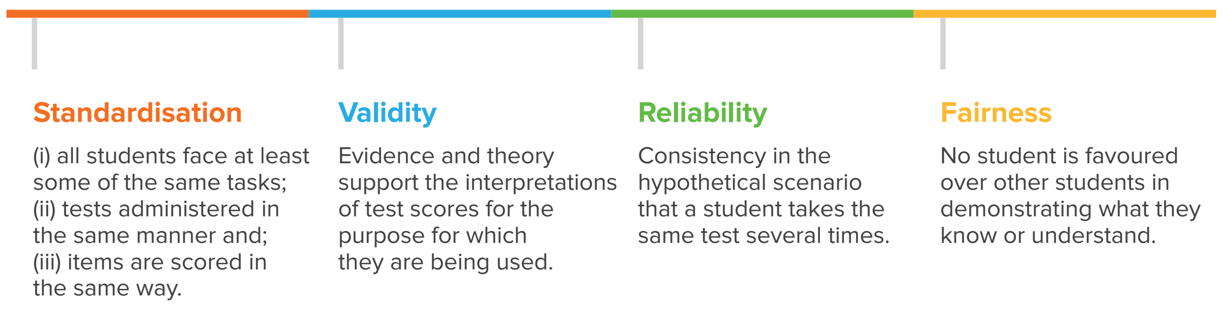
Finding a range of specific measurable aspects which can be used to understand the whole of what a child knows and can do.

Figure 5 Making Inferences about Learning

4 Important concepts in learning assessment

This section describes the core concepts for learning assessments. It also discussed the importance and centrality of teachers, which is a particular focus for Schools2030.

DEFINING RIGOUR IN ASSESSMENT



Standardisation

(i) all students face at least some of the same tasks; (ii) tests administered in the same manner and; (iii) items are scored in the same way.

Validity

Evidence and theory support the interpretations of test scores for the purpose for which they are being used.

Reliability

Consistency in the hypothetical scenario that a student takes the same test several times.

Fairness

No student is favoured over other students in demonstrating what they know or understand.

Consideration of teachers

Teachers are key agents in achieving impact of assessment, have a wealth of experience to contribute to assessments and are significantly affected by disruptions caused by assessments.

Figure 6 Summary of Contributors to Rigour

Standardisation

Standardisation is the process of ensuring that
(i) all students face at least some of the same tasks;
(ii) tests are administered in the same manner and;
(iii) items are scored in the same way.

Validity

Validity is a critical concept in assessment and is closely related to the concept of fitness for purpose. Validity refers to the degree to which evidence and theory support the interpretations of test scores for the purpose for which they are being used.[2] The validity of an assessment is dependent on the validity of each stage in the process. Key threats to validity are: (i) Construct Irrelevant Variance (CIV) - this is where factors other than the construct of interest affect the scores that test takers achieve and; (ii) Construct Under-Representation (CUR) – this is where elements of the domain are not given due prominence within the final test score.

Reliability

Reliability is the consistency of an assessment score in the hypothetical scenario that a student takes the same test several times.

Fairness

Fairness is the ability of an assessment to ensure that no student is favoured over other students in demonstrating what they know or understand. This is where two students with the same level of proficiency in particular domain are able to achieve the same score. For example, a child in a rural setting may not score well in a task that assumes knowledge of living in an urban area. Differential Item Functioning (DIF) can be used to test for this. DIF is a psychometric analytical method to test if particular items are biased towards particular groups.

It is important to **consider teachers** in the assessment process as they: (i) are key agents for change and crucial actors in achieving the impact from assessments; (ii) are significantly affected by the disruption and pressure imposed by assessments and; (iii) have a wealth of understanding about students, teaching and learning practices and challenges that can strengthen the assessment process.

[2] - Standards 1999

4.1 Assessment purposes

Summative assessments need to be standardised. Standardisation of assessments ensures that everyone who takes the assessment does so under the same conditions so that the results have the same meaning for all test-takers. **All should:**

- Face the same tasks
- Be administered in the same manner, and
- Be scored in the same way

This is important because it avoids irrelevant factors (anything other than the test-takers' skills, knowledge and/or values of interest) which could affect test scores and distort inferences drawn from the results. For the vocabulary test example, if test-takers are given lists of different difficulty or if some are able to see the word lists ahead of the assessment, it might be inferred that those with easier word lists or prior sight of the word lists have a larger vocabulary than they really do. Standardisation avoids these discrepancies between test takers.

Note that in some cases, standardisation can entail providing allowances for some students, such as additional time for students who need it and adaptions or additional resources for students with physical disabilities (e.g., braille test papers, scribes, etc.). This is discussed more in section 4.5.

Standardisation of Assessments in Schools2030

During the development of assessments for Schools2030, partners should consider what standardisation of the three points above means in their context. In particular, they should consider whether there will be learners who need additional resources to allow them to take part in assessments. As assessments will be administered by teachers, it will be of particular importance that clear instructions and a simple approach to scoring are used, to ensure standardisation across the second two areas.

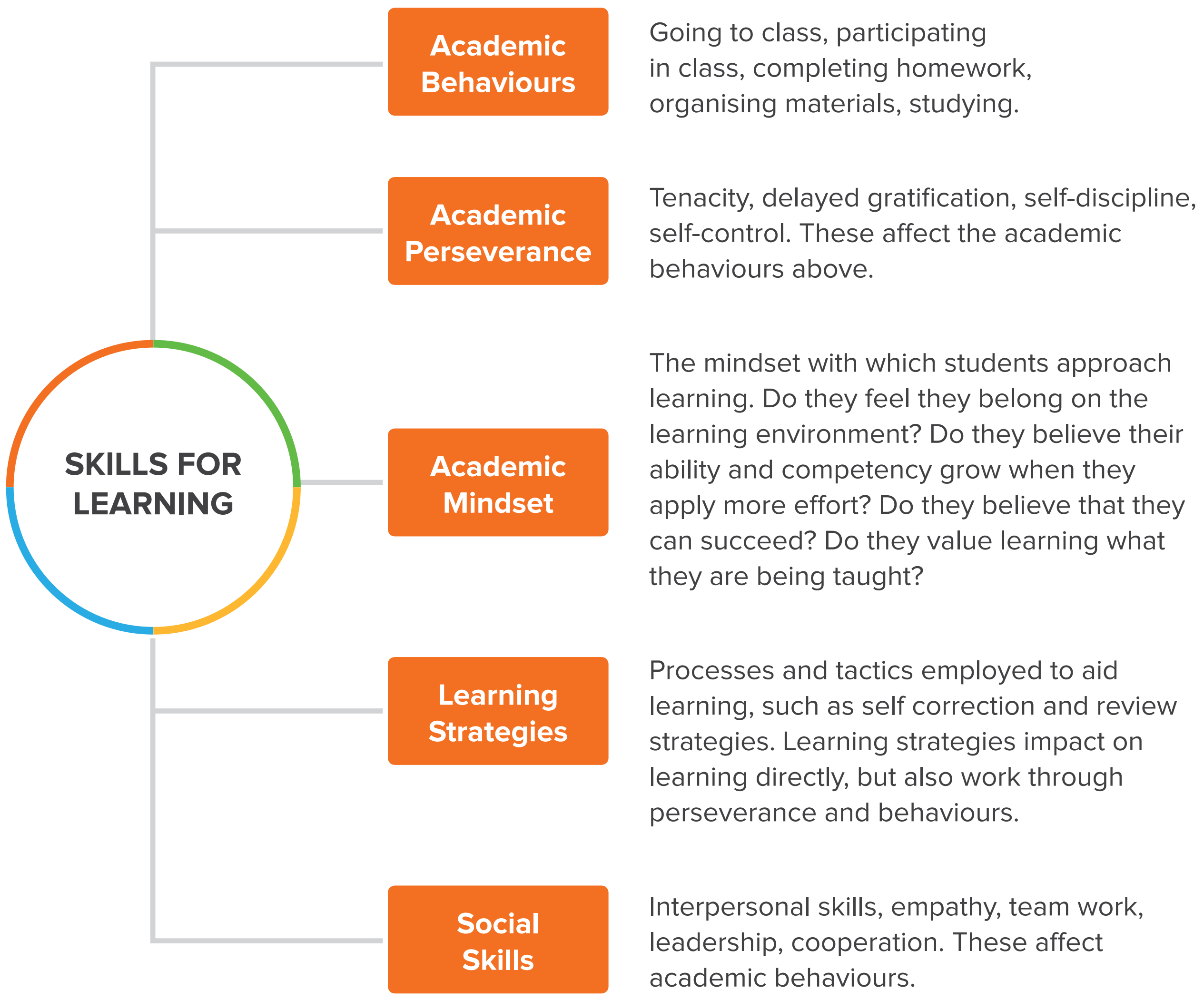
4.2 Constructs

A construct is a concept or topic for study that a learning assessment measures, often containing a number of elements. Schools2030 categorise constructs in two groups: academic and non-academic.

Academic constructs are those traditionally detailed in national curricula and assessed in standardised tests. They normally fit within a subject area. They include literacy, numeracy, football skill, human anatomy, understanding of Islam, etc.

Non-academic constructs cover a wide range of skills, strategies, attitudes and behaviours. They do not fit within any one traditional academic subject and are often more nebulous and more difficult to measure. Examples include creativity, problem solving, persistence and teamwork.

Non-academic constructs include concepts often referred to as non-cognitive skills and 21st Century skills. These have been conceptualised in a variety of ways. For example, Farrington et al. (2014) categorise skills for learning (a subset of the non-academic constructs) into five groups ordered according to their impact on school performance.



SKILLS FOR LEARNING

Academic Behaviours: Going to class, participating in class, completing homework, organising materials, studying.

Academic Perseverance: Tenacity, delayed gratification, self-discipline, self-control. These affect the academic behaviours above.

Academic Mindset: The mindset with which students approach learning. Do they feel they belong on the learning environment? Do they believe their ability and competency grow when they apply more effort? Do they believe that they can succeed? Do they value learning what they are being taught?

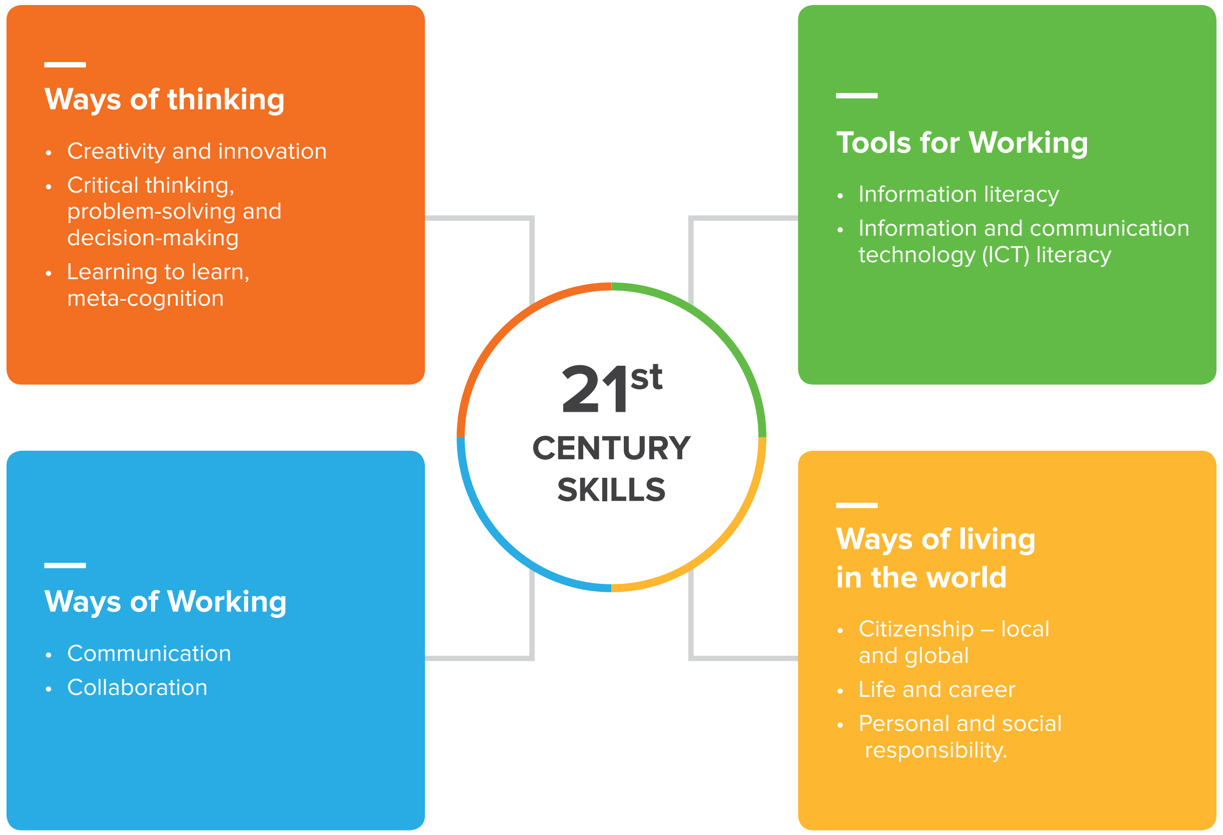
Learning Strategies: Processes and tactics employed to aid learning, such as self correction and review strategies. Learning strategies impact on learning directly, but also work through perseverance and behaviours.

Social Skills: Interpersonal skills, empathy, team work, leadership, cooperation. These affect academic behaviours.

Figure 7 Framework of Skills for Learning

Gutman and Schoon (2013) provide a broader conceptualisation that also includes meta-cognition and creativity. Although these might sometimes be counted as cognitive skills, they fit within the non-academic categorisation. Meta-cognition is thinking about one's thinking. It is a student's ability to plan and monitor their own understanding, which requires deep awareness of their thinking and learning as well as the ways in which they think and learn. Creativity is the ability to develop new and useful ideas. It requires the ability to combine information and ideas.

There is not complete consensus about the definition of 21st Century skills either. However, they are generally understood to include four categories[3]:



21st CENTURY SKILLS

Ways of thinking
- Creativity and innovation
- Critical thinking, problem-solving and decision-making
- Learning to learn, meta-cognition

Tools for Working
- Information literacy
- Information and communication technology (ICT) literacy

Ways of Working
- Communication
- Collaboration

Ways of living in the world
- Citizenship – local and global
- Life and career
- Personal and social responsibility.

Figure 8 Framework of 21st Century Skills

From these definitions of non-cognitive skills and 21st Century skills, it is clear that there is significant overlap, but with different conceptualisations and categorisations. 21st Century Skills include the ways of thinking, which are sometimes counted as cognitive rather than non-cognitive. Non-academic skills can include all constructs included in either non-cognitive skills or 21st Century Skills.

[3] - ATCS. www.atc21s.org

Non-Academic Skills in Schools2030

For Schools2030, 27 proficiencies and competencies have been defined (see Fig 9). These are divided into four categories, aligned to the OECD's skills for 2030 learning compass. **The four areas are:**

Knowledge

Knowledge encompasses the established facts, concepts, ideas and theories about certain aspects of the world. Knowledge usually includes theoretical concepts and ideas as well as practical understanding based on the experience of having performed certain tasks.[4]

Skills

Skills are the ability and capacity to carry out processes and to be able to use one's knowledge in a responsible way to achieve a goal.[5]

Attitudes

Attitudes are underpinned by values and beliefs and have an influence on behaviour. It reflects a disposition to react to something or someone positively or negatively and attitudes can vary according to specific contexts and situations.[6]

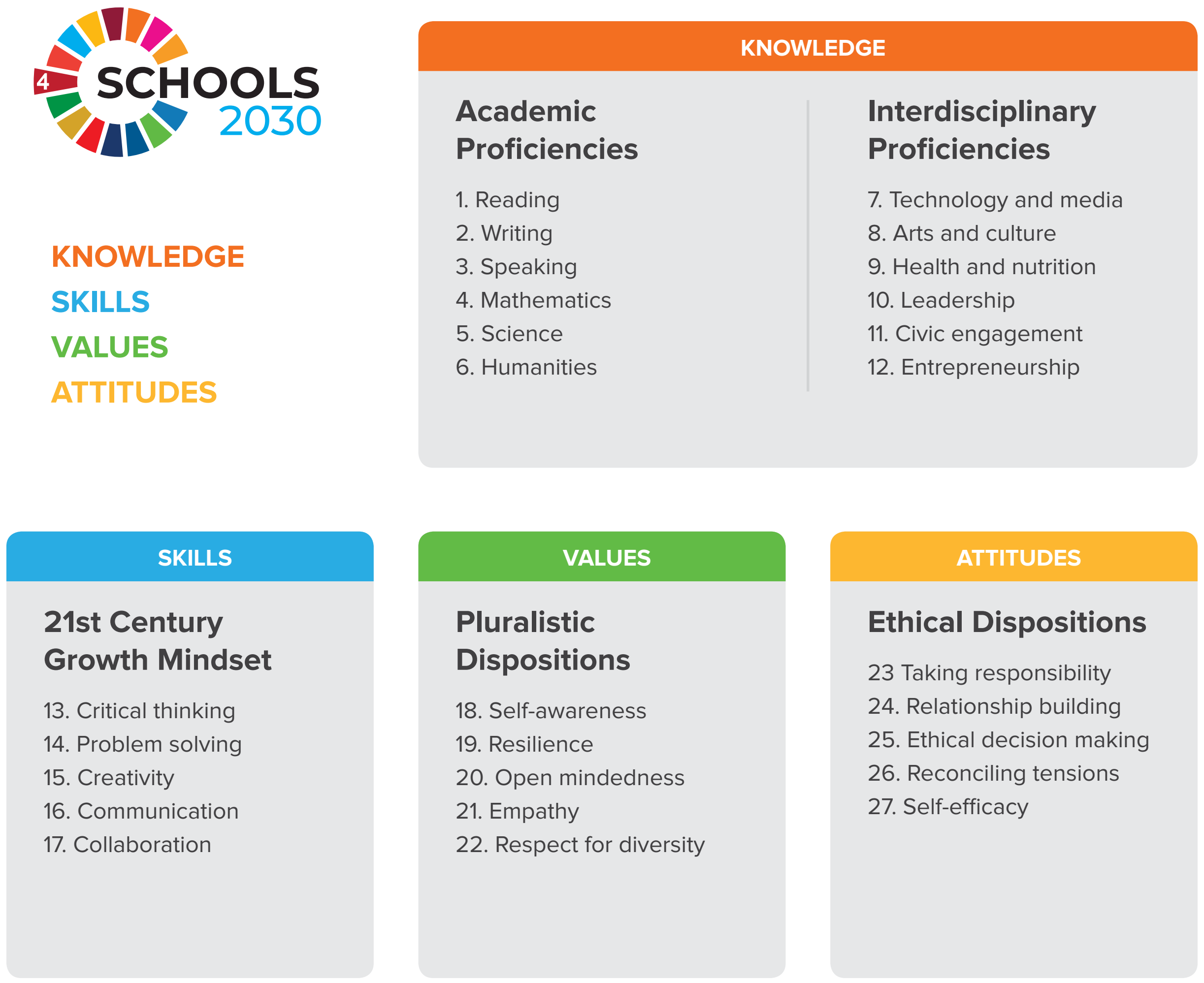
Values

Values are the guiding principles that underpin what people believe to be important when making decisions in all areas of private and public life. They determine what people will prioritise in making a judgement, and what they will strive for in seeking improvement.

[4] - See OECD concept note on Knowledge: https://www.oecd.org/education/2030-project/teaching-and-learning/learning/knowledge/Knowledge_for_2030_concept_note.pdf

[5] - See OECD concept note on Skills: https://www.oecd.org/education/2030-project/teaching-and-learning/learning/skills/Skills_for_2030_concept_note.pdf

[6] - See OECD concept note on Values and Attitudes: https://www.oecd.org/education/2030-project/teaching-and-learning/learning/attitudes-and-values/Attitudes_and_Values_for_2030_concept_note.pdf



KNOWLEDGE

Academic Proficiencies

1. Reading
2. Writing
3. Speaking
4. Mathematics
5. Science
6. Humanities

Interdisciplinary Proficiencies

7. Technology and media
8. Arts and culture
9. Health and nutrition
10. Leadership
11. Civic engagement
12. Entrepreneurship

SKILLS

21st Century Growth Mindset

13. Critical thinking
14. Problem solving
15. Creativity
16. Communication
17. Collaboration

VALUES

Pluralistic Dispositions

18. Self-awareness
19. Resilience
20. Open mindedness
21. Empathy
22. Respect for diversity

ATTITUDES

Ethical Dispositions

23 Taking responsibility
24. Relationship building
25. Ethical decision making
26. Reconciling tensions
27. Self-efficacy

Figure 9 Schools2030 Holistic Learning Framework

While these 27 domains were selected at the global level, the core of the selection process occurs at the national level. Schools2030 countries, as part of the assess phase, go through a rigorous and participatory process of selecting and defining domains of relevance for each country and age cohort. While the names of domains may be taken from the 27 suggested here, the definitions and constructs will be heavily influenced by content. This is crucial to ensure that what is measured reflects the understanding of, and importance given to, different domains in each country.

4.3 Validity

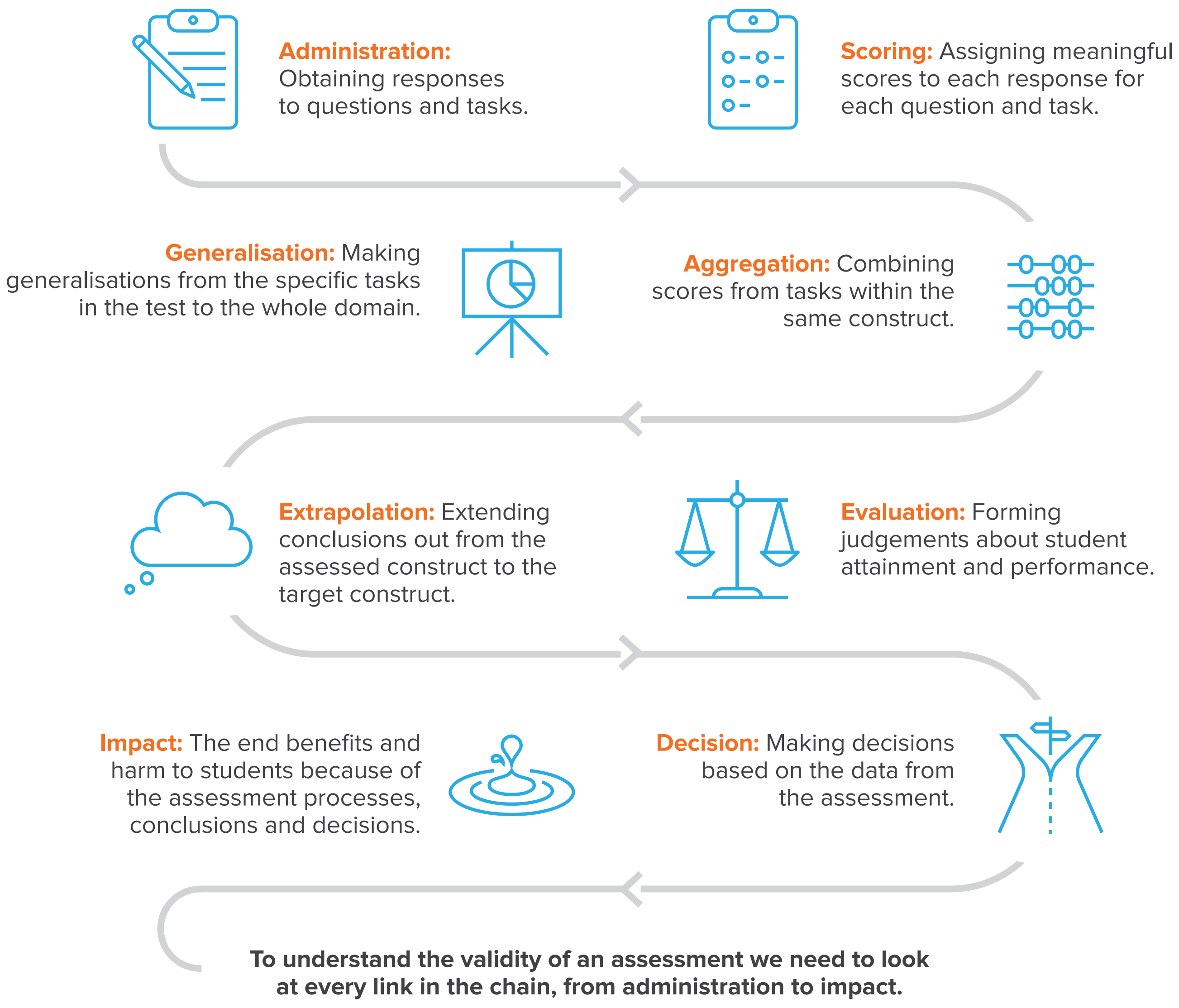
Validity is a critical concept in assessment and is closely related to the concept of fitness for purpose. "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests."[7]

[7]- Standards 1999

It “is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.” [8]

The validity judgement is centred on the purpose of the assessment – what is it intended to be used for? The judgement is an integrated evaluative judgement, so it takes into account a range of evidence to arrive at a judgement. It is not a single statistic or objective score. Rather, it requires drawing on the statistical evidence, assessment and learning theory and experience and expertise.

Examining Validity of Assessments



Administration: Obtaining responses to questions and tasks.

Scoring: Assigning meaningful scores to each response for each question and task.

Aggregation: Combining scores from tasks within the same construct.

Generalisation: Making generalisations from the specific tasks in the test to the whole domain.

Extrapolation: Extending conclusions out from the assessed construct to the target construct.

Evaluation: Forming judgements about student attainment and performance.

Decision: Making decisions based on the data from the assessment.

Impact: The end benefits and harm to students because of the assessment processes, conclusions and decisions.

To understand the validity of an assessment we need to look at every link in the chain, from administration to impact.

Figure 10 Examining Validity of Assessments

[8] - Messick

Evaluating validity requires examination of the whole process of learning assessment design from item development through to analysis, drawing inference and using the data in decision making. All elements need to be considered against a clearly defined purpose. Crooks, Kane and Cohen (1996) liken the stages of assessment laid out above (they describe eight stages) to links in a chain. Each link needs to be examined in turn and any weaknesses in any links result in weaknesses to the whole chain.

A chain is only as strong as its weakest link and an assessment is only as valid as the weakest stage in the process. Crooks, Kane and Cohen identify illustrative threats to validity at each step of the process.

Administration

Obtaining responses to questions and tasks.

Threats to validity:

- Low motivation of students to do as well as they can in the assessment.
- Anxiety about the assessment, which prevents students from performing as well as they could.
- Inappropriate assessment conditions – there are factors in the administration (distractions, unclear instructions, etc.) which cause errors that are not related to the student’s attainment. Alternatively, factors that artificially inflate their score, such as coaching by administrators or posters on the wall.
- Task or response not communicated – failure to complete the task is counted as an inability to do so when it was actually caused by ineffective communication or the student’s failure to communicate their response.

Scoring

Assigning meaningful scores to each response for each question and task.

Threats to validity:

- Scoring fails to capture important qualities of task performance.
- Undue emphasis on some criteria, forms or styles of response. The scoring may place too much significance on elements of responses that have little or no relevance to the target construct (e.g., spelling, grammar and handwriting where this is not the construct).
- Lack of intra-rater and/or inter-rater consistency. Scorers scores are not consistent through time so that they do not reward the same score to answers of equal quality (intra-rater consistency). Alternatively, different scorers award different scores to answers of equal quality (inter-rater consistency).
- Scoring is too analytical (or granular) or too holistic. Scoring that is too analytical would divide the response into too many aspects so that the aggregated score does not reflect the overall quality of the response. Conversely, scoring that is too holistic ignores useful information about performance in the construct.

Aggregation

Combining scores from tasks within the same construct.

Threats to validity:

- Aggregated tasks are too diverse. If the tasks are not related to each other, their scores will not be sufficiently correlated, which will harm all future steps in the process.
- Inappropriate weights given to different aspects of performance. When the construct is divided into constituent parts (sub-domains), each part should have a reasonable weight in the total score to avoid the score being skewed towards less significant sub-domains.

Generalisation

Making generalisations from the specific tasks in the test to the whole domain of similar questions and tasks.

Threats to validity:

- Conditions of assessment too variable so that factors outside of a student's attainment (e.g. time allowed for completion of tasks, time of day, materials available) have significant influence of individual scores.
- Too few tasks so that the selection of questions from all possibilities has significant influence over a student's score.

Extrapolation

Extending conclusions out from the assessed construct to the target construct.

Threats to validity:

- Conditions of assessment are too constrained so that students may not be expected to perform to a similar level if they faced tasks in another format.
- Parts of the target construct are not assessed or are given little weight. If there are parts of the construct that are not assessed and performance in these areas is not related to the parts that are assessed, conclusions can only be made about the assessed construct and not about the target construct.

Evaluation

Forming judgements about student attainment and performance.

Threats to validity:

- Poor grasp of assessment information and its limitations. The evaluator needs to understand what information the assessment provides and where its limits are.
- Inadequately supported construct interpretation. There can be a temptation to over-state the conclusions or to make leaps that are not supported by the evidence or theory.
- Biased interpretation or explanation. Often, evaluators can carry an agenda, or pre-conceived ideas or may have an incentive to report certain findings. These undermine validity.

Decision

Making decisions based on the data from the assessment.

Threats to validity:

- Poor decision-making can occur for a range of reasons. It could be that the conceptual framework in which decisions are made is incorrect or weak. Alternatively, evidence can be used to support pre-existing agendas even when they should not. Decisions may not take account of other factors or evidence to a sufficient degree.
- Decision makers may not use the assessment data.

Impact

The end benefits and harm to students because of the assessment processes, conclusions and decisions.

Threats to validity:

- Positive consequences not achieved.
- Serious negative impact occurs.
- The risks and assumptions involved in achieving desired consequences and avoiding harm should be considered and mitigated. If they cannot be mitigated to a sufficient degree it may be appropriate to cancel the assessment.

The decision-making link in the chain can be a stage at which many assessments fall down. It needs to be considered and planned for from the beginning. Trust in and familiarity with the assessment need to be developed by including decision makers in the process and engaging with them to determine what they need and what would increase their trust. People are more likely to use data they know and understand, and they are more likely to trust assessments when they know how they have been designed to meet their requirements. It may improve the validity of an assessment to make concessions that weaken the technical strength of the instrument but improve the trust placed in it.

The table above illustrates particular threats to validity. Broadly, we can think of two key threats to validity: Construct Irrelevant Variance (CIV) and Construct Under-Representation (CUR).

The first (CIV) is the threat that factors other than the construct of interest – the skills, knowledge and values being assessed – affect the scores that test takers achieve. That is, two students with the same ability in the construct (e.g., reading) could achieve different scores because of other differences between them. This can occur in many ways.

For example, if a test uses stimuli that draw on experiences that are familiar to some children and not others, it will likely favour those who are familiar. A test might use examples from farming and agriculture to test literacy skills. This would favour children from rural areas or farming backgrounds for whom the concepts and vocabulary would be more familiar. Alternatively, test questions that closely resemble the style and layout of questions and instruction provided in schools would favour children who have learnt in that setting over children who have developed the skills in other countries or in non-school settings. Outside of test development, other factors that could affect test scores include differences in the settings for tests (children taking them in quieter settings may perform better) or issues with marking (markers may be more generous for children with better handwriting, when this is not part of the construct).

The second key threat, CUR, is that elements of the domain are not given due prominence within the final test score. Once the domain is defined, the test design and scoring methodology should reflect all sub-domains within and the relative weights given to it. Failure to do so means that the assessment does not measure what it purports to.

Ensuring the Validity of Schools2030 Assessments

The assessments produced for Schools2030 will be used directly by teachers to design and showcase solutions for their classrooms. To ensure that the solutions will be as effective as possible, it is essential that the assessments used are providing a valid description of learning levels.

While the table in this section gives generic threats to validity across the cycle of assessment, specific risks should be considered. The table below can be used as a template for identifying risks and mitigation strategies across the process. For each of the headings the team developing the assessment should write in specific risks, and strategies to minimise them. These can be both strategies in the development of tools, but also in the development of guidance materials and training approaches for teachers. This is important as it will be teachers who will complete many of these steps, particularly in the administration of assessments, and the analysis of results.

SECTION	THREATS	MITIGATION STRATEGY
Administration		
Scoring		
Aggregation		
Generalisation		
Extrapolation		
Evaluation		
Decision		
Impact		

4.4 Reliability

The idea of reliability is related to validity. Indeed, it is often considered to be part of validity as an unreliable assessment limits the validity of inferences drawn. "If a student attempts a test several times, even if no learning takes place, the student will not get the same score each time – the student might not feel very 'sharp', the marker may be more or less generous, or the handwriting might be a little bit clearer so the marker can understand the answer." [9] Reliability is the consistency of an assessment in this hypothetical scenario.

Section 3 describes how test questions are drawn from a larger set of all possible questions. This question selection is a source of inconsistency and therefore a threat to reliability if the questions do not perform similarly (i.e., students of a similar level of attainment are not equally likely to provide the correct answer).

Less reliable assessments provide less precision so that for a student with a given level of attainment (unobserved) may score within a wide range either side of that true attainment. Expressed the other way, a student with a given assessment score may have an actual level of attainment within a wide range of that value. Lower precision makes it more difficult to draw inference as observed differences could be a result of the assessment's imprecision rather than differences in the true attainment of student.

Reliability of Assessments in Schools2030

A core principle of Schools2030 is to place the power for assessment in the hands of teachers. This means that Schools2030 teachers will become the determining factor for the reliability of assessments developed for Schools2030.

In this context, what can be done to maximise both the inter-and intra-rater reliability of assessments?

To maximise the reliability of both inter- and intra-rater assessments the clarity of instructions and training materials should be considered. It is important to remember that teachers are (in many cases) not professional enumerators or test scorers. Understanding teachers' previous experience and expertise in assessment will help determine how to best approach developing instructions for assessments to ensure as much consistency in how tests are administered and scored as possible. Materials should be accompanied by a thorough training process for teachers, to help them understand their role in the assessment process.

The analysis of pilot data will be essential for reflecting on the reliability of assessments. It may be considered useful to accompany the piloting of tools with some qualitative data collection on teachers' experiences of delivering and scoring assessments, as this may help to diagnose any areas of confusion which are causing reliability issues.

[9] - William (2001) *Reliability, validity and all that jazz.*

4.5 Bias and fairness

Learning assessment fairness is a technical question related to validity, but it also considers the wider impacts of the assessment on societal equality. All students need to have the same chance to demonstrate their ability and the assessment should not harm the students or have outcomes that are disadvantageous to the society.

“A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. It would consider the technical properties of tests; the ways test results are reported and the factors that are validly or erroneously thought to account for patterns of test performance for groups and individuals.” [10]

There are individual and group dimensions to fairness.[11] Individually, if a test is fair, no student is favoured over other students in demonstrating what they know or understand. Students must face standardised conditions (see Standardisation section) and be treated comparably. Comparable treatment does not necessarily require that all students be treated equally. Rather, some students will have recognised difficulties that prevent them from demonstrating their attainment under the same conditions. For their scores to be comparable to other students’, accommodations may be required to enable them to demonstrate their attainment. These may include additional time, translated materials, additional or alternative materials (such as braille test papers). The aim is always to make assessment results comparable and ensure that no student is advantaged or disadvantaged relative to the rest of the cohort.

“Another aspect of individual fairness involves treating test takers with dignity and sensitivity”[11] . The assessment must be conducted ethically, with due consideration for the welfare of all students involved. Assessment instruments should be reviewed to ensure that there is no content that may be culturally offensive, support negative stereotypes or cause emotional or psychological difficulties for some students. This may entail some difficult decisions. For example, the prevalent gender roles in a culture may be that women tend to stay home and look after the house and children while men are more likely to work. Therefore, there may be a validity argument that these gender-norms should be reflected in an assessment so that students are not confused by gender roles with which they are not familiar. However, it could be argued that children need to see that gender norms are not necessarily destiny for them and that girls could work and/or boys could be the main care-giver when they become adults. Representing men in work situations and women in domestic situations may act to reinforce gender norms and discourage children from considering the alternatives. These are difficult judgements that need to be made carefully by people who understand and belong to the context.

Statistical methods to identify unfairness operate at the group level. Groups of interest need to be determined and defined such that the groups that may be treated unfairly are not combined with those that are not, potentially hiding the unfairness when results are analysed.

A threat to inter-group fairness that can be analysed statistically is Differential Item Functioning (DIF), which occurs where students in two groups, A and B, with the same attainment, have a different probability of answering a question correctly. For example, it may be that students who live in rural areas (group A) are more likely to correctly answer a question that uses rural farming scenario as a stimulus than students who live in urban areas (group B) when comparing students with the same overall attainment.

Bias and Fairness for Assessments in Schools2030

As with other concepts outlined here, bias and fairness should be considered when developing assessments for Schools2030. In particular, it may be useful to map out different groups present in Schools2030 schools. Whether these are language groups, cultural or ethnic groups, or other groups defined by background characteristics there may be a biasing effect. Considering this during the initial development phase will avoid some adjustments that may be necessary after piloting.

It will also be apparent from the analysis of pilot data carried out by the Global Assessment Partner whether there are any items for which learner background characteristics have a biasing effect. These can then be examined and adjusted before implementation

4.6 Considering teachers in the assessment process

Teachers are central to all Schools2030 programming and should always be considered in assessment processes.

- • Teachers are key agents for change and crucial actors in achieving the impact from assessments;
- • Assessment processes can also disrupt their work, place pressure on them and misdirect their attention away from important issues;
- • Teachers have a wealth of understanding about students and teaching and learning practices and challenges that can strengthen the assessment process.

[10] - Srandards

[11] - Camilli 2013 Ongoing issues in test fairness

It is therefore important that due consideration is given to how to minimise negative consequences for teachers and how to make best use of their skills, knowledge and position as agents for change. Teachers can often find that policies and programmes are done to them, but assessment processes should be conducted for and with them.

This must be achieved in a way that neither hinders progress in learning assessments nor undermines the validity of the assessment for its intended purposes. Therefore, engagement is likely to stop short of consultation and co-creation in most cases.

Integrating teachers into assessments in Schools2030

As mentioned throughout this handbook, teachers are the key party involved in delivering assessments and interpreting their results. As mentioned above they should not necessarily be involved in co-creating assessment instruments, as this may bias the results. However, a process is in place to ensure that teachers' realities and interests are given importance during the process. **This includes:**

1. **Involvement of teachers in develop domain and construct definitions**. It is important that a range of voices are included in the definition of constructs to ensure that they reflect the contextual realities of learning. Teachers are a key voice in this process.
2. **Interpretation of assessment data.** Assessment data is a key input for the "explore" phase of the human centred design (HCD) process in Schools2030. Establishing teachers with the support needed to interpret the results of assessments will ensure that the solutions developed reflect both the results of assessments and teachers' understanding of the causes of learning levels.
3. **Teacher generated assessments.** As mentioned in the introduction, alongside the assessment tools that this handbook covers, there will also be space for teachers participating in Schools2030 to develop their own approaches to assessment. While the tools developed at the national level will be used at the beginning and end of each year, other approaches can be used during the year to track and iterate the solution. These can be driven by teachers to help them understand the effects of their classroom solution.

Together these approaches will ensure that the input of teachers is taken into account in Schools2030, and that their insights and understanding can complement the data generated through assessments.

Conclusions

This handbook has outlined a number of key questions and concepts which should underpin the development of any assessments, whether for academic or non-academic constructs. Before starting tool development, and throughout the development process the following questions should be considered:

Key Questions

1. Why is the assessment being conducted? What should it achieve in terms of the data it can provide and the decisions that it can help teachers making?
2. What are the constructs that we are aiming to measure? What are the observable or measurable constructs that can be used to represent learning?
3. Looking at the process of assessment, what are the factors that could influence the validity, reliability and fairness of the assessment results?
4. How will teachers be involved in the process? How can we ensure that the assessment process will not take away from their teaching work and will produce data that is relevant for them?

Bearing these questions in mind, the process of developing assessments should begin. Following this handbook are two others in the series.

The first will walk through the steps and considerations involved in developing assessments for academic constructs. The second will repeat the process with the steps and considerations unique to the development of assessments for non-academic constructs.

DESIGNING LEARNING ASSESSMENTS

SCHOOLS 2030

Oxford MeasurEd