# DESIGNING LEARNING ASSESSMENTS

## HANDBOOK 2

Designing Academic Assessments

# Contents


**1. Introduction 6**

**2. Assessment blueprint and planning 8**
2.1 Assessment purpose 9
2.2 Assessment users 10
2.3 Assessment design process and risks 10
2.4 Definition of construct and assessment content 12
2.5 Assessment specifications 13
2.6 Testing format and question specifications 14
2.6.1 Choosing a testing format 14
2.6.2 Developing Question Specification 15
2.7 Number of items and time constraints 16
2.8 Method of administration 16

**3. Developing items 17**
3.1 Validity in item development 18
3.2 Considering how students respond to test items 19
3.3 Types of items and item components 20
3.3.1 Constructed response questions: Short- and long-answer questions 20
3.3.2 Selected response: Multiple choice questions (MCQs) 23

**4. Test design and assembly 27**
4.1 Test design | Ceiling Effects | Construct Irrelevant Variance 27
4.2 Item analysis 30

**5. Conclusion 32**


# Figures


**Figure 1** | Assessment in Schools2030 6

**Figure 2** | Options for Sequencing of Assessment 11

**Figure 3** | Target versus Assessed Constructs 12

**Figure 4** | Example of a test specification from British Council 15

**Figure 5** | Rules for Selected Response Questions 25

**Figure 6** | Example of an Item Characteristic Curve 31


This handbook is second in a set of three designed to support learning assessment processes in Schools2030. Together, the purpose of the handbooks is to provide a common understanding of the way that learning assessments work and to provide guidance for Schools2030 partners to conduct effective assessments. These reference documents are designed to be used by Schools2030 National Assessment Partners, but may also be used by a range of stakeholders involved in various stages of the assessment cycle.

This handbook walks through the process of developing learning assessments of academic skills. It describes the main steps and discusses the main considerations. Throughout, validity, reliability and fairness are at the core of decision making.

# Summary

Where to start on designing a new assessment tool? This handbook will take you through the key three stages: planning, developing items, and test design and assembly.

1

## Assessment Blueprint and planning

Before we start we need to plan the way ahead. What should we consider before thinking about designing an assessment?

**Things to consider:**

Assessment Purpose
Assessment Users
Assessment design process and risks
Defining constructs and content
Assessment specifications
Test format and question specifications
Number of items and time constraints
Method of Administration

2

## Developing Items

The first step is to generate some items (questions). How do we ensure they will measure what we want to measure? What options do we have and what do they look like?

**Things to consider:**

Validity in item development
Considering how students respond to test items
Types of items and item components

3

## Test Design and Assembly

How do we bring it all together? What can we learn about items that will help us build a useful and usable assessment?

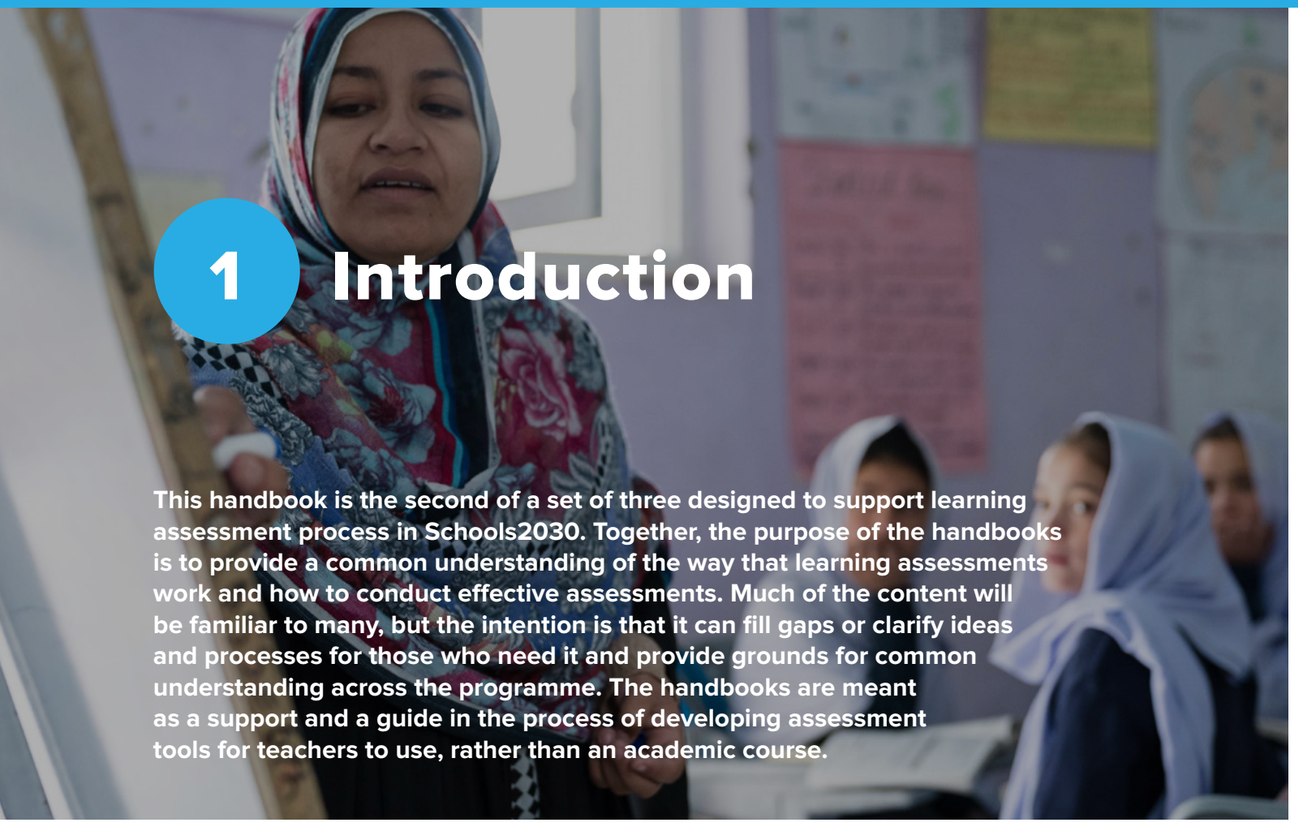
**Things to consider:**

Test Design
Item Analysis



# Key terms

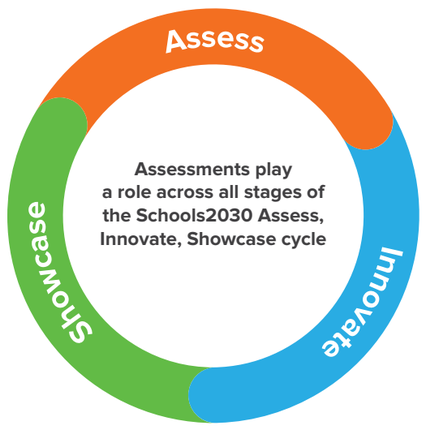
| TERM | DEFINITION |
|---|---|
| **Academic/Non-Academic Knowledge and Skills** | For the purposes of this document, academic skills refer to Academic Proficiencies in reading, writing, speaking, mathematics, Science and Humanities. Non-academic knowledge and skills refer to the range of knowledge, skills, values and attitudes outside of these limited academic subjects. For Schools2030 these include 21st Century growth Mindset skills, pluralistic dispositions (values) and ethical dispositions (attitudes). |
| **Item** | An item refers to one question in an assessment. The best approach to developing quality assessment is to use the information from piloting about individual items to construct a collection of items (assessment tool) that represents balance in the domain, an appropriate spread of item difficulty and items that perform well under psychometric analysis. |
| **Assessment/Tool** | Assessments/assessment tools and tools are a collection of items administered to measure an overall construct or domain. |
| **Constructed response** | An answer that the student has to produce, rather than selecting from options. E.g. short answer questions, essay questions |
| **Selected response** | An answer provided by selecting from given options. E.g. multiple choice questions. |
| **Pivot item** | A question or task used to assign students to assessment instruments of differing difficulties. |
| **Computer Assisted Personal Interviewing (CAPI)** | The use of tablets or mobile phones to collect data in the field. |
| **Item Response Theory (IRT)** | An applied statistical discipline used for learning assessments. |
| **Ceiling effects** | Ceiling effects occur where all items are too easy so that most students answer all or most correctly. This results in a lack of information about where students' ability ends. |
| **Floor effects** | Floor effects occur where all items are too difficult so that most students answer all or most incorrectly. This results in a lack of information about what students are able to do. |
| **Item discrimination** | The ability of an item to distinguish between students of higher and lower ability. |
| **Item Discrimination Index** | A score of item discrimination ranging from -1 to 1. |
| **Item Characteristic Curve** | A graph produced using IRT that provides useful information about items by plotting the probability that a student will answer the item correctly against the ability of the student. |
| **Stop rules** | Conditions for the assessment to end early if certain criteria are fulfilled (e.g. a set number of incorrect responses are provided). |



# 1 Introduction

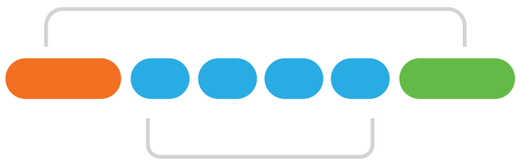
**This handbook is the second of a set of three designed to support learning assessment process in Schools2030. Together, the purpose of the handbooks is to provide a common understanding of the way that learning assessments work and how to conduct effective assessments. Much of the content will be familiar 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 in the process of developing assessment tools for teachers to use, rather than an academic course.**

Figure 2 outlines how teacher implemented, and teacher generated assessment tools and approaches are used across the Schools2030 programme 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

**The three handbooks are:**


**Core concepts**

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 process for assessing academic skills**

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

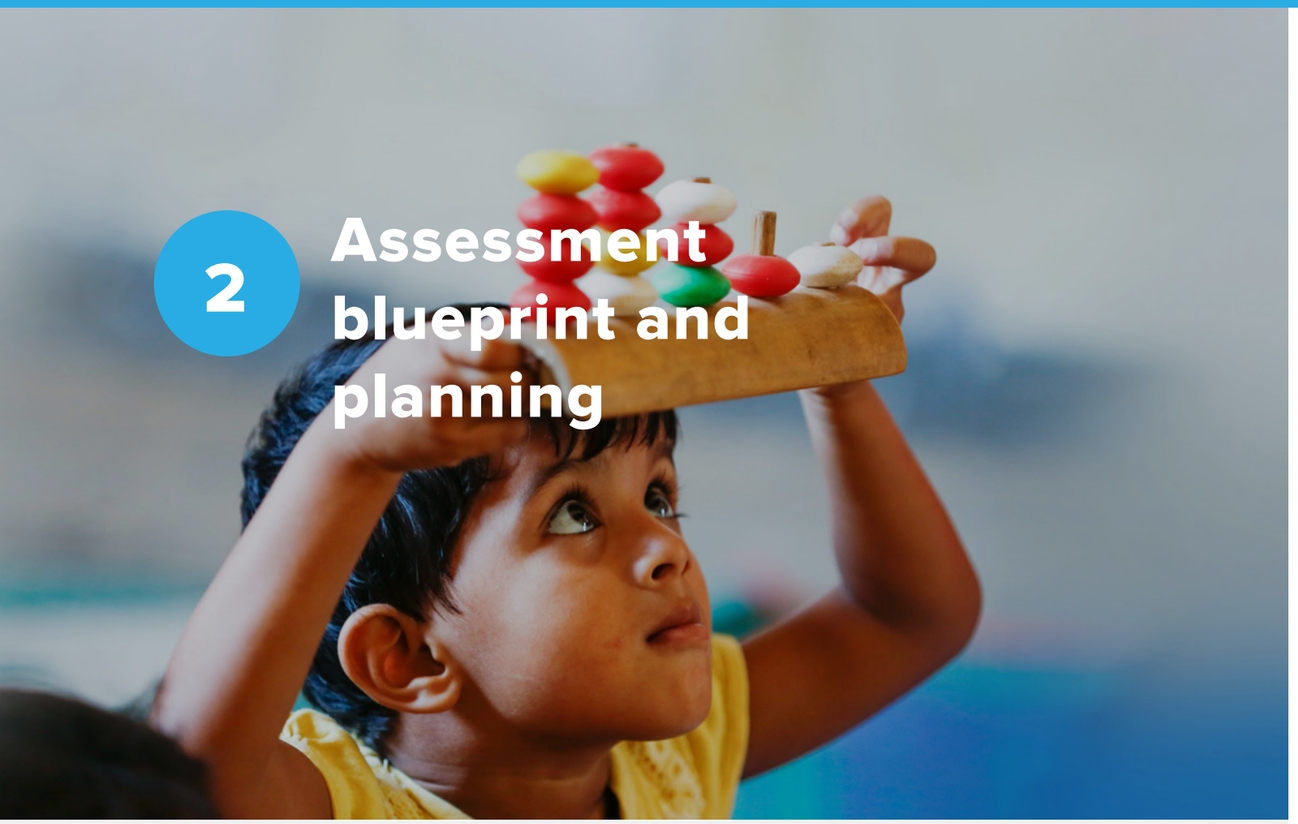
**The process for assessing non-academic skills**

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 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 in may be helpful to return 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 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 useful.**



# 2 Assessment blueprint and planning

## Summary

- All assessments need an overall plan.
- The purpose and users need to be clearly defined, the external constraints need to be established and key features of the assessment need to be determined.
- Most of the fundamental decisions about the assessment need to be made at the beginning, before test development starts in earnest. They guide and shape decisions later in the process.

## 2.1 Assessment purposes

The purpose of the assessment is the first decision. It shapes many decisions and trade-offs throughout the assessment process so it needs to be considered carefully and articulated clearly. A common error within learning assessments is to use assessments for purposes that they were not intended for and are not suitable for. Similarly, purposes can become confused during the assessment process when stakeholders want to add or change purposes and new problems arise. Such confusion can seriously undermine the validity and usefulness of the assessment and may result in wasted resources.

**To determine the purpose, consider:**

**What is the end impact or benefit from the assessment?**

**How will it help improve students' learning?**

**What decisions will be supported to achieve this impact?**

**When will these decisions be made?**

The core handbook provides more information on assessment purposes. It may be helpful to consider the three levels: result, impact and decision. The result is the appraisal of a student's learning attainment (e.g. pass or fail; attainment levels). The impact is the consequences of the assessment within education and society (e.g. changes in teaching practice and teacher behaviour, motivating students, political awareness). At the decision level, what decisions are intended to be supported and who makes the decisions.

## 2.2 Assessment users

The assessment users are the people who will make decisions (described in the purpose) using information from the assessment. They will look at the results and use them to shape their understanding of the situation, problems, priorities and what should be done next.

Assessment users should be identified based on the purpose. Once designers have determined how the assessment is going to improve learning, the assessment users become apparent. They are the people who need to make the decisions and/or change their practice based on the assessment results.

They should be considered throughout the process to ensure that they know about the assessment, understand what it does and what its limitations are, and that they trust the results that the assessment produces. This cannot wait until results are disseminated because building trust and familiarity takes a long time. Key users should be engaged at the very outset to ensure that the assessment does what they need and (just as important) that they believe that the assessment does what they need and is a reliable source of information.

## 2.3 Assessment design process and risks

**All the stages of the assessment development process should be planned with timelines, who is responsible and who is involved.**

The risks should be identified, and mitigation should be planned and monitored using a risk register. Risks include risks to the validity of the assessment – anything that could prevent the end impact from being achieved – and risks to the delivery of the assessment – factors that could prevent the assessment from being delivered on time and to budget.

The timing and frequency of the assessment need to be selected so that the assessment provides the information required at the time that is required. For example, to evaluate the impact of an intervention, it may be necessary to conduct a baseline assessment before commencement of the intervention (or programme) and an endline assessment after the intervention. Further assessments may be helpful between these rounds, particularly if the information is intended to assist programme implementation.

Another element of timing is how long to wait after instruction. It is commonly found that learning attainment is lower after a long break from study than before the holiday (known as “summer effects”). Therefore, it may be beneficial to assess learning at the end of the academic year, rather than the beginning.



Conversely, if longer term retention of learned knowledge, skills and values is important, a test too soon after instruction would not identify what might be forgotten in the following weeks. In order to avoid over-estimating long-term attainment, it may be better to leave more of a time gap between instruction and assessment.

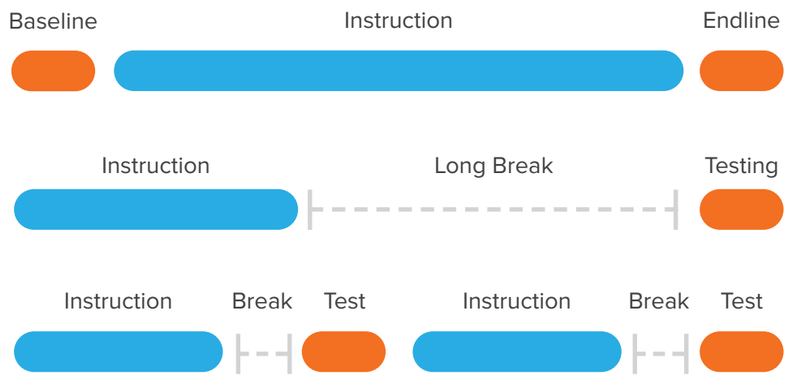



Figure 2 Options for Sequencing of Assessment

For formative assessment, information is needed very quickly in a manner that is accessible and usable by teachers. The assessment needs to fit within the teaching and learning process of the class being assessed. For summative assessments used for evaluative or policy making purposes, timelines are less urgent, but information still needs to feed into fixed policy or programme design cycles. The impact on teachers’ ability to fulfil their role should be considered and minimised. They should not impinge on other significant times within the academic calendar, such as preparation for major examinations.

**These factors usually provide the fixed points, from which timelines need to be drawn:**

When the assessment needs to be administered, based on timing of instruction and the academic calendar.

When the decisions need to be made based on information from the assessment.

There can be tensions between the two that need to be worked out with the compromises that result in the greatest assessment validity, while minimising negative impacts on teachers and students.

## 2.4 Definition of construct and assessment content

Closely related to the assessment purpose is the decision about the definition of the target construct. Constructs are defined in the core handbook and constructs have already been selected within Schools2030. For completeness, it is important to note the importance of this step here. The construct needs to be selected in line with the assessment's purpose and defined in detail, including all sub-domains.

Assessment designers then need to consider what will be assessed. The core handbook describes the distinction between the target construct and the assessed construct. The target construct is the skills, knowledge and values that test users need to draw inference about in order to assist decision making. The assessed construct is the combination of elements from the target construct that are included within the assessment. For constructs of limited scope, these may be the same. For larger constructs (e.g. curriculum content for a subject over 2 years' schooling), the assessed construct will necessarily be smaller.

The assessed construct should be selected such that results can be extrapolated to apply to the whole target construct. That is, the scores for the assessed should be the same (or very close to) the scores that would have been achieved if the whole target construct were measured (if that were possible). In order to achieve this, learning attainment of excluded elements should be closely related to learning attainment of included elements. Further, the selection should consider how important each element is, this may be considered in terms of the impact on future learning or other applications.

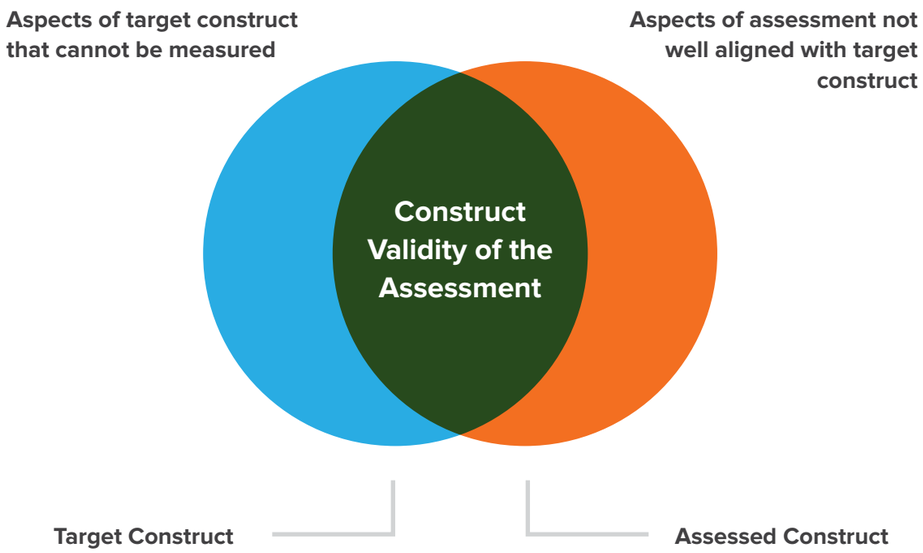



Figure 3 Target versus Assessed Constructs



Ultimately, these decisions require expert judgement from people who understand assessment, the construct and the test-takers. For higher stakes tests (such as those for selection or certification), this process is particularly important and will need to be more carefully defended than for lower stakes assessments.

The assessments we are developing for Schools2030 are low stakes assessments, in that they do not determine future educational pathways (such as examinations) and are not individually reported back to students or families.

## 2.5 Assessment specifications

The test specifications provide a complete operational plan for the assessment instrument. The specifications include the type of testing format, the number and breakdown of items, languages used, whether or not the items will include visual stimuli, the expected item scoring rules, the method of administration, time constraints for each item and/or the test as a whole and how test scores will be interpreted.

The table below provides an example of the types of information that might be provided in a test specification.

| SPECIFICATION CRITERIA | SAMPLE SPECIFICATION |
|---|---|
| **Number of questions** | 50 questions: 30 Selected Response Questions, 20 Constructed Response Questions. |
| **Time allowed for students to complete test** | 1 hour. |
| **Types of question** | Selected response – matching, Multiple Choice Questions with 4 options and one correct answer. Constructed response – short answer questions. |
| **Content covered (assessed construct)** | Letter recognition, blending, reading fluency, reading paragraphs, comprehension. |
| **Method of administration** | 1-to-1 using CAPI and a test booklet. |
| **Difficulty level** | Designed to capture ability from emerging literacy skills through to Grade 2 reading level. |

# 2.6 Testing format and question specifications

## 2.6.1 Choosing a testing format

The main categories of test format are constructed response (CR) and selected response (SR). Constructed response questions require respondents to provide a response using their own words (or choosing how to solve a mathematics question). They are often referred to as open-ended questions. There are a large – possibly infinite – number of possible responses. Essay questions, short answer questions and mathematics problems are examples of constructed response questions. Selected response questions are typically multiple choice questions. Here we discuss the relative merits of each, but guidance on developing both are discussed further in section 3.3.

Different types of questions have different strengths and weaknesses in assessing different constructs so the construct selected will partly determine these decisions.

Benefits and risks of using **constructed** response items


The purpose of the assessment may also shape the nature of the information required and therefore guide the choice of test format. The cohort of students that are being tested will also factor into the decision. Designers should consider whether the assessment purpose requires that testing should be similar in look and format to instruction, whether it should reflect real-world scenarios and applications or whether it should be novel.



Benefits and risks of using **selected** response items


## 2.6.2 Developing Question Specification

Question types should be defined as precisely as possible. For example, it is often possible to specify the action words involve (e.g. solve,  explain, define, describe, etc.). The length of the question should also be described so that item writers have a clear sense of the expectations. This will provide items of a uniform (and therefore  comparable) nature. Items should be long enough to clearly portray the task and avoid confusion. However, they should not be excessively long as this places a greater burden on reading and comprehension of the question and makes it more likely that a test taker will miss or forget elements of the question.

| Features of the Input Text | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Word count** | 125-135 words (including target words for gaps) | | | | | | | | | |
| **Domain** | Public | | | Occupational | | | Educational | | | Personal |
| **Discourse mode** | Descriptive | | | Narrative | | | Expository | | Argumentative | Instructive |
| **Content knowledge** | General | | | | | | | | | Specific |
| **Cultural specificity** | Neutral | | | | | | | | | Specific |
| **Nature of information** | Only concrete | | | Mostly concrete | | | Fairly abstract | | | Mainly abstract |
| **Presentation** | Verbal | | | | | Non-verbal (i.e. graphs) | | | Both | |
| **Lexical Level** | K1 | K2 | K3 | K4 | K5 | K6 | K7 | K8 | K9 | K10 |
| | | | | | | | | | | |
| **Lexical Level; further criteria** | The cumulative coverage should reach 95% at the K3 level. No more than 5% of words should be beyond the K3 level. (See Guidelines on Adhering to Lexical Level for more information). | | | | | | | | | |
| **Grammatical level** | A1-B1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level) | | | | | | | | | |
| **Avg sentence length** | 13-15 (This is an average figure. Individual sentences will span a range above and below the average) | | | | | | | | | |
| **Topic** | From topic list for B1. | | | | | | | | | |
| **Text genre** | Magazines, newspapers, instructional materials (such as extracts from textbooks describing important events or people). The texts are adapted to the level. Although not intended to be authentic, they should reflect features of relevant texts from the TLU domain. It should be possible to answer the questions: *where would a reader be likely to see a text like this outside the test?* and *is the genre relevant to TLU tasks important for Aptis General test takers at B1 level?* | | | | | | | | | |
| **Writer/Reader relationship** | The relationship is not specified. The texts will typically be written for a general audience, not a specific reader. | | | | | | | | | |

Figure 4 Example of a test specification from British Council [1]

Figure 4 shows an extract from a reading assessment specification for the British Council. It would be accompanied by some further explanation about the terms used, but it demonstrates the level of detail involved.

1 - http://events.cambridgeenglish.org/alte-2014/docs/presentations/alte2014-john-tucker.pdf

The specification shows the level of vocabulary and grammar to use as well as the genres. It specifies the levels of difficulty and the word count. In this example, a significant element of the item is the text provided to read so muchof the details to provide the criteria by which the reading samples should be selected. A mathematics assessment specification may focus on operations, number of digits, types of fractions that can and cannot be used, etc.

## 2.7 Number of items and time constraints

The number of items in a test depends on how long it takes students to complete each one and how long the test takers can reasonably be expected to concentrate on the task (which depends on their age and the nature of the assessment).

It also depends on the scope and range of abilities that need to be covered within the assessment instrument. Where there is a wide range of ability level among the test-takers piloting with a set of question or tasks (or a small set of tasks) to determine their approximate level of ability will be important.

To plan the number of items required, consider what is required to assess each sub-domain (component of the construct) validly and reliably.

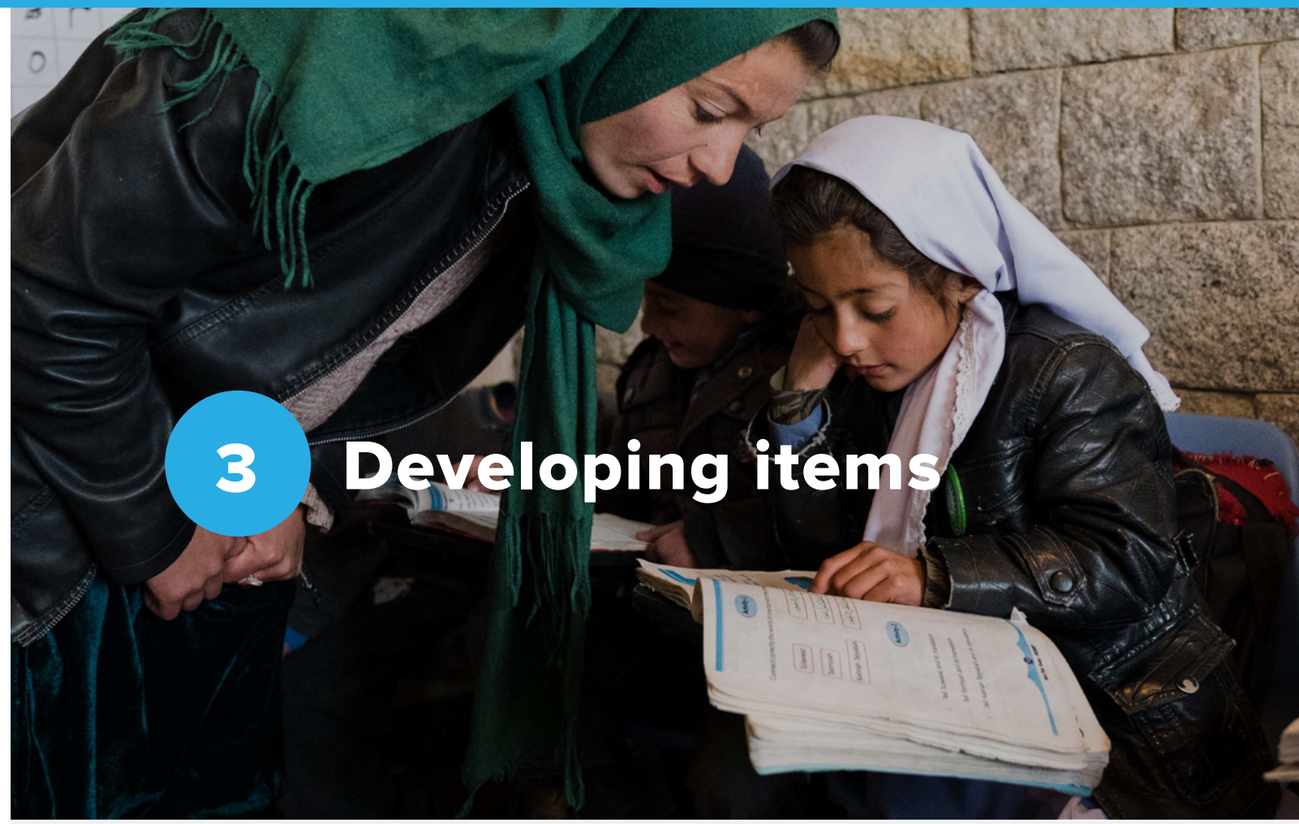
Consider whether time will be restricted or not. If it is not, students can take as long as they need, which may result in overly lengthy tests for the students and in difficulties in completing all assessments within the required timeframe. However, it avoids complications from considering how to score questions that are not attempted and how to treat them in the analysis.

Ethical considerations are also important. Children should not be made to feel uncomfortable or stressed. Tests should therefore be kept as short as possible. Steps should also be taken to avoid children having to respond to too many questions that they are unable to answer correctly. Stop rules can be used for this purpose, terminating the test or part of the test when a sequence (of set length) of incorrect responses are given.

## 2.8 Method of administration

Assessments can be administered one-to-one, in pairs or groups, or to whole classes or grades. Items can be asked or tasks can be set verbally or in writing. Responses can be provided in writing, verbally or by other means (such as computers or activities). Responses may need to be recorded by enumerators separately. This can be paper-based or technology based, using a tablet or mobile phone (CAPI). Paper-based administration to full classes of children usually relies on the children being proficient in reading so they can understand the instructions. One on one enumeration is usually needed where children are not proficient readers. This allows the enumerator to provide instructions verbally to the child and note down the child's response.





# 3 Developing items

## Summary

Developing items that can measure important content at the appropriate level of difficulty is one of the greatest challenges in assessment development. The Schools2030 Global Assessment Partners are available to provide support throughout this process. **We do this by:**

- Providing validity checks on your developed items (and can provide support throughout the item planning and writing process, if needed).
- Providing guidance materials on the item piloting process
- Analysing pilot data and providing recommendations to National Assessment Partner on how items can be improved

Broadly, there is a preparation phase, a writing phase and a review phase. During the preparation phase, item writers collect ideas and material. Depending on the nature of the assessment, this may be from reviewing tuition materials or identifying examples from real life that could serve as stimuli. For example, an assessment of literacy based on real life applications may require examples of articles, forms, adverts to assess understanding. These materials need to be gathered and should be from the context in which the assessments will be implemented. The framework also needs to be developed as above.

There is then a process of drafting and reflection. The item writers have to apply the framework, using materials that they have already identified or that they identify as they go. They need to consider every part of the question ensuring that they perform as intended. Checking the validity and fairness implications: does it measure the assessed construct? Will other factors about the student influence their performance (e.g. other skills or background)? Are some groups likely to perform better than other?

Item writers should plan time to reflect on items that they draft and make refinements. The nature of this work schedule depends on what works best for the item writer, but it is important that item writing is seen as a process of refinement and not as an activity that can be completed in a single effort.

Finally, it is useful to bring in independent reviews so that items are considered from other perspectives. Reviewers need to be knowledgeable about the cohort of test takers, how they learn and express the target construct and differences between students of differing ability. Most often, this requires some teaching experience.

## 3.1 Validity in item development

The core handbook defines assessment and validity and describes the main drivers of validity issues: Construct Irrelevant Variance (CIV) and Construct Under-Representation (CUR). The latter is largely a question for test compilation, but CIV needs to be considered by item writers.

Writers have to be very careful that the question is testing the intended construct. Consider how students might tackle the question and whether there are ways that students who are weaker in the target construct might perform better in the item. There will be validity issues if there are other skills or knowledge that could influence students' performance and that would vary within the cohort of test takers.



**Example: Construct Irrelevant Variance**

The classic example of CIV is when a child is administered a mathematics question in a language they do not understand. What is being tested becomes language, rather than mathematics knowledge and skills in this situation.

The content of the item needs to be considered. To make the test more responsive to instruction (that is, teaching more directly influences test scores), content should resemble the materials used by teachers. In order to test real-world applications, they should be typical of scenarios found within the culture.

## 3.2 Considering how students respond to test items

Item writers should use their experience to consider how students might respond to the question.

**In general the process for answering a test question is:**

1. Learning the material
2. Reading the question
3. Searching the memory
4. Matching their memory to the task
5. Generating the answer (which may be simple recall or may involve processes)
6. Writing an answer

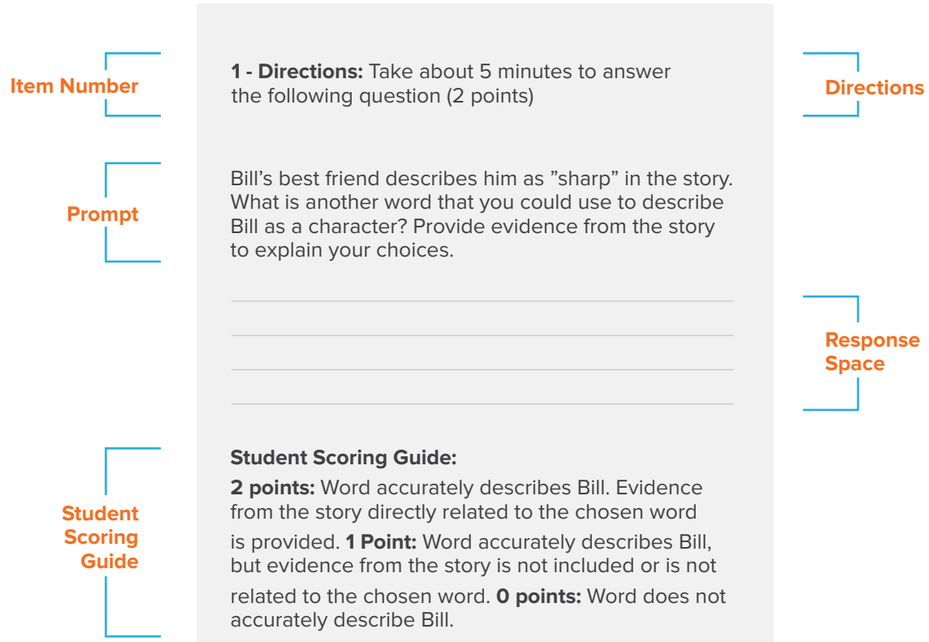
We should consider the elements of a question that can influence the complexity and difficulty for students. Some of these will be related to the construct and are therefore the difficulty that is necessary to ascertain information about students' attainment in the construct. Other contributing factors to difficulty may be unrelated to the target construct and should therefore be minimised.

**Factors affecting difficulty include:**

- The complexity of the knowledge required to answer the question.
- The number of steps involved in an answer.
- The level of familiarity and prior knowledge that a student may have about the content, application or procedures required to answer the question.
- Familiarity with the question format. Have students seen questions that look similar and use the same command words?
- The number and range of elements that the student has to draw together (from stimuli and prior knowledge) in order to answer the question.
- Complexity of language and grammar used.

**These should all be considered to reduce complexity and difficulty that is not related to the construct.**

**Anatomy of a Constructed Response Item**




## 3.3 Types of items and item components

### 3.3.1 Constructed response questions: Short- and long-answer questions

Constructed response questions are any questions where the student has to provide the answer without having any options to choose from. The alternative is selected response questions typified by multiple choice questions, which are discussed in the next section.

Constructed response items consist of the task/question and marking scheme. The task or question may refer to stimuli, such as text, pictures, diagrams, tables or graphs.

Constructed response questions can be usefully divided into short- and long-answer questions. Short-answer questions require an answer of one or a few words (or numbers). This may be an objective response where there is one correct answer or the marking scheme may allow some variation in responses.

Short answer questions can be more reliable than long answer questions because of the limited scope of correct responses. They can also enable a relatively large range of content to be assessed in less time.

Long answer questions, such as essay questions and mathematical problems requiring multiple steps, can be better suited to assessing some constructs, such as reasoning and evaluative skills. They can provide more space for students to demonstrate what they know and the skills that they possess. They generally require more complicated marking schemes, that require greater marker judgement. Ensuring reliability is a greater challenge and often requires additional or more-involved steps such as greater marker training, stronger invigilation processes and double marking.

Questions should be as brief and clear as possible. Language should be familiar to students and as simple as possible so that students that have the target construct skills and knowledge are not inhibited from answering the question correctly because of limited vocabulary.

Scenarios, examples and applications should be familiar to students and should avoid favouring some students over others (see the section on fairness in the core handbook).

The marking scheme needs to be designed together with the question as this is integral to the way that the question performs in eliciting evidence about the students' attainment in the target construct and encoding this into a format that can be analysed quantitatively.

Marking schemes can vary in the degree of structure. Highly structured mark schemes define how to assign scores point by point. There is a clear definition for what should receive a mark and what should not. The marker identifies everything that should be credited with a mark and adds up all the marks to provide a score.

Less structured marking schemes are more levels-based, whereby the marker is required to make judgements about which level a student's answer sits in by some set characteristics (e.g. persuasive writing, creativity, clarity of communication, etc.). The scoring rubric provides clear criteria for each level and each characteristic (an answer may be judged according to a variety of characteristics). The marking scheme may include model answers to demonstrate what the criteria listed may look like in students' responses.

Although assigning marks relates to the weighting of a single question within the test paper, this should not be considered at this stage. Weights can be adjusted when the test is compiled. Therefore, do not assign more marks to harder questions or to questions that are of greater importance (as defined in the test blueprint discussed in Section 1). Instead, the number of marks for the question should be the number required to distinguish between different levels of performance. Hence, single-answer questions are usually only assigned one mark, even if they are difficult or important.

**An extended mathematics problem would be assigned more marks in order to distinguish between those that knew the approach, but made an error, those that knew some of the process, but not all and those that could answer the question completely.**



### 3.3.2 Selected response: Multiple choice questions (MCQs)

Selected response items may be more appropriate for use in the Schools2030 programme, because teachers are using assessments to take action to improve learning. The time and effort available for the teacher to undertake the assessment and mark the assessment is limited.

Multiple choice questions are any questions for which students select one or more answers from a number of options.

**They consist of three parts:**


**1 - The Stem**
The question (or task), such as reading comprehension questions or mathematics problems. This may include a stimulus, such as some text or a picture or diagram, which is referred to by the question.

**2 - The Key**
The correct answer.

**3 - The Distractors**
incorrect answers provided as options.

They are often considered to be best for testing information recall, but they can be equally effective at assessing higher order skills, when designed carefully.

**There are a number of benefits to using multiple choice questions:**

- They are quick and easy to mark, particularly using computer marking.
- They can be more reliable to mark because answers are unambiguous and require no interpretation.
- They do not require the student to write, so writing skill and hand-writing do not influence performance.
- They can be faster to administer because recording a response takes very little time (although deciding on which answer to choose may take longer). Therefore, they can be used to cover a lot of content within a relatively short time.

**However, they also have problems and limitations:**

- They generally lend themselves to a binary score (correct or incorrect) so there is no information if a student can get part of the way to an answer, but is not able to complete the process and find the correct answer. Students do not get partial credit and teachers may not gain information about why students got the answer wrong.
- They can be difficult to design because there needs to be enough plausible incorrect answers.
- They require care in interpreting results because it is possible to get answers correct by guessing (rather than knowing).
- Although MCQs can be used to assess a range of skills, it is more difficult to use them for some constructs.

Commonly, the stem is a question and there is one correct answer (key) from a list of four or five options (three of four distractors). However, there are a number of variations and related formats.

True or false questions provide statements and students have to decide whether they are true or false. It can be difficult to find uses that are the right level of difficulty and do not act as 'trick questions' that try to lure students into giving incorrect responses. Sometimes true or false questions can be formatted as multiple choice questions, but where the student can select more than one answer. **For example, the question might be:**

**"Which of these are fruits? (select all that apply)"**

- Banana
- Carrot
- Mango
- Potato
- Beans

**This is equivalent to asking true or false questions using statements "The ___ is a fruit" for each of the options given.**

Matching questions require students to match items from two sets. For examples, the two sets could be pictures of shapes and shape names or dates and events. Students would be required to match the shape to its name or the event to the year that it happened.



**Best answer questions provide a list of possible answers in the same way that multiple choice questions do, but students are asked to select the best answer instead of the one correct answer. One situation where this may be helpful is to assess understanding of processes.**

Students can be given scenarios and be asked to select the best course of action. The test designer may have an order of preference for the non-optimal options and can choose scoring criteria accordingly.

As mentioned above, multiple choice questions can be difficult to write and require a great deal of care. It is important to keep in mind how the question will feel for the students answering it and how they might approach the task of deciding on an answer. Taking this into account, the stem must be as clear as possible without unnecessary words or details.

They should not attempt to 'wrong-foot' the respondent by providing information that is not required for the task. It may be helpful to mark important words in bold to help ensure students to understand the question.

In general, questions are better than sentences to complete or fill in the gap. For example, "What would you expect to happen to a sealed empty plastic bottle when you place it in water?" is preferred to "When placed in water, a sealed empty plastic bottle would be expected to...". This is because the 'complete the sentence' format can require the student to repeat the question with each of the options to check whether the statement is true or false.

Similarly, to avoid confusion, avoid questions that are framed negatively. For example, "Which of these are fruits?" is preferable to "Which of these are NOT fruits?".

Distractors need to be chosen carefully so that students who do not know the answer are equally as likely to select the distractors as the key. To do this, distractors should ensure that there are no effective strategies for students to use other than to know the answer because of attainment in the target construct. The most important way to do this is to ensure that all distractors are plausible, but incorrect – they are answers that someone might give if they do not have the skills or knowledge being assessed. Answers that are clearly ridiculous can help a student to arrive at the correct answer even if they do not have the required skills and/or knowledge. If there are not enough plausible wrong answers to provide the desired number of distractors, it is better to have fewer distractors than include implausible distractors.

For the same reason, the correct answer must be assigned to random positions in each question so that students cannot simply select ‘D’ for all questions and get 100%. Similarly, the distractors should be similar to the key in length and style so that students cannot get the correct answer by selecting the option that is different from the rest.

Grammar can give away the correct answer so ensure that all distractors are grammatically correct. This particularly applies to ‘complete the sentence’ format stems, but can also be a consideration for questions.

**Duplicating teaching and learning materials (such as textbooks) should also be avoided because students may select the correct answer because they recognise the text and not because they understand it.**

**3 RULES FOR SELECTED RESPONSE QUESTIONS**


Figure 5 Rules for Selected Response Questions

In general, it is best to avoid “all of the above” and “none of the above” options. Discounting one option will also discount the “all of the above” option so it does not act in the way that most distractors would. Conversely, a student only needs to identify two correct answers to be able to select “all of the above”. “None of the above” can only be used when the answer is absolutely correct, such as arithmetic problems or historical dates. In these cases it is preferable to provide the correct answer so that the student has to positively identify it rather than simply ruling out alternatives. It should not be used when the stem is negatively framed as this would create confusion for the student.



# 4 Test design and assembly

## 4.1 Test design

The next step is to compile items into a test instrument that can be administered. This will involve developing multiple test instruments to pilot and then using the information from the pilot to design the final instrument. Guidance about how to pilot assessments will be provided by the GLACP separately and support during the piloting process will be available if needed.

When compiling tests we need to consider key threats to validity. Firstly, we need to avoid construct under-representation (CUR). The test needs to cover all elements of the assessed construct, with scoring matching the intended weightings as defined in the blueprint.

It is better to plan item writing to ensure sufficient coverage (allowing for the fact that some items will not perform as intended when they are piloted).Therefore, more items than needed should be developed and piloted. If there are gaps at this stage, more items will need to be written or poorly performing items will need to be fixed.

## Ceiling Effects

In order for a sub-domain to be satisfactorily covered, items need to cover a range of difficulty to match the range of attainment/ability of students being tested. Ceiling effects occur where all items are too easy so that most students answer all correctly. This means that the test will provide information about what students can do, but it does not tell us where their attainment in the construct ends.

### Example: Ceiling Effects

If a test focuses on number recognition and counting and all students can answer all of the questions correctly, the test will not provide any information about whether they can also perform addition and subtraction.

The converse problem is floor effects, where the items are all too difficult so that most students answer very few correctly. In this instance, we know what students cannot do, but we do not know what they can do and where their capacity ends.

### Example: Floor Effects

A test might contain only addition, subtraction, multiplication and division, but students can answer very few items correctly. We would gain no information about lower-level skills such as number recognition and counting. Both problems represent a wasted opportunity to obtain useful information.

## Construct Irrelevant Variance

Construct Irrelevant Variance (CIV) is the other threat to validity to consider. Anything that might make the test more difficult to complete should be avoided. This includes making sure that the layout is clear, and the overall instructions are clear. Fonts should be easily readable and there should be space on the page to avoid it having a daunting or stressful effect on the student. Questions should not overlap pages.

### Example: Construct Irrelevant Variance

Text to read and comprehension questions based on it should be on the same double page so that students do not have to repeatedly flick between the two pages, holding information in their heads as they do. Instructions should help to alleviate stress, particularly for low-stakes assessments.



Another factor that can cause stress or confusion for students is a strange order of questions. Questions that jump around in difficulty within the same sub-domain can cause confusion and frustration for students as they might expect questions to be of increasing difficulty and question whether they have misunderstood an easy question. Alternatively, they may struggle to answer a difficult question early on and become discouraged affecting their motivation for following questions. As a result, questions tend to start with the easiest and get progressively difficult within the same sub-domain. Alternatively, all questions may be able to assess the full range of ability. The principle here is to consider the test paper from the perspective of a student.

Test length can also cause CIV problems because fatigue becomes a factor so students with better stamina (or speed) score better than students with the same level of ability, but who become tired sooner.

Sometimes there is a wide range of ability levels in a test-taking cohort. This can cause challenges to CUR and CIV. The range of ability levels requires an equivalent range of item difficulty which can make it difficult to cover the full assessed construct without the test becoming too long so there may be a problem with CUR. CIV is also a risk because students may become exhausted answering questions that are too easy, which affects their performance at more difficult questions. The risk for students with lower attainment may be greater as it can be very uncomfortable and stressful to face lots of questions that they are unable to answer.

**When instruments are developed they should be checked for face validity, which simply means validity at its face. This asks whether the assessment instrument appears to be valid by looking at the instrument alone. It entails expert reviews of the instrument to check for common causes of validity problems. As the GLACP, Oxford MeasurEd will provide face validity checks.**

### Face Validity Checklist

| | Yes | No |
|---|---|---|
| • The length of the test is appropriate for the use. | ○ | ○ |
| • The test is clearly laid out and easy to read. | ○ | ○ |
| • Items do not cross multiple pages. | ○ | ○ |
| • Stimuli, materials and language appear to be suitable for the use and context. | ○ | ○ |
| • The test is likely to measure the intended construct and not other constructs. | ○ | ○ |
| • The test appears likely to behave similarly for different groups of students. | ○ | ○ |

## 4.2 Item analysis

Oxford MeasurEd will provide item analysis on pilot data in a clear and easy-to-interpret manner, highlighting the implications of any information that comes out of the analysis. Recommendations will be provided for NAPs to discuss with National Leads, in order to make final decisions on the test.

Item Response Theory (IRT) provides statistical methods that can be used to analyse items to help inform selection when constructing the instrument. Oxford MeasurEd will provide IRT analysis after piloting to elicit information on how the items have performed and provide recommendations to NAPs. The most useful two measures that it provides are item difficulty and item discrimination.

The difficulty of an item can be assessed by simply looking at the percentage of students who got the answer correct. In IRT, the measure of difficulty takes into account the 'ability'[2] of the students who answered correctly. In other words, IRT takes account of which items each individual got right, not just how many items the individual got right.

IRT can place item difficulty and student achievement on the same scale. This useful because achieving 75% correct in a test is only meaningful if the difficulty of the test is known. IRT is a probabilistic model, whereby if an item and a person sit on the same place on the scale, the student has at 50% chance of answering that item correctly.

**The difficulty of all items can be mapped to ensure that the full scope of difficulty levels is included in the test. Item difficulty measures can also be used to avoid ceiling and floor effects.**

Item discrimination is the ability of an item to distinguish between students of higher and lower ability. Discrimination index scores fall between - 1 and 1. A positive score means that higher ability students are more likely to answer correctly than lower ability students.

A negative score means that lower ability students are more likely to answer correctly than higher ability students. A score of 0 means that higher and lower ability students are equally likely to answer correctly. Generally, a discrimination index score of 0.3 is fair and an item with a discrimination index score of 0.5 discriminates well.

2 - by which we mean the attainment in the target construct

IRT can produce Item Characteristic Curves – such as the graph shown below – for all items piloted or with previous uses. The graph shows the probability that a student will answer the question correctly on the vertical axis (Y-axis). The ability of the student (in the assessed domain) is on the horizontal axis (X-axis). For this item, the lower ability students to the left are less likely to answer the question correctly (at the bottom) than the higher ability students to the right. As students' ability gets greater, so does their probability of answering the question correctly.

The further to the right that the curve crosses the 0.5 probability line, the more difficult the item is. The steeper the curve, the greater the discrimination. The difficulty can and should vary within an assessment instrument. The slope of the curve should not get too flat as this indicates that the item does not provide much information about the ability of the student. Curves with two humps or plateaus may suggest that the item is not working as intended. Downward sloping sections of curve indicate that at that ability range, higher ability students are less likely to answer correctly so there is likely something in the question that is sending them in the wrong direction or confusing them.

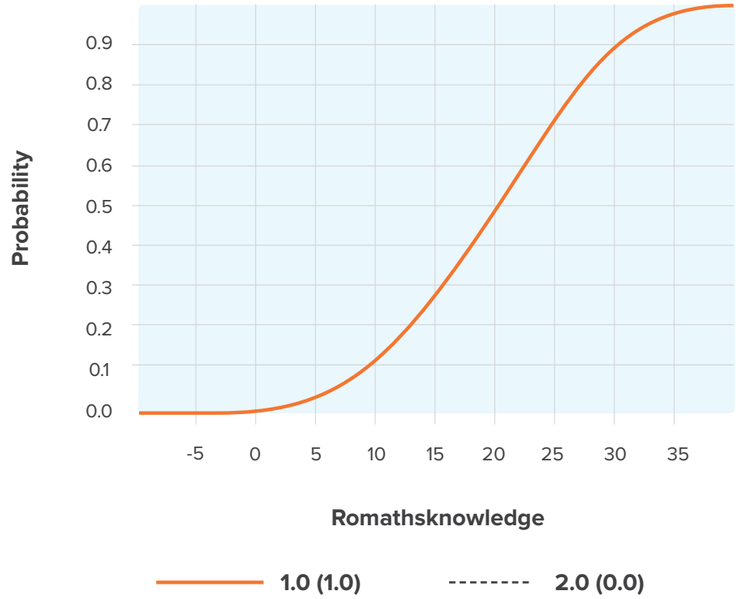
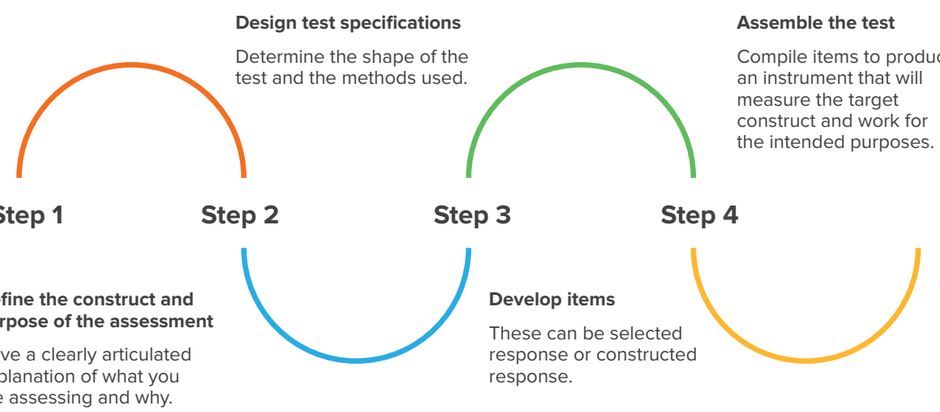



Figure 6 Example of an Item Characteristic Curve

# 5 Conclusion

**This handbook has provided guidance for developing assessment of academic skills. The assessment development process is as follows:**




The test can then be piloted and administered.

Throughout the process, it is important to consider the validity implications of decisions. This requires maintaining sight of the target construct and the intended purpose and being aware of how students might respond to items and the test as a whole.

**The handbook is intended as a guide and we hope that it is a useful companion for the process. Oxford MeasurEd is available to provide support and the Assessment Hub is a useful starting place to discuss experiences and seek support.**



# DESIGNING LEARNING ASSESSMENTS